UNIVERSITÉ DE STRASBOURG

Ecole Doctorale des Sciences de la Vie et de la Santé
IGBMC - CNRS UMR 7104 - Inserm U 964

THÈSE présentée par :
Benjamin LINARD

soutenue le : **15 octobre 2012**

pour obtenir le grade de : **Docteur de l'université de Strasbourg**

Discipline/ Spécialité : Bioinformatique

Développement de méthodes évolutionnaires d'extraction de connaissance et application à des systèmes biologiques complexes

(Development of evolutionary knowledge extraction methods and their application in biological complex systems)

THÈSE dirigée par :

  Mme THOMPSON Julie   Directeur de recherche, Université de Strasbourg

RAPPORTEURS :

  M VANDEPOELE Klaas   Directeur de recherche, VIB / Ghent University
  M GIBRAT Jean-François   Directeur de recherche, MIG-INRA

EXAMINATEURS :

  Mme DESPONS Laurence   Maître de conférences, Université de Strasbourg

Remerciements

Je tiens à exprimer ma sincère reconnaissance à M. VANDEPOELE Klaas, M. GIBRAT Jean-François et Mme DESPONS Laurence pour l'honneur qu'ils me font de juger cette thèse.

Cette section est sûrement la plus difficile à écrire car elle concrétise en quelque sorte la fin d'une aventure… Je vais essayer de n'oublier personne, mais j'ai tellement de personnes à remercier.

Commençons par le laboratoire. Merci à tous de m'avoir si bien accueillis durant toutes ces années. J'ai toujours été motivé pour venir travailler au laboratoire, ce n'est pas seulement grâce au sujet mais aussi grâce à la bonne ambiance qui a toujours régné dans les bureaux et aux petites blagues qui fusaient régulièrement... Merci Luc, Raymond et Laëtitia pour tout le temps que vous m'avez consacré pour m'expliquer les méandres des serveurs de l'institut. Merci Wolfgang, Nicolas et Nicodème pour vos conseilles avisés sur divers outils statistiques. Merci Odile et Fréderic pour avoir été les premiers à me permettre d'intégrer ce labo et pour avoir initié la petite étincelle qui m'a donné l'envie de continuer. Merci Jean pour tous tes conseils et les petites références star wars du quotidien. Merci Hoan pour ton aide sur DB2 et pour toutes les nouvelles mondiales bizarres que tu dénichais régulièrement. Merci Laurent et Véronique pour tous vos conseils. Merci Dao pour m'avoir montré ce qu'est vraiment la persévérance et pour tous les plats vietnamiens louches que j'ai goûté. Merci à Alexis, Raphaël, Can et Marc pour leurs contributions respectives à OrthoInspector. Merci également à Alin, Xavier et Vincent pour toutes les discussions autour de nos repas au RU.

Enfin il reste deux personnes du laboratoire pour qui les mots de remerciement me manquent. Olivier et Julie, je pense que beaucoup vous ont déjà remercié pour tout ce que vous avez fait pour eux scientifiquement et professionnellement parlant. Alors c'est pour un autre point que je vais vous remercier. Je ne sais pas si vous vous en rendez toujours compte, mais vous avez un don que peu de gens possèdent et ça, je l'ai compris en commençant à discuter avec d'autres personnes dans le monde scientifique et en particulier en observant d'autres « chefs ». Vous savez motiver les troupes, vous savez partager votre passion et vous êtes capable de tirer le meilleur des compétences de chacun en rendant globalement tout le monde satisfait après les efforts. J'ai vu trop peu de personnes capables de réaliser ça durant ces trois ans et c'est pourtant tellement important pour donner à des débutants comme moi l'envie de persévérer et d'aller plus loin. C'est pour celà que je veux avant tout vous remercier et pour ce sourire que vous avez toujours affiché en regardant mes arcs-en-ciel de barcodes. ;)

Quelques mots également pour tous les gens de l'institut ou d'ailleurs qui ont eu beaucoup d'influence sur ma thèse. Merci à Serge, Guillaume et Damien pour leur aide sur les serveurs et leur disponibilité. Merci aux membres du service communication pour les fêtes de la science que nous avons menés ensemble. Merci à Georges Labouesse et Laurence Drouard pour m'avoir choisi pour les enseignements d'OpenLAB et merci à toute l'équipe d'OpenLAB pour les inoubliables souvenirs que l'on s'est forgé lors de nos interventions (ah là là, ces ados…).

Une pensée aussi pour tous les membres du SPB, vous êtes nombreux, il me faudrait 1 page rien que pour vous remercier. On a monté une association super en à peine 1 an et on a réussi à convaincre beaucoup de hautes instances que c'est une chose utile ! Je pense que rien que pour ça, on a gagné notre pari.

Enfin un petit mot plus personnel pour certaines personnes…

Merci à Nicolas, Ismail, Tao et Claire R. Je n'ai pas à les écrire, vous savez toutes les choses pour lesquelles vous avez compté. Mickaël et Damien, c'est la même chose. En plus grâce à moi vous êtes célèbres à l'IGBMC maintenant ! Un petit mot également pour Lena, Claire B., Ebe, Florianna, Deepika, Thomas, Jérôme, Adrien, Rose-Marie, Justine, Sara, Laure, Caroline, Morgane, Isabelle, Mélanie, Nadia, Terrence, Twix, Firas, Anna, Iskander et Katarzyna pour les tous les bon moments que l'on a partagé ensemble autour d'un café, d'une bière fraîche ou en allant au sport.

Pour finir, merci à ma famille et surtout à Jonathan et Marjolaine et vous deux papa et maman. Cette réussite est aussi la vôtre et rien de tout cela n'existerait si vous n'aviez pas toujours été là pour tant donner et pour m'aider à réaliser mes projets.

Résumé de la thèse (version française)

La biologie des systèmes, une opportunité pour les études évolutionnaires

L'évolution, principe inhérent à toute forme de vie, est un aspect fondamental de la biologie. La vie ne connaît pas la stabilité et la fixité, toute espèce se transforme subtilement avec le temps. Les indices de ces transformations peuvent s'observer à tous les niveaux. Tout d'abord au niveau moléculaire, avec la fixation de mutations génétiques, la conservation/la modification de structure protéiques 3D, l'apparition/la perte de gènes ou de familles entières de gènes… Mais les principes de l'évolution ne s'appliquent pas qu'au niveau moléculaire, l'évolution induit des remodelages qui s'opèrent à tous les niveaux biologiques. La biologie des systèmes s'est beaucoup développée ces dix dernières années, confrontant plusieurs niveaux biologiques (molécule, réseau, tissu, organisme, écosystème…). Elle a démontré que le gène n'est pas seul responsable de l'apparition d'un phénotype. Notre compréhension du vivant s'est étendue au rôle des phénomènes épigénétiques, au rôle de la dynamique des processus intra et inter cellulaires jusqu'à une modélisation des interactions tissulaires. Toutes ces études se sont réalisés dans un contexte couramment appelé la biologie des systèmes, qui tente de décrire et de prédire le comportement d'un phénomène biologique en tenant compte de tous les niveaux biologiques.
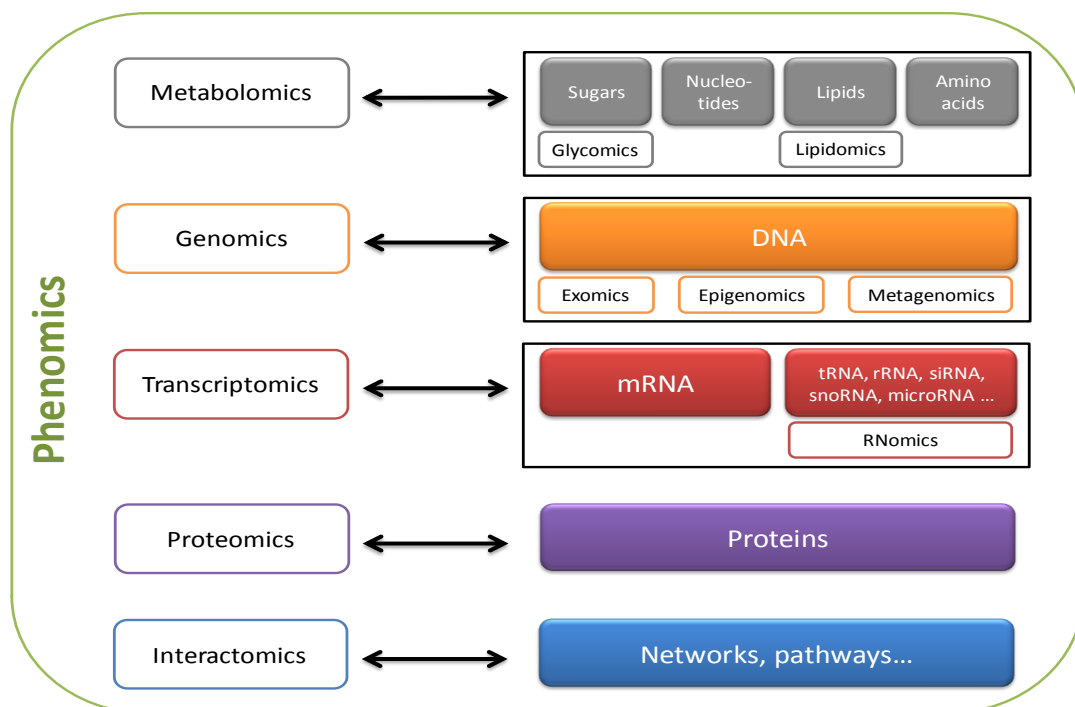
Figure 1. Vue générale des principaux omics utilisés en biologie des systèmes.

Pour comprendre ces phénomènes, ce domaine en plein essor produit de très nombreuses données à haut-débit, multipliant l'apparition de nouveaux « omics » et plaçant la bioinformatique comme une discipline indispensable à l'ère post-génomique (figure 1). Du point de vue de l'étude de l'évolution, la biologie des

systèmes offre de nombreuses possibilités. Les études évolutionnaires peuvent enfin ne plus se restreindre à la seul variation du gène mais s'étendre à la variation des systèmes cellulaires. Ce contexte a tout récemment donné naissance à un nouveau domaine, encore balbutiant, mais riches en opportunités : la « biologie évolutionnaire des systèmes » (evolutionary systems biology). Son but est de réunir la mécanistique détaillée de la biologie des systèmes avec le domaine plus ancien de l'évolution pour comprendre comment un organisme répond à des perturbations. En effet, ce qui est conservé par l'évolution est généralement essentiel pour le fonctionnement d'un organisme et les variations sont autant d'innovations potentielles. Etudier l'évolution d'un système biologique permet donc de comprendre la façon dont s'exercent les pressions évolutives au niveau des éléments d'un processus, les variations induisant ou non des transformations au niveau d'un organisme.

La biologie évolutionnaire des systèmes

Les premières études de « biologie évolutionnaire des systèmes » ont permis de mettre en évidence plusieurs tendances évolutives au niveau d'un organisme. Concernant la diversité inter-espèce, la plasticité des réseaux biologiques a été mise en évidence : l'inactivation d'un gène peut être compensée par d'autres processus biologiques et les gènes d'un processus peuvent être recrutés dans de nouvelles voies selon les espèces considérées. A l'inverse il a été démontré que certains modules d'interactions biologiques sont conservés par l'évolution, indépendamment des mutations génétiques. Eventuellement dormant dans une espèce, des processus biologiques entiers peuvent être réactivés par mutagénèse ou lors de changements liés à l'environnement. Au niveau intra-espèce, la démocratisation des techniques de séquençage à haut-débit a permis de démontrer la grande diversité des transcriptomes liés à un même phénotype, et ce en particulier dans le cas des maladies génétiques multi-géniques. Ces quelques exemples reflètent le nouveau consensus qui commence à émerger de la biologie évolutionnaire des systèmes : la variation s'applique à tous les niveaux biologiques. De nombreuses études se sont alors intéressées à ces variations d'un point de vue de l'évolution et ont définies de nombreuses nouvelles variables décrivant un système : propension à la perte d'un gène, nombre de paralogues, niveaux d'expression, centralité dans un réseau d'interaction... Toutes ces variables biologiques sont interdépendantes et démontrent l'aspect multidimensionnel d'un système biologique complexe (figure 2). Etudier l'évolution d'un système biologique nécessite donc d'intégrer toutes ces variables multidimensionnelles de plusieurs espèces dans un cadre unificateur, exploitable par de puissantes méthodes formelles de fouille de données et d'extraction de connaissances.

—	NP	PPI	GI	EL	CAI	PA	KE	PGL	ER
NP	*								
PPI	++	*							
GI	++	+	*						
EL	+++	+++	−	*					
CAI	ND	+++	ND	+++	*				
PA	ND	+++	ND	+++	+++	*			
KE	+	+++	−	+++	+++	+++	*		
PGL	NS	− −	NS	− − −	ND	ND	− −	*	
ER	− −	− − −	−	−	−	−	−	+++	*

Figure 2. Corrélations du point de vue de l'Evolution entre plusieurs variables biologiques.
Une corrélation positive est indiquée par un signe + et une corrélation négative par un signe - . CAI, codon adaptation index; EL, expression level; ER, evolutionary rate; GI, number of genetic interactions; KE, lethal effect of gene knockout; NP, number of paralogs; PA, protein abundance; PGL, propensity for gene loss; PPI, number of physical protein–protein interaction partners. ND, not determined; NS, not significant. Adapté de Koonin and Wolf, 2006.

Vers l'intégration multidimensionnelle en biologie évolutionnaire des systèmes

Depuis la découverte de la structure de l'ADN, la génomique, très centrée sur le gène, a apporté de nombreuses réponses sur les mécanismes évolutifs qui induisent la lente transformation des gènes et leur transmission, mais peu de réponse pour les autres niveaux biologiques. Durant ma thèse, je me suis intéressé au développement de nouvelles méthodologies et de nouveaux outils pour étudier l'évolution des systèmes biologiques tout en considérant l'aspect multidimensionnel représenté par les variables liés à plusieurs niveaux biologiques (génome, protéome, réseau, phylum…). Pour la première fois, des techniques formelles d'extraction de connaissance sont appliquées à la fois au niveau génomique et systémique pour l'étude de l'évolution. De ce fait, cette thèse tente de palier un manque méthodologique évidant pour réaliser des études haut-débit dans le récent domaine de la biologie évolutionnaire des systèmes. La considération de l'aspect multidimensionnel des processus biologiques nous permet ainsi de décrire de nouveaux messages évolutifs liés aux contraintes intra et inter processus. En particulier, mon travail a permis (i) la création d'un algorithme et un outil bioinformatique dédié à l'étude des relations évolutives d'orthologie existant entre les gènes de centaines d'espèces, (ii) le développement d'un formalisme original pour l'intégration de variables biologiques multidimensionnelles permettant la représentation synthétique de l' histoire évolutive d'un gène donné, (iii) le couplage de cet outil intégratif avec des approches mathématiques d'extraction de connaissances pour étudier les perturbations évolutives existant au sein des processus biologiques humains actuellement documentés (voies métaboliques, voies de signalisations…).

(i) L'inférence, la visualisation et l'analyse des relations d'orthologie

L'orthologie est une relation d'homologie liant deux gènes partageant le même ancêtre commun et issus d'un évènement de spéciation. Cette relation est centrale en génomique comparative et évolutionnaire car on considère généralement (mais pas exclusivement) que deux gènes orthologues sont fonctionnellement similaires. Définir les relations d'orthologie est donc une étape essentielle pour pouvoir comparer des

caractères moléculaires entre espèces, ces caractères allant du résidu jusqu'au réseau de gènes décrivant un processus. J'ai donc développé OrthoInspector, une suite logicielle permettant l'inférence des relations d'orthologie existant entre de nombreux génomes. OrthoInspector intègre un nouvel algorithme dont l'originalité est de considérer l'inparalogie comme base pour détecter l'orthologie, cette approche étant applicable à grande échelle tout en gardant une bonne sensibilité/spécificité. De plus, contrairement à la plupart des méthodes existantes, OrthoInspector est accompagné de nombreux outils de visualisation et d'analyse de ces relations. Ce programme répond donc à un besoin de plus en plus pressant à l'ère post-génomique d'outils intégratifs permettant l'analyse de très grandes quantités de données tout en fournissant des outils permettant de résumer ces données de manière efficace et informative. A partir de cet algorithme nous avons pu produire une base de donnée contenant les relations d'orthologies des protéomes complets de nombreuses espèces eukaryotes (Figure 3). Cette base de donnée est disponible en ligne pour la communauté.
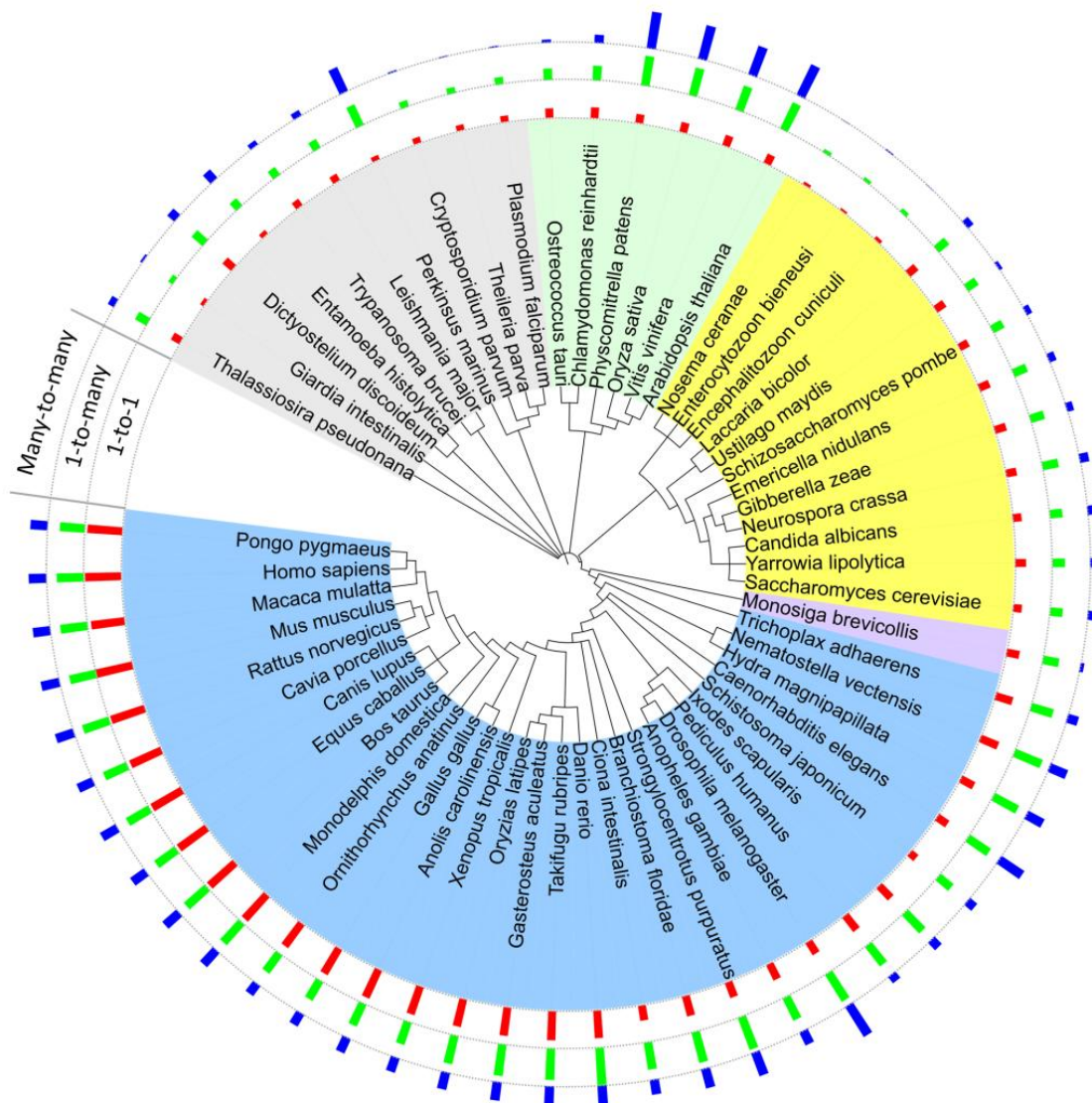
Figure 3. Les 59 espèces vertébrées composant la base de données OrthoInspector. *Les couleurs correspondent aux principaux clades: viridiplantae (vert), fungi (jaune), choanoflagellida (violet), metazoa (bleu) autres eukaryotes (gris).*

(ii)   Un formalisme original pour représenter l'histoire évolutive des gènes

La seconde étape de ma thèse a été de développer une méthodologie originale permettant d'étudier le scenario évolutif d'un gène qui a aboutit à son état actuel. Ce travail a donné naissance au concept de l'EvoluCode (Evolutionary Barcode ou code barre évolutif) : un profil décrivant l'histoire évolutive d'un gène à une échelle évolutive donnée. L'EvoluCode est une représentation synthétique qui permet d'intégrer, de visualiser et d'analyser des paramètres diverses issus de multiples niveaux biologiques (génomique, protéique, réseau…) et de multiples organismes. Ces codes barres peuvent s'adapter à tout type de paramètre biologique et peuvent être facilement mis à jour. De plus, leur structure matricielle permet de les comparer facilement selon divers métriques et de façon automatique, permettant ainsi leur utilisation dans des études à haut-débit. Un travail particulier a été fourni pour décrire la typicité des différents paramètres utilisés dans leur « contexte évolutif ». En effet, pour chaque espèce, l'état particulier d'un paramètre peut être considéré comme typique ou atypique par rapport à la valeur communément observée chez cette espèce. L'intégration de multiples paramètres de différents niveaux biologiques définit alors une combinaison complexe de valeurs typiques ou atypiques et permet de décrire un scénario évolutif complexe qui ne pourrait pas être résumé, par exemple, par une simple analyse de séquences.

Nous avons appliqué cette nouvelle méthodologie des EvoluCodes sur le protéome humain en reconstituant les histoires évolutives géniques au sein des vertébrés. Pour ce faire, nous avons compilé des variables de plusieurs échelles, telles que le contexte génomique, l'organisation et la conservation protéiques ou encore, la distribution phylogénétique à partir des données d'OrthoInspector (figure 4). L'intégration de toutes ces variables dans une structure facilement exploitable et l'étude au niveau du protéome complet nous a permis plusieurs conclusions. Premièrement, nous avons mappé nos EvoluCodes sur le génome humain et nous avons observé, grâce à un outil de visualisation dédiée, des clusters d'histoires similaires qui correspondent souvent avec des clusters de gènes connus. Deuxièmement, nous avons exploités notre formalisme pour appliquer une méthode d'extraction de connaissances, la classification non-supervisée, afin de regrouper les EvoluCodes similaires et d'étudier des enrichissements fonctionnels potentiels. Cette démarche nous a permis de mettre en évidence des tendances évolutives liées à des groupes de protéines partageant une fonction ou une localisation cellulaire commune. Cette seconde conclusion reflète l'intérêt d'étudier l'évolution des gènes dans le contexte de leurs processus biologiques et de ne pas se contenter du seul gène.
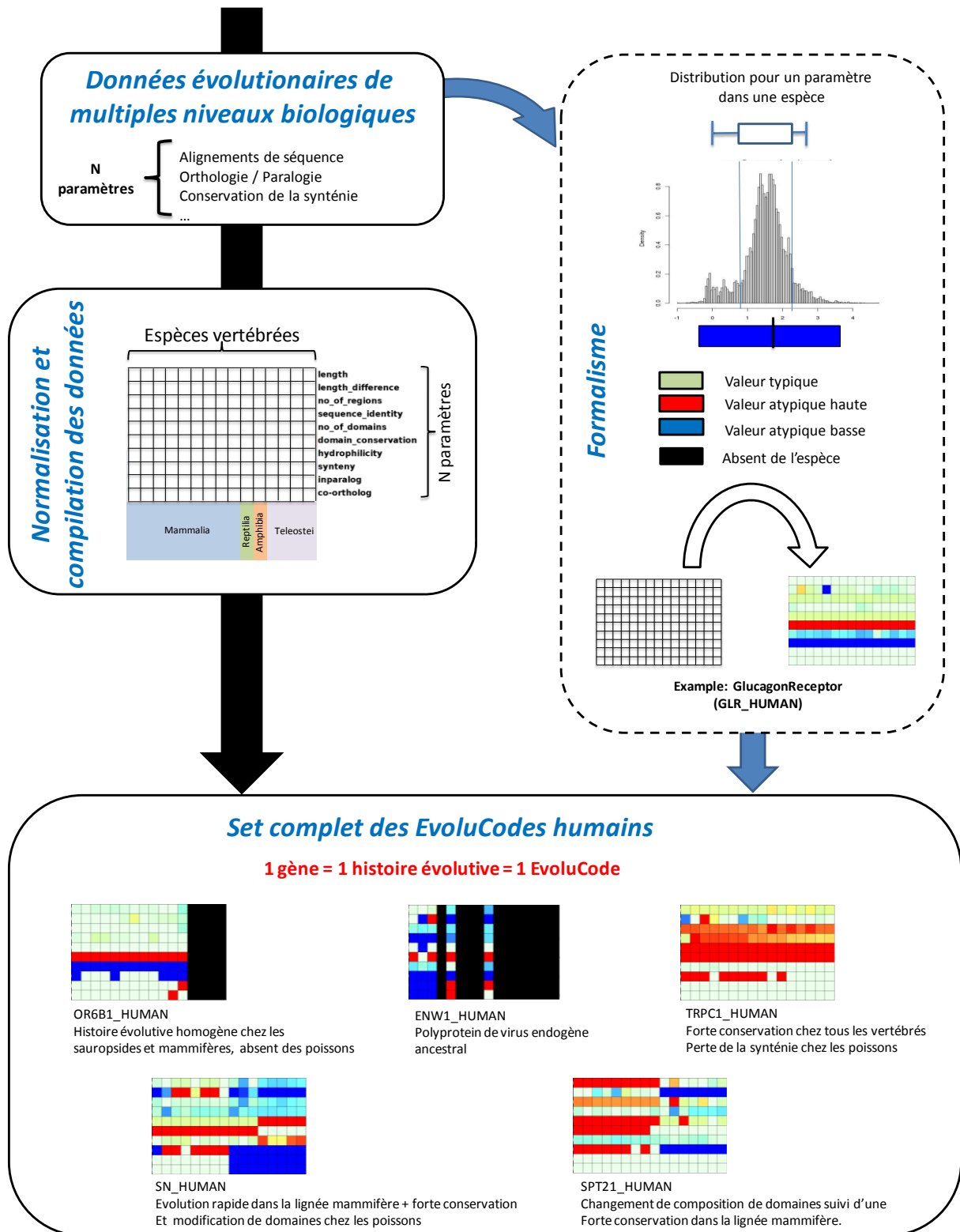
Figure 8-4: Vue d'ensemble du processus de construction des EvoluCodes. *Différents paramètres évolutionnaires sont compiles depuis différentes sources et sont organisés dans le formalisme du code barre. Parallèlement, la variation de ces paramètres est décrite statistiquement en les comparant avec la référence humaine. Le modèle statistique est utilisé pour coloriser les evolucodes, permettant ainsi une visualisation directe des différents profiles évolutifs caractérisant différant gènes.*

(iii)   Etude multidimensionnelle de l'évolution des réseaux biologiques humains

Les résultats encourageants apportés par les EvoluCodes appliqués au protéome humain nous ont motivés à étudier l'histoire évolutive de l'ensemble des processus biologiques humains. Nous avons donc localisé typologiquement nos EvoluCodes sur les réseaux biologiques humains actuellement documentés. Nous ne nous contentons plus d'analyser l'histoire évolutive d'un gène, nous voulons explorer l'histoire d'un processus défini par un réseau de gènes et lié à une réponse biologique précise. Nous avons mis en place un protocole intégrant un algorithme de détection d'anomalies pour identifier les gènes présentant une histoire évolutive originale par rapport aux autres gènes impliqués dans un processus. Cette atypicité d'un gène est définie via nos EvoluCodes, elle intègre donc des paramètres de différents niveau biologiques dans l'étude de l'évolution des processus biologiques humains (figure 5).
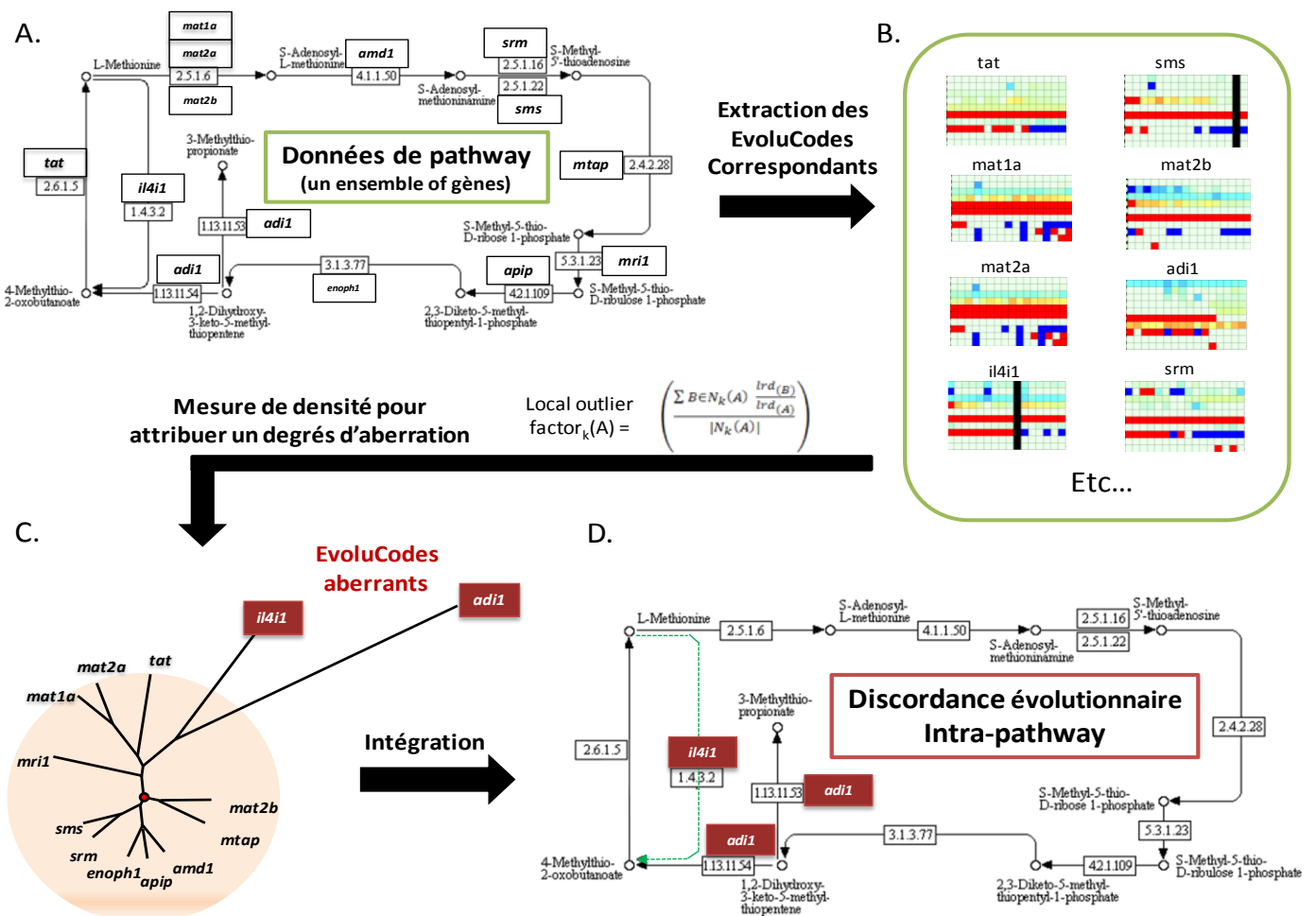
Figure 5. Méthodologie d'attribution d'un degrés d'aberration à l'histoire évolutive d'un gène dans le contexte de sa voie biologique. *A. Les données de voie biologique sont extraites de la base de donnée KEGG. B. Les EvoluCodes des gènes correspondant à la voie sont extraits et définissent le contexte évolutionnaire de cette voie. C. Une méthode d'extraction de connaissance, le Local Outlier Factor (LOF), est utilisé pour mettre en évidence la discordance évolutionnaires existant dans la voie biologique. D. L'analyse du contexte évolutionnaire permet d'extraire de nouveaux messages évolutifs.*

Une telle intégration multidimensionnelle s'est révélée puissante pour l'étude de l'évolution des réseaux cellulaires. Nous avons ainsi pu dégager plusieurs messages évolutifs liés aux systèmes humains. Premièrement, dans le cas des voies métaboliques, l'originalité des histoires évolutives est liée à la topologie du réseau. Par exemple, les gènes catalysant des réactions aboutissant à des composés impliqués dans de multiples voies, ou à l'inverse non métabolisables, se révèlent posséder des histoires évolutives très particulières relativement au reste de la voie. Nous nous sommes également intéressés aux gènes possédant un haut niveau de distribution et qui sont impliqués dans de nombreux processus cellulaires. L'histoire évolutive de tels gènes peut en effet être considérée comme typiques dans un processus mais atypique dans un autre. Leur référencement nous a permis de dresser une carte des processus cellulaires liés par le comportement évolutif différentiel de leurs gènes. Ce réseau représente notre premier aperçu des contraintes évolutives liées au maintien / à la modification des processus cellulaires, ainsi qu'un aperçu des recrutements de gènes qui se sont déroulés durant l'histoire des vertébrés entre ces processus. Comprendre ces contraintes est essentiel, car elles définissent le champ des possibilités d'évolution des processus cellulaires vertébrés.

Conclusion et perspectives

Les travaux décrits dans cette thèse représentent les premières étapes méthodologiques pour permettre une étude de l'évolution des processus biologiques à un niveau cellulaire. Nous avons développé un nouveau formalisme permettant de décrire de façon synthétique l'histoire évolutive d'un gène : les EvoluCodes. Le concept des EvoluCodes permet des analyses à grande échelle et l'utilisation d'outils formels d'extraction de connaissances. Nous avons ainsi généré les histoires évolutives pour l'ensemble du protéome humain à l'échelle des vertébrés. Contrairement aux études précédentes, construisant l'histoire évolutive de chaque famille protéique par la construction d'arbres phylogénétiques, nous avons créé un outil puissant, facilement exploitable et intégrable dans de nombreux projets. L'utilisation de techniques de classification super-paramagnétique et d'autres méthodes d'extraction de connaissances nous ont permis d'étudier l'histoire décrite par nos EvoluCodes dans le cadre des processus cellulaires humains. Nous avons ainsi mis en évidence les premiers indices relatant la façon dont les réseaux biologiques des vertébrés ont évolués et en particulier comment cette évolution a aboutit aux processus cellulaires humains. Cette étape clé est importante pour comprendre jusqu'à quel point la modularité des réseaux biologiques et leur transformation au cours du temps peut modifier un phénotype de façon spectaculaire ou par contre, ne pas produire de phénotype.

L'intégration de nombreuses données multidimensionnelles s'est révélée riche en message biologiques dès les premières analyses et nous a permis dans un premier temps de dégager des messages évolutifs globaux pour les processus biologiques humains. Tout d'abord, l'intégration d'autres types de donnés peut être imaginés. À l'avenir, un certain nombre d'améliorations est déjà prévu, tel que l'intégration d'autres types de données décrivant différents aspects des systèmes complexes. Par exemple, l'intégration de données décrivant l'expression des gènes pourrait être une étape clé pour comprendre le rôle du nombre de transcrits et des variations temporelles et/ou tissulaires dans le cadre de l'évolution des processus biologiques. D'ailleurs, de nouveaux EvoluCodes décrivant le protéome d'autres espèces permettraient également de comparer directement l'histoire évolutive des processus de différentes espèces.

L'approche EvoluCode et son intégration dans un contexte de biologie des systèmes est riche en potentiel. Au sein du laboratoire, les EvoluCodes ont été intégrés dans une base de données dédiée aux liens existant entre mutation et maladies génétiques humains. Le module des EvoluCodes aide à la priorisation de gènes pour

obtenir une liste de gènes candidats répondant à une question biologique précise. Au final, l'approche des EvoluCodes et leurs applications aux systèmes biologiques se révèlent un outil de choix pour mettre en évidence des messages évolutifs dans différents organismes et peuvent donc contribuer de manière significative à la biologie des systèmes et en particulier au domaine émergeant de la biologie évolutive des systèmes.

Thesis summary

Systems biology, an opportunity for evolutionary biology

Evolution is a principle inherent to life and a fundamental aspect of biology. Life is not fixed or stable, instead all species are slowly transformed over time. These transformations can be observed at the molecular level, with the selection of genetic mutations, the conservation/modification of 3D structures or the gain/loss of new genes or gene families. However, evolutionary principles are not only restricted to the molecular level, they also induce a constant remodelling at all biological levels.

Systems biology has developed enormously over the 10 last years, with studies covering diverse biological levels (molecule, network, tissue, organism, ecology…). In this context, the gene is no longer considered the sole element responsible for a phenotype. Our understanding of living systems has grown to incorporate the role of epigenetic phenomena, the role of intra- and inter- cellular processes, and even the modelling of tissue interactions. All these studies are regrouped in the field commonly known as systems biology, with the goal of deciphering and modelling biological phenomena at multiple biological levels.

To understand these phenomena, high-throughput technologies are now employed on an everyday basis, producing a deluge of "omics" data, and making bioinformatics an essential discipline in the post-genomic era. From an evolutionary point of view, systems biology provides unequalled opportunities. Evolutionary studies are no longer restricted to gene variation, and can be extended to the study of variations in cellular systems. This context recently gave rise to a new field, still in its infancy but full of possibilities, the so-called "evolutionary systems biology". The goal is to combine the detailed mechanistics of the recent systems biology with the more mature field of evolutionary biology in order to understand how organisms respond to perturbations. Indeed, what is evolutionarily conserved is generally essential for an organism, while variations represent potential innovations. Studying the evolutionary behaviour of a biological system is therefore the key to understanding how evolutionary pressures affect the individual components of biological processes and how variations induce potential transformations at the organism level.

Evolutionary systems biology

The first evolutionary systems biology studies emphasized several important evolutionary trends at the organism level. Inter-species comparisons have highlighted the plasticity of biological networks: the inactivation of a gene can be compensated for by other pathways and genes from one process can be recruited to another one in a given species. On the other hand, some interaction modules are conserved during evolution, in spite of genetic mutations. Entire processes of closely-related species can be inactivated during evolution and reactivated by mutagenesis or during environmental changes. In intra-species studies, high-throughput technological advances have allowed to observe a large diversity of transcriptomes associated with the same phenotype, in particular in the context of multigenic diseases. These examples reflect the new consensus that is now emerging from evolutionary systems biology: variation is experienced at all biological levels. Consequently, numerous studies have focused on these variations and have defined numerous new parameters to describe a system: propensity for gene loss, number of paralogs, expression levels, network centrality, etc. All these biological parameters are interdependent and demonstrate the multidimensional aspect of a complex biological system. Therefore, the evolutionary study of a

biological system requires integration of all these data from multiple species in a unifying framework, thus enabling the application of formal data mining and knowledge discovery methods.

Towards multidimensional integration in evolutionary systems biology

Since the discovery of the structure of DNA, the biological sciences have been gene-centric and have provided numerous answers to questions concerning the evolutionary mechanisms that induce the slow transformation of genes and their transmission. However, so far they have produced little information about other biological levels. During my thesis, I have developed new methodologies and tools to study the evolution of biological systems, taking into account the multidimensional properties of biological parameters associated with multiple levels (genome, proteome, network, phylum...). For the first time, formal knowledge discovery techniques have been applied at both the genomic and systems level to study evolution. Thus, this thesis addresses the clear need for novel methodologies specifically adapted to high-throughput evolutionary systems biology studies. By taking account the multi-level aspects of biological systems, we highlight new evolutionary trends associated with both intra and inter-process constraints. In particular, my thesis work includes (i) the development of an algorithm and a bioinformatics tool dedicated to comprehensive orthology inference and analysis for hundreds of species, (ii) the development of an original formalism for the integration of multi-scale variables allowing the synthetic representation of the evolutionary history of a given gene, (iii) the combination of this integrative tool with mathematical knowledge discovery approaches in order to highlight evolutionary perturbations in documented human biological systems (metabolic and signalling pathways...).

(i) The inference, visualisation and analysis of orthology relations

Orthology is a homology relation linking two genes that share the same ancestor and that result from a speciation event. This relation is essential in comparative and evolutionary genomics because it is assumed that two orthologous genes generally share a similar function. Thus, defining orthology is a key step in the comparison of molecular characters between species, where the characters can range from the single residue to the gene network. This motivated the development of OrthoInspector, a software suite dedicated to the inference of orthology relations between hundreds of genomes. OrthoInspector is based on a novel algorithm that uses the inparalogy relation as the basis for orthology inference. This approach is applicable to large-scale studies and maintains a good balance between sensitivity and specificity. In contrast to most existing methods, the OrthoInspector suite also provides numerous tools for the analysis and visualisation of complex orthology relations. The program thus represents a complementary approach to existing methods, responding to the growing need for integrative tools for the analysis of large-scale data in the post-genomic era, at the same time providing tools to summarize the data efficiently and informatively.

(ii) An original formalism for describing gene evolutionary histories

The second stage of my thesis involved the development of an original methodology to describe the evolutionary scenario leading to the current state of a gene. This work led to the concept of the EvoluCode (Evolutionary Barcode): a profile describing the evolutionary history of a gene at a given evolutionary scale. The EvoluCode is a synthetic representation for the integration, the visualisation

and the analysis of diverse parameters extracted from multiple biological levels (genomic, proteomic, network, etc.) and multiple organisms. The barcodes can be adapted to any kind of biological parameter and can be easily updated. Moreover, their matrix structure means that automatic comparisons can be easily performed with diverse metrics, thus facilitating their use in high-throughput studies. A key feature of the barcode formalism is the ability to describe the specific state of the parameters in their "evolutionary context". Thus, for each species, the state of a given parameter is defined as typical or atypical with respect to the generally observed state in the same species. The integration of multiple parameter types from different biological levels thus defines a combination of typical or atypical states, and allows to describe a complex evolutionary scenario that could not be resumed, for example, with a simple sequence analysis.

We applied the EvoluCode methodology to the human proteome by reconstructing the evolutionary histories for all genes at the vertebrate level. To achieve this, we combined parameters from several biological levels such as the genome context, the organization and conservation of protein domains or phylogenetic distributions extracted from the OrthoInspector predictions. The integration of these data in the EvoluCode structure and their study at the complete proteome level revealed a number of interesting conclusions. First, we mapped our EvoluCodes onto the human genome and with a dedicated visualization tool we observed clusters of similar evolutionary histories, most of them corresponding to known gene clusters. Second, we exploited our formalism in a knowledge discovery protocol based on a super-paramagnetic clustering algorithm, to group similar barcodes and study their functional enrichment. This approach highlighted several evolutionary trends linked to groups of proteins with similar functions or cellular localizations. This second conclusion illustrates the potential of studying gene evolutionary histories, not only as independent objects, but also in the context of their biological processes.

(iii)   Multidimensional study of the human gene networks

The encouraging results provided by the study of the EvoluCodes associated with the human proteome motivated us to explore the evolutionary history of all human biological processes. Consequently, we performed a topological mapping of our EvoluCodes on documented human networks. Here, we no longer restricted our study to the evolutionary history of a single gene, we wanted to explore the history of a biological process, defined by a gene network and corresponding to a particular biological response. We created a protocol that incorporates an anomaly detection algorithm to identify genes with an unusual evolutionary history compared to the other genes implicated in the process. This gene state (typical/atypical evolutionary history) is defined based on our EvoluCodes and thus integrates their multi-scale biological parameters in the evolutionary analysis of the human cellular networks. Using this powerful integrative approach, we were able to identify several evolutionary trends characterizing the human systems. First, in the case of metabolic pathways, evolutionary perturbations are linked to the network topology. For example, genes catalyzing metabolic reactions that produce either compounds implicated in multiple pathways or non-metabolised compounds have very specific evolutionary histories compared to others in the same pathway. Second, we studied a set of widely distributed genes, i.e. genes implicated in multiple cellular processes. The evolutionary histories of such genes can be considered as typical in one biological process but atypical in another one. We used this gene set to create a map of human cellular processes linked by the differential evolutionary behaviour of their shared genes. This map is the first overview of the evolutionary constraints governing the modification/maintenance of cellular

processes, as well as an insight into the inter-process gene recruitments that operated during vertebrate evolution. Understanding these constraints is essential since they define the range of possibilities for the evolution of vertebrate cellular processes.

Conclusion and perspectives

The work described in this thesis represents the first methodological steps towards the study of the evolution of biological processes at the cellular scale. We have introduced a new formalism, the EvoluCode, to describe the evolutionary history of a gene in a synthetic manner. The EvoluCode concept facilitates high-throughput analyses and the use of formal knowledge extraction tools. We generated evolutionary histories for the complete human proteome at the vertebrate scale. In contrast to previous studies, which mainly described evolutionary histories of protein families by constructing phylogenetic trees, we have created a powerful tool that is easily exploitable and can be integrated in numerous projects. The use of super-paramagnetic clustering and other knowledge extraction techniques, allowed us to study the evolutionary histories described by the EvoluCodes in the context of human biological processes. This is a first step towards a better understanding of how these networks were modelled during evolution and how this evolution led to the current human biological processes. This knowledge should contribute to a better understanding of the modularity of biological networks and how their transformation over time can affect the final phenotype.

The integration of multidimensional data performed here has allowed us to highlight several global evolutionary trends in the context of human pathways. In the future, a number of enhancements are already envisaged, such as the integration of other data types describing different aspects of complex systems. For example, the integration of gene expression data will provide insight into the role of the number of transcripts or temporal/tissue transcriptional variations in the context of the evolution of cellular processes. Another potential enhancement would the generation of EvoluCodes corresponding to the complete proteomes of other species , which would allow us to directly compare evolutionary histories between species.

The EvoluCode approach has many potential applications in systems biology studies. In the laboratory, the human EvoluCodes have already been exploited in the KD4v database dedicated to the study of the relationships between genetic mutations and human genetic diseases. The EvoluCode module of the database is used to facilitate gene prioritization, i.e. the selection of the best gene candidates relevant to a specific biological question. This application illustrates the power of the EvoluCodes and their potential contribution to the emerging fields of systems biology and evolutionary systems biology.

THESIS TAG CLOUD

A dedication to Olivier and Julie, for all the time spent looking at original visualization tools. Image generated with the content of this thesis in the website http://www.wordle.net with the content of this manuscript.

SUMMARY

LIST OF FIGURES

LIST OF TABLES

ABREVIATIONS

AED   Asymmetric Evolution after Duplication

BIPS   BIoinformatique Platform of Strasbourg

BLAST   Basic Local Alignment Search Tool

DNA   DesoxyriboNucleic Acid

EvoluCode   Evolutionary barCodes

GO   Gene Ontology

GRN   Gene Regulatory Network

KEGG   Kyoto Encyclopaedia of Gene and Genomes

LGT/HGT   Lateral/Horizontal Gene Transfer

MACS   Multiple Alignment of Complete Sequences

MACSIMS   Multiple Alignment of Complete Sequences Information Management System

ncRNA   non-coding RiboNucleic Acid

PPI   Protein-Protein Interaction

RNA   RiboNucleic Acid

SAGE   Serial Analysis of Gene Expression

TCA   TriCarbocylic Acid cycle

WGD   Whole Genome Duplication

XML   eXtended Markup Language

1 EVOLUTION: AN ESSENTIAL PRINCIPLE FOR UNDERSTANDING LIFE

1.1.1 Why study evolution?

"In the whole history of thoughts no transformation in men's attitude to Nature – in their 'common sense' – has been more profound than the change in perspective brought about the discovery of the past" (Toulmin and Goodfield, 1965)

Since the beginning of written history, the 'question of our origins' has been a central theme in all cultures. For thousands of years, metaphysical beliefs and religions provided the only answers. It is only in the last two centuries that humanity has experienced one of its most important revolutions: a new conception of our origins supported by scientific reasoning and called 'Evolution'. This idea is young in the history of man and remains fragile, as testified by its many opponents. It is even more fragile since it conflicts with some of our more natural beliefs, such as anthropocentrism, essentialism or fatalism. The theory of Evolution implies that our natural attitude to explain things by "meaning" and "purpose" are merely human projections, simple expressions of the values we choose to give to our own existence. Such a revolution in human thinking is a tough task. The main challenge for the concept of Evolution may not be to explain the origins of life, but to withstand the species that created it.

The evolutionary revolution has had a huge impact on science, particularly in the biological domain. By definition, biology focuses on living beings and in principle, all living beings are the result of millions of years of evolution. It is interesting to note that telling the history of life through evolution is a reverse process and what we observe is the end of the history; all the organisms currently living on Earth. Evolutionary principles underlie all biological studies, because they have fashioned the systems implicated in biological mechanisms at all levels, from the molecular to the ecological level. They are not even restricted to biology, since the current enrichment of oxygen in our atmosphere or the constant impact of Life on the Earth's geology also result from Evolution. Thus, the study of Evolution increases our understanding of life and its environment.

1.2 Birth of the theory of Evolution

1.2.1 Classification and essentialism

"L'histoire n'est que l'évolution de l'idée de Dieu dans l'humanité."
(Alphonse Esquiros, 1814-1876, Les martyrs de la liberté)

The idea that living beings can be compared in some way originated in the ancient world with the first classifications. Theophrastus (Θεόφραστος, 371–287 BC) described the first taxonomy blueprint of the plant kingdom in '*Enquiry into Plants',* a book classifying plants by their modes of generation, geographical localization, size or use. This work was partially reused by a naturalist of the Latin world,

Pliny the Elder (Gaius Plinius Secundus 23–79 AD), in his books 'Historia naturalis'. These books covered several fields such as anthropology, botany, zoology, mineralogy or pharmacology, and extended the classifications from plants to animals.

These ancient works were relatively untouched during the Middle Ages, with only sporadic additions by middle age botanists and zoologists. However, the consecutive addition of comments and new figures showed the limits of the proposed classifications. Finally, during the 16th century, three botanists - Fuchs (1543), Gesner (1541) and Camerarius (1586) – proposed a new classification based on alphabetical order of plant names! Despite its limitations, this decision broke 1500 years of an untouched classification. During the 16th and 17th centuries numerous classifications were developed, based on criteria such as size, leaf/root shapes… In 1694, Joseph Pitton de Tournefort (1656-1708) was the first to understand that species can be reunited into genus, introducing the concept of hierarchical levels. This hierarchy was then extended by Karl von Linné (Carl Nilsson Linnæus, 1707-1778), who introduced the basis for the current traditional ranks: *regna, classes, ordines, genera, species, synonymis locis*. He also introduced the binomial nomenclature (genus species) in his major work '*Systema Naturae'*, a book classifying thousands of animals and plants.

We note that these major advances in classification were achieved within a scientific community that was difficult to separate from theologians. Essentialism was predominant among botanists and zoologists, such as Karl von Linné, who believed in a world created by God and organized according to his will. Consequently, from this point of view, all species are seen as they were originally created. Variations within species or malformed animals were considered divine amusements or "Monsters". This conception was first debated by paleontologists such as Georges Cuvier (1769-1832) or William Smith (1769- 1839). According to them, fossil species had totally disappeared and could not be linked to current species. This idea gave birth to the concept of 'catastrophism', which hypothesized that major natural disasters were regularly responsible for the extinction of several species and were followed by migrations of surviving populations. This theory was not accepted by naturalists who observed that mammals were absent from older geological layers. Thus, some of them reformed this hypothesis by arguing for a continuous creation of new species following major extinctions. Despite the idea that species could appear or disappear with time, the essentialist theory remained during all the first half of the 18th century.

1.2.2 Birth of transformism

"*Nevertheless, it is even harder for the average ape to believe that he has descended from man.*" (Henry Louis Mencken / 1880-1956)

During the second half of the 18th century, several botanists suggested that species might change by transmutation. Jean-Baptiste Pierre Antoine de Monet (1744-1829), better known as Lamarck, defended the idea that ancient species did not disappear and that a continuity existed between fossil and current forms of life. He supported his idea by introducing physical factors, responsible for a general progress and diversification of life with time. More precisely, he hypothesized that species transformed with time by the use or uselessness of their organs. This represented the birth of the 'transformist' movement. Unfortunately, Lamarck missed the importance of intrinsic changes of the

species and gave no credit to intra-species variations for organism transformations. This is the fundamental difference between Lamarck and Darwin. Following Lamarck, the transformist theory progressively extended through Italy, France and England, but also inspired many opponents defending the essentialist idea. This was particularly true in England during the first decade of the 19th century, with the work of the theologist William Paley (1743-1805) and his book 'Natural theology'. According to him, God created a providential perfect world and the laws of Nature - *expression of its perfection-* had to be deciphered by humans. Paley thus integrated science in theology, with the goal of proving God's omniscience. It was in this intellectual environment that Charles Darwin performed his studies.

1.2.3   Charles Darwin: the father of the theory of Evolution

"Those whom we called brutes had their revenge when Darwin shewed us that they are our cousins." (George Bernard Shaw, 1856-1950)

In 1859, Charles Darwin (1809-1882) published '*The Origin of Species by the means of natural selection*'. He developed a transformist point of view, but gave more credit to intra-population variations. The second hypothesis that he introduced was the process of natural selection and it was only in the 6th publication of his theory that he introduced the word 'Evolution'. His arguments for the theory of Evolution can be resumed in 5 observations:

1. There are variations (physical characteristics, capabilities) among sexually compatible individuals. Independently from the source of this variation, there is a natural variability inside what we designate as species.
2. Humans can artificially select and model species for their needs. Consequently, there is a natural capacity for species to be selected. This implies a second notion: variations can be inherited, allowing artificial selection.
3. Species can reproduce if they find necessary resources. When the reproduction rate reaches some limit, resources are exhausted or other factors such as predators will limit the size of population. Consequently, there is a natural capacity for overpopulation.
4. A wild environment is populated with multiple species, despite the overpopulation capacity of species. Thus, there are natural equilibriums. Each species is limited by the extension of other species. Species can be selection agents.
5. The success of reproduction for a species depends on physical and chemical optimums (temperature, humidity, pH, odorant molecules…). These factors are a second selection agent.

These observations lead to the evolutionary hypothesis of Darwin (figure 1-1). Individuals with advantageous variations relative to the physical, chemical and biological environment will produce more individuals in the following generation. If these conditions are maintained, the frequency of the advantageous variations will expand to the entire population. The whole species will change slightly over time, i.e. it is not stable. If the environmental conditions change, new variants will be positively selected. This phenomenon was named 'natural selection' by Darwin and implied a differential reproductive success.
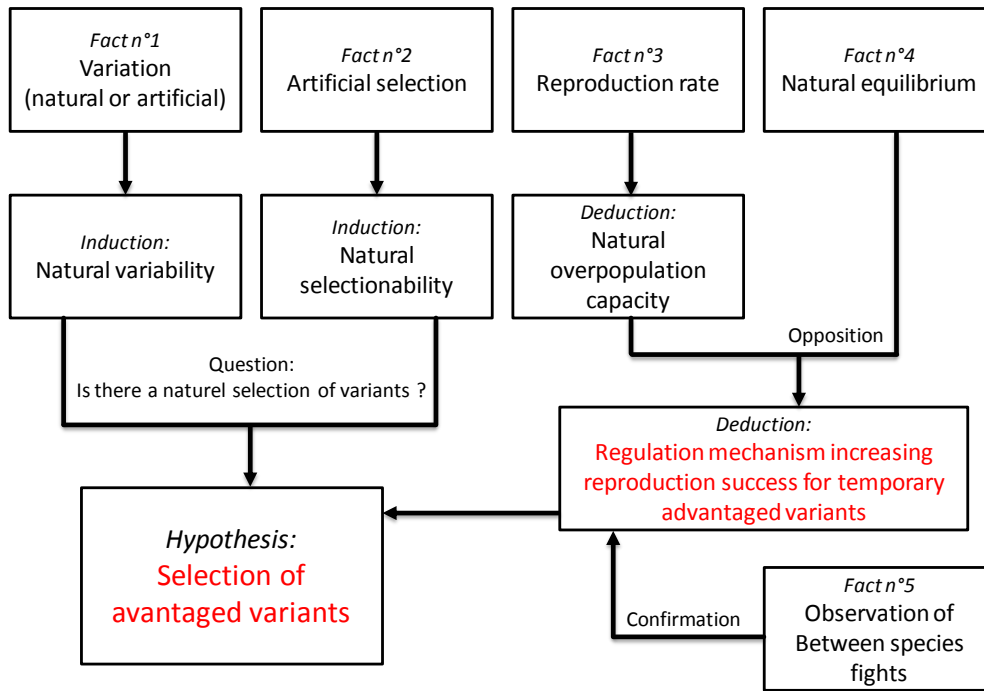
Figure 1-1. The reasoning proposed by Charles Darwin in 'The origin of species'.
Adapted from (Patrick, 2000)

During the 70 years following the publication of 'The origin of species', Darwinian theory was progressively accepted in scientific communities and was extended with new ideas. Darwin wrote two other major books in 1871: '*The descent of Man'* and '*Selection in relation to sex*', in which he extended the transformist theory to the human species with the proposal that humans are rooted in the Tree of Life with catarhinian monkeys. Later, he extended the theory of natural selection from organic variations to instincts and social behaviours.

1.3   Modern evolutionary synthesis

1.3.1   Neodarwinian evolutions

From 1860 to 1930, several key scientists greatly contributed to the extension of Darwin's theory. Gregor Johann Mendel (1822-1884) performed the first agricultural experiment of heredity by crossing different varieties of beans and published the three fundamental laws of heredity in 1866. However, his results were mainly forgotten by the scientific community. At the beginning of the 20th century, the first genetic studies appeared, initiating new debates between naturalists and geneticists. At the same time, Mendelean laws were re-discovered independently in Germany, Austria and the Netherlands. In 1902, Walter Sutton (1877-1916) described the chromosome pairs as a potential physical basis for Mendelean laws. A few years later in 1909, he introduced the notions of 'gene' and 'mutation'. These discoveries created a new philosophical thinking in the geneticist community: 'saltationism'. For geneticists, a 'discontinuous' variation was the source of Evolution:

genetic mutations induce speciation jumps and natural selection plays a minor role. On the other hand, the naturalist community continued to defend the natural selection hypothesis.

These contradictory ideas were finally reunited with the experimental work of Thomas Hunt Morgan (1866-1945). His genetic studies on *Drosophilia melanogaster* demonstrated that most mutations have a limited effect and induce a gradual transformation of the population. In the 1930's, the population genetics field appeared, definitively reconciliating geneticists and naturalists. The first neodarwinian synthesis appeared: mutations with limited impact appear randomly in populations and the modification of the mutation frequency in the population initiates speciation.

1.3.2   Development of the modern evolutionary synthesis

"Nothing in Biology Makes Sense Except in the Light of Evolution."
(Theodosius Dobzhansky, 1900–1975)

The hundredth anniversary of Darwin's theory provided the occasion to reunite all evolutionary knowledge into a consensus theory called the *'Synthetic evolutionary theory'*. From 1900 to the 1970's, numerous experimental works confirmed many aspects of this theory. We can cite the work of Bernard Kettlewell (1907-1978) who directly observed natural selection phenomena. He confirmed that dark populations of the butterfly *Biston betularia* were positively selected in a polluted environment due to lower bird predation. During the 1940's, Maxime Lamotte (1920-2007) and Gustave Malécot (1911-1998) confirmed the role of random fluctuations of alleles in populations by studying populations of *Cepea nemoralis* snails. The smaller the population is, the higher the probability to randomly fix an allele in this population, without any natural selection intervention. In the 1950's, Philippe l'Héritier and Georges Tessier confirmed many aspects of the Synthetic theory through several studies of *Drosophilia melanogaster,* in particular:

- Selection creates novelty: when the environment is unstable, natural selection extends genotype variations and composition. The opposite is observed in a stable environment where a conservative selection is observed.
- Allele selection depends on the genetic context: one gene can be advantageous depending on the other loci that are linked to it. Interestingly, this is the first systemic view introduced in evolutionary theory, but during the following decades these interactions were restricted to gene level studies.
- Selection is frequency dependant: individuals exploiting a resource not used by the majority of the population are positively selected and escape competition. Consequently, rare genotypes are positively selected until their frequency grows because competition grows at the same time.

In 1962, V.C. Wynne-Edwards (1906-1977) introduced the notion of 'group selection' to explain the altruism observed in animal species. Finally, mathematical tools derived from game theory were applied to genes by John Maynard-Smith (1920-2004) and George R. Price (1922-1975), launching the debate on levels of selections. All these studies contributed to a more complex and refined theory of evolution, the so-called *'Synthetic evolutionary theory'*. However, during the following fifty years,

most biological fields that elaborated this theory (anatomy, morphology, zoology, botany…) slowly declined. This fact can be partially explained by the frenzied emergence of molecular biology, accompanied by an all-powerful entity: the gene.

1.3.3 Expansion of the theory of Evolution in gene-centric biology

"Biology will relate every human gene to the genes of other animals and bacteria, to this great chain of being." (Walter Gilbert)

1.3.3.1 The Big Bang of molecular biology

The second half of the 20th century was a revolution for evolutionary theories. The discovery of the structure of DNA by James D. Watson and Francis Crick (1916-2004) initiated in 1953 the beginning of the molecular biology era. This led to the development of biology focused on the gene and the genome, providing a common basis for comparisons of all life forms. In the evolutionary field, the first significant results were produced in 1977 by Carl Woese and Gary J. Olsen with the comparison of ribosomal genomes in prokaryotes, differentiating the Bacteria from the Archea. In 1977, Sanger & al. introduced the well known DNA sequencing technique (Sanger and Coulson, 1975). In 1983, Kary Mullis developed the PCR technique that would later become a routine technique for molecular analysis (Bartlett and Stirling, 2003). These multiple technical advances finally allowed complete genome sequencing. The very first complete genome was sequenced in 1976 by Walter Fiers with the publication of the complete nucleotide sequence of the bacteriophage MS2 (Fiers et al., 1976). In 1995, *Haemophilus influenzae*, was the first completely sequenced bacterial genome (Fleischmann et al., 1995). In 1996, the 16 chromosomes of the eukaryote *Saccharomyces cerevisiae* were sequenced (Goffeau et al., 1996). In 2001 the first draft genome of our own species was published (McPherson et al., 2001; Venter et al., 2001). Since the nineties, hundreds of genomes have been fully sequenced, providing incredible new opportunities for evolutionary studies and more remodelling of phylogenies in less than 30 years than 2000 years of classifications! Molecular innovation descriptions have complemented the physiognomic or physiologic innovations that were described during the 18th and 19th centuries. Phylogenetic studies confirmed many genetic mechanisms of genome evolution. Many cross-validations were performed between phylogenetic, ecological and geologic studies, giving new insights into the common histories of Earth and Life. This booming of the molecular biology field thus gave rise to several new fields in the evolutionary domain.

1.3.3.2 Phylogenetics and comparative genomics

During the last decades, evolutionary studies have progressively shifted to a molecular description of Evolution, with DNA and protein sequences being the most studied biological entities. In 1927, Motoo Kimura (1924-1994) already observed the large number of enzymatic polymorphisms inherent to the same species. He proposed that modifications of macromolecules are "selectively neutral" and that modified genes can be fixed in the population if the mutation does not affect the global structure of the protein. For Kimura, what natural selection can "see" is not the sequence itself but the shape and the function of the molecule. This was the first intuition of the importance of the link between sequence/structure/function. When DNA and protein sequences became accessible

for most laboratories, the gene 'function' became a central subject and many biologists believed that understanding the function of all genes would be the key to understanding life. The evolutionist community made a link between the conservation/modification of molecular function and the process of natural selection, giving rise to the idea of functional adaptations. They understood that comparing DNA sequences between organisms is the key to understanding the functional adaptations that were selected during Evolution. This idea was the basis of phylogenetics. In 1937, Dobzhansky constructed one of the first molecular phylogenies by comparing chromosome rearrangements in 17 Drosophila Pseudo-obscura strains. In 1964, one of the first human DNA phylogenetic trees was published, highlighting the main migrations that human populations followed during prehistoric times (Edwards and Cavalli-sforza, 1964). Today, gene variation is used to decipher the molecular mechanisms of evolution and phylogenetic trees are an everyday tool for evolutionary biologists. Interestingly, this period was fruitful for interdisciplinary cross-validations of models of evolution. The model of tectonic plates, despite being hypothesized Alfred Wegener (1880-1930), was finally confirmed during the sixties. Animal phylogenies, fossils and models of continental surface fragmentations together explained the divergence of continental ecosystems and the appearance of endemic species. This kind of study now corresponds to the specific field of 'biogeography' (Springer et al., 2011).

Simultaneously with the expansion of gene related studies, new advances in physics during the fifties opened up a completely new view of biological systems. Several authors described the structural conformation of proteins (Edsall, 1956; Mizushima et al., 1949; Ramachandran et al., 1963). Rapidly, the relation between protein structure and function was recognized and the role of molecular structures was later integrated in evolutionary models (Goldstein, 2008). The evolution of the modular organization of proteins in 3D domains and the link between residue mutation and 3D structure modification were major breakouts in molecular evolutionary theory (Liberles et al., 2012). The role of key residues in catalytic sites or protein interfaces was discovered (Worth et al., 2009). Later, the identification of structured catalytic RNA molecules was another striking result. It initiated a new hypothesis for the origin of life in which pre-biotic life emerged in a RNA world, supporting both genetic code and biological functions (Melendez-Hevia, 2009). This hypothesis was complemented by the fact that most protein folds are found in bacteria, archaea and eukaryotes, hypothesizing a Last Universal Common Ancestor possessing a complex protein repository (Abeln and Deane, 2005).

All these revolutions were accompanied by an increasing amount of gene and protein data for many species. Phylogenetic approaches were powerful but mainly restricted to a single protein family. With the help of the emerging bioinformatics sciences, performing a multi-species comparison of genomic data provided an opportunity to study genome evolution at a larger scale. This was the birth of comparative genomics, a field focusing on the relationship of genome structure and function across different biological species or strains. Such approaches exploit both similarities and differences in biological sequences to understand the evolutionary mechanisms that modeled genomes. The comparison of multiple organisms can highlight similarities conserved through time, as well as divergent elements. Contrary to phylogenies, comparative genomics is not restricted to similar sequences and is particularly useful for elucidating the functional and evolutionary aspects of biological systems (Hardison, 2003). For example, the genome-scale detection of homologous

relationships is used for the functional annotation of new genomes (homology and orthology are discussed in chapters 2 and 3). Functional regions of chromosomes can be detected by phylogenetic shadowing (Boffelli et al., 2003) and conserved DNA regulatory elements can by highlighted by phylogenetic footprinting (Aerts, 2012; Zhang and Gerstein, 2003). The structures of genes, in particular the differential conservation of intron/exon sequences, was also highlighted (Lander, 2011). Today, many properties of genomes continue to be discovered by comparative genomics approaches. For example, the role of ultra-conserved non-coding regions in the human genome is beginning to be understood (Pollard et al., 2006).

1.3.3.3   Emergence of the Evo-Devo field

Interestingly, of molecular evolution concepts are central not only in molecular biology but also in developmental biology. This fact is a heritage of the anatomy and comparative embryology fields that compared body structures between animal embryos and slowly declined during the 19th century. The molecular knowledge of evolution is now giving a second life to this approach. The field of evolutionary developmental biology, named 'Evo-Devo', combines data concerning the genetic control of development with data from experimental and evolutionary comparative embryology. Its origin comes from the particular type of mutations observed in 1894 by William Bateson. He described 'homeotic' mutations in insects: mutations changing the position of appendix pairs along the body plan. Later in 1978, Edward B. Lewis (1918-2004) published a model of the evolution and function of several clusters of genes governing embryologic development: the bitorax complex (Dang et al., 1998). The Evo-Devo field now focuses on two main research axes:

- The identification of the genes controlling embryogenesis and their functions.
- The analysis of the repartition of these genes in the metazoan lineage by comparing their sequence and their expression. Effectively, genes controlling development have particularly conserved chromosomal localizations. Such analyses could lead to a new interpretation of organ homologies, parallelism and convergence between animals.

Such approaches could highlight the importance of developmental constraints in evolution and explain the main morphological differences existing between zoological groups. In particular, some morphological characters can be conserved in animal phyla because of developmental constraints, despite a lack of apparent usefulness. One striking example is the origin of hiccups (Straus et al., 2003). In mammals, the phrenic nerve describes a complex trajectory from the bottom of the skull to the diaphragm. This complexity can cause an inflammation of the nerve, inducing hiccups. This anatomical configuration is a legacy from the bony fishes and the functional aspect of hiccups, despite losing its meaning for mammals, can be found in early amphibians that gulp air and water across their gills via a motor reflex. Moreover, the same authors observed that the cellular pathways enabling hiccupping are activated prior to cellular pathways enabling normal lung ventilation during foetal development.

1.3.3.4   (re)-discovering phenotypic variation

The gene-centric biology developed during the second half of the 20th century was highly motivated by biotechnological and biomedical research mainly focused on the mechanistic aspects of genes and proteins. Thus, the biology of the last 50 years was mainly focused on the gene 'entity', an immutable object that is present/absent, activated/inhibited in a cell and shared or not shared by species. It is

interesting to note how the variation intrinsic to any gene has often been ignored by molecular biologists who have focused more on the functional features and mechanistic aspects of biological processes. Consequently, Evolutionary mechanisms are not taken into account in many molecular biology studies, despite being the origin of this intrinsic variation.

In the post-genomic era, new genetic and systemic knowledge is providing an opportunity to finally explain phenotypic variations at every biological level, from the genetic mutation to biological network dynamics. Today, with the democratization of high-throughput sequencing, the increasing role of epigenetic regulation, the role of ncRNA or the understanding of network dynamics, the gene is slowly losing its all-powerful role in biology. Genotype is no longer directly linked to phenotype and the attention given to intermediate biological levels is increasing. In particular, biology is reconsidering the importance of variations and their impact on function at all levels. Interestingly, it is biomedical research - one of the main motors of the functional view- that is motivating this new transition by providing new data about intra-population variation in humans. Biomedical research is also linking molecular studies and phenotypes, through new ideas such as personalized medicine (Neal and Kerckhoffs, 2010) or high-throughput sequencing as a routine diagnostic tool (Ku et al., 2012). In this context, re-introducing evolutionary concepts at the system level is a great opportunity to study the importance of variations and their dynamics. By integrating evolutionary data from numerous biological levels (genomic context, protein level, interactions, pathways, signaling…), evolutionary systems biology can study numerous questions linked to specific phenotypes or syndromes. Finally, through evolutionary systems biology, evolutionary genomics is returning to the analysis of phenotype variation, a view that was partially overshadowed by the analysis of gene variation.

2   DEFINING HOMOLOGY, THE BASIS FOR EVOLUTIONARY STUDIES

The evolutionary sciences have 150 years of history, evolving from a direct observation of Nature to the concept of genetic inheritance, followed by the molecular description of genetic mechanisms. Today, the theory of evolution is complemented by the mechanisms that model biological processes at the system level. A common factor in all evolutionary studies is that they compare multiple organisms based on a shared character. This shared character is sometimes referred to as a homology. Evolutionary studies first considered body shapes, physiology or life environments, but the molecular biology revolution has now placed the gene and the genome as the reference characters. Effectively, DNA is a common character of all Life forms, giving hope for the establishment of an evolutionary history of Life since its beginnings.

2.1   Similarity and homology

Etymologically, the word 'homolog' derives from the Greek word '*homologos'*, meaning 'equivalent relation'. Use of homology to describe life is as old as comparative anatomy or paleontology, but its definition has changed over time. In 1843, Richard Owen coined the term 'homology' as 'the same organ in different animals under every variety of form and function'. Owen noticed the similarities between certain structures in different organisms and imagined an ideal plan describing similarity structure among groups of animals. Its evolutionary meaning began to be accepted in the scientific community during the same period. However, because the term homology was coined in the pre-evolution era, its meaning can be ambiguous and today most zoologists prefer the term 'synapomorphy'. Moreover, in an evolutionary framework, similarity in structure can be due to either from common ancestry or from convergent evolution. Convergence results from similar evolutionary pressures and constraints that induce similar structures in order to perform similar functions. Birds' and bats' wings are an example of convergence (figure 2-1). To test whether a characteristic is homologous or the result of convergence, the scientific reasoning is the following:

- Assume homology and make a prediction based on that assumption (homology is a bet).
- Search for sources of evidence that support the assumption or refute it (repeated tests confirming homology increase its confidence).

The notion of homology was transferred from the organism level, based on anatomical characters such as number of legs or organ shapes, to the molecular level during the second half of the 20th century. It is now well established in all scientific communities that homolog refers to a character deriving from a common evolutionary ancestor. In the post-genomic era, homology is mainly used in comparative anatomy, evo-devo and molecular biology. There is no degree of homology: characters are either homologous or not (Tautz, 1998).
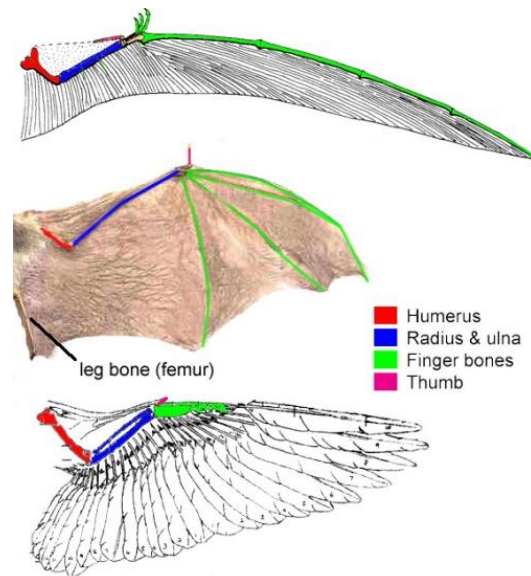
Figure 2-1. Differentiating homology and convergence. *Pterosaurs, bats and birds produced wings with functionally similar shapes from the forelimb. The bones in each wing are homologous, but the arrangement of bones within the wing and the wing itself appeared independently in each group (convergence). Image by J. Rosenau.*

2.2 Homology in molecular biology

In molecular biology, homology -a qualitative denomination- is mainly inferred by estimating a quantitative similarity, usually based on sequence residue identities. Indeed, as homologous sequences derive from a common ancestor, key residues and sequence motifs are often conserved during evolution (generally linked to function). Consequently, sequence similarities are used to describe the relatedness of sequences and infer homology. However, two sequences can be homologous without sharing significant residue similarity. It is important to note that sequence homology must be applied to the sequence itself and not to higher biological concepts such as genes. A gene could result from the fusion of several genes or the addition of new domains (portions of other genes). Following strictly the homology definition, such a gene is homolog to both genes sharing the same similar regions. This particular case highlights the fact that homology is not a transitive definition: the fused gene is homologous to two other sequences, but these latter sequences are not homologous.

At the molecular level, homology was first used to describe similar functional phenomena. For example, myoglobin and hemoglobin were first described as 'homologs' based on their similar chemistry (Kendrew, 1961). This idea still remains in several biological fields, ignoring the evolutionary definition of homology and focusing on the functional implications. One of the pioneers of DNA and protein comparison in an evolutionary framework, W.M. Fitch, introduced the idea that nucleotide replacements account for the divergent descent of a set of genes, given a particular topology for the tree depicting their ancestral relations (Fitch, 1970). Since then, similarity measurement between biological sequences has become a standard means of establishing homology. Comparative analyses, functional annotation or evolutionary studies require a transfer of

information between organisms and homology is one of the most popular concepts used to address this problem. Beginning with small sets of genes, these analyses grew to large-scale analysis of complete genomes or meta-genomes. The increasing use of comparative studies based on homology highlighted the need for more specialized homology definitions, including orthology or paralogy.

2.2.1 Orthology/paralogy

Orthologs are homologous genes that diverged from a single ancestral gene in their most recent common ancestor via a speciation event. Paralogs are homologs resulting from gene duplications. The distinction between orthologs and paralogs refers exclusively to the evolutionary history of genes and does not have functional implications *stricto sensu* (Peterson et al., 2009). However, from an operational point of view, it is widely accepted that two orthologs generally share the same function (Brown and Sjolander, 2006). This hypothesis is supported by domain architecture conservation analysis showing that function conservation between orthologs demands higher domain architecture conservation than other types of homologs, relative to primary sequence conservation (Forslund et al., 2011). In contrast, it is generally considered that paralogs can diverge more rapidly and new functions can emerge as the result of mutations or domain recombinations. The most frequent outcome after a gene duplication is that one of the paralogous genes becomes a pseudogene. This phenomenon is known as nonfunctionalization (Lynch and Force, 2000; Maere et al., 2005). The sequence of a pseudogene degrades over time by including more and more mutations, until the gene is no longer recognizable by sequence similarity searches. Alternative fates for duplicated genes include a positive selection for both paralogous copies (neofunctionalization) or the specialization of one gene copy compared to the ancestral gene role (subfunctionalization) (Taylor and Raes, 2004).

2.2.2 Inparalogy/Outparalogy

The multiplication of available genomes in the post-genomic era has highlighted the necessity to distinguish two subtypes of paralogs: inparalogs and outparalogs (Koonin, 2005). Inparalogs are produced by duplication subsequent to a given speciation event, while outparalogs result from an ancestral duplication (relative to the given speciation event). To resume these relations, 'in-paralogy' and 'out-paralogy' are concepts relative to the species under comparison (figure 2-2). The distinction is crucial in evolutionary studies since sets of inparalogs derive from orthologs by lineage-specific expansions and thus can be considered to be co-orthologs, while outparalogs do not have orthologous relationships at all.
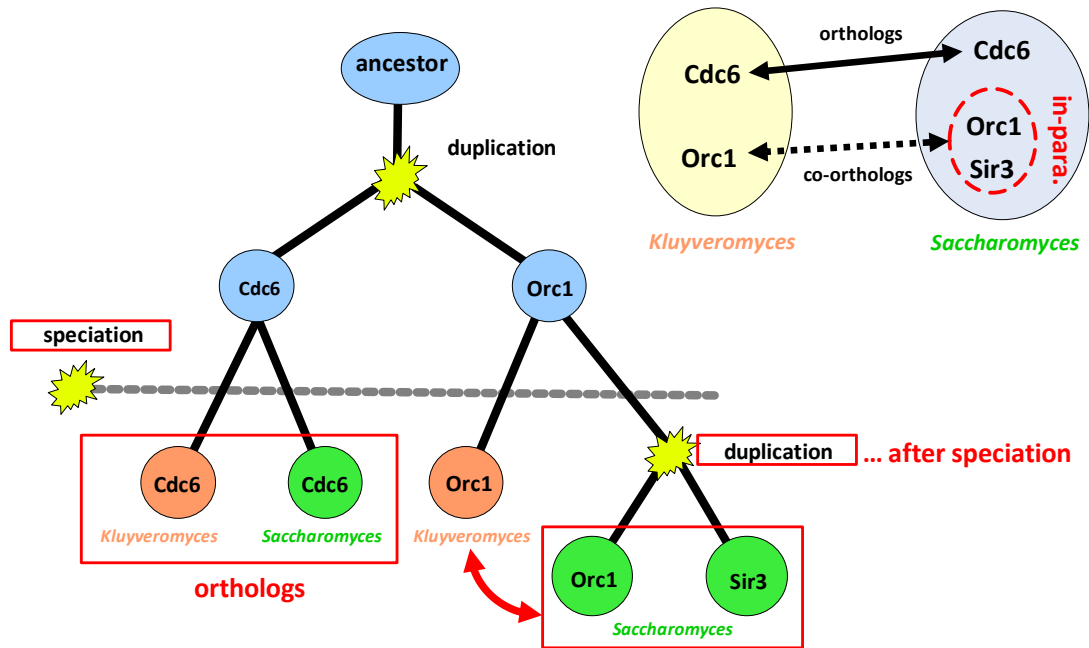
Figure 2-2. Schematic representation of an inparalogy relation. *Inparalogs are defined relative to a duplication event that occurred after a speciation event. Data from Manolis K., 2004.*

2.2.3 Xenology

Xenologs are homologous sequences found in different species because of lateral gene transfer (LGT, also called horizontal transfer) instead of speciation. The term was introduced by E.V. Koonin in the context of prokaryotic LGT studies (Koonin et al., 2001). During the last decade, numerous cases of inter-prokaryotic xenology have been described, identifying LGT as a major contribution to prokaryotic evolution (Boucher et al., 2003; Ochman et al., 2000). Some rare cases of xenologs between mammals and bacteria (Goulas et al., 2011) or endosymbiont and host (Timmis et al., 2004) have been described, introducing a LGT role even in eukaryotic evolution (Andersson, 2005; Ros and Hurst, 2009).

2.2.4 Functional aspects of orthology/paralogy

The putative relation between function and orthology/paralogy is still actively debated in the Evolution field. However, it is generally accepted as proven in the molecular biologist community, which is an error considering that orthology has a strict evolutionary definition. In 2011, Gharid & Robinson-Rechavi reviewed literature cases of orthologs with functional divergence between mouse and human (Gharib and Robinson-Rechavi, 2011). They highlighted the lack of systematic exploration of functional differences between orthologs and the fact that this kind of research appeared only

recently in the scientific literature. Nevertheless, they estimated that a divergence of gene expression, alternative splicing or mutant phenotypes each affected about 10–20% of ortholog pairs. These human-mouse orthologs with strong differences would clearly affect many pathways and biological processes of interest (Gharib and Robinson-Rechavi, 2011). Another recent study by Nehrt et al. (Nehrt et al., 2011), comparing human and mouse functional genomic data concluded that paralogs are often a much better predictor of function than orthologs, even at lower sequence identities. This paper warned about the general use of the 'ortholog conjecture' (where orthology is synonymous to function conservation) and led to hot debate in the orthology community, with the subsequent publication of several new studies. GO consortium members demonstrated that the bias noted by Nehrt et al. between different classes of homologous genes in human and mouse, is more likely to reflect a global bias in the GO annotations for all human and mouse genes (Thomas et al., 2012). In support of the orthology conjecture, other authors have shown that in general, small molecule binding is conserved for pairs of human to rat orthologs (Kruger and Overington, 2012). Finally, a recent paper moderates all these conclusions by demonstrating that orthologs generally have more similar functional annotations than paralogs, but the difference between orthologs and paralogs is weaker than expected under a naive understanding of the orthology conjecture, especially when GO Molecular Function and Biological Process are considered separately (Altenhoff et al., 2012).

2.2.5 Some extended definitions: Ohnology, gametology

Several other specialized homology definitions have appeared recently, but are mainly used in specific biological domains. Here are some examples:

- **Ohnology:** Ohnologs are paralogous genes that duplicated through a process of genome duplication (Wolfe, 2000). This term was originally introduced by K. Wolfe in honour of Susumu Ohno who proposed that whole genome duplications (WGDs) are key evolutionary transitions in chordate evolution (Ohno et al., 1968). Consequently, this homology definition is closely related to genome context and synteny conservation and is mainly used in evolutionary studies concerning species that are known to have experienced multiple genome duplications. Example studies include animal models such as *Danio rerio*, where two WGDs modelled the *teleostei* lineage (Postlethwait, 2007) or *Saccharomyces cerevisiae* where ohnologs represent 17% of genes (Byrne and Wolfe, 2005).

- **Gametology:** This term was introduced to describe homologous regions in opposite sex chromosomes. Gametologous genes arise via non-recombination and differentiation of sex chromosomes (Garcia-Moreno and Mindell, 2000). Thus, genomic context is the key to differentiating gametologs in the particular case of recombination barriers occurring between portions of opposite sex chromosomes. These barriers are considered similar to the lineage splitting and gene duplication that are used for orthology and paralogy inference.

2.3 Approaches to establish sequence homology

As stated above, homology is a binary relation and two sequences are said to be homologous if they share a common ancestor. Unfortunately, we do not generally have access to information about the ancestors and therefore homology cannot be determined explicitly. Thus, sequence similarity is often used to predict ancestral states and to hypothesize homolog, ortholog and paralog relationships. Many bioinformatics approaches have been developed in genomics and phylogenetics to estimate sequence similarities.

2.3.1 Sequence alignments

During evolution, genes encoding RNA and proteins can be mutated. These mutations can concern only one residue or hundreds of residues if a whole part of sequence is deleted, inserted or undergoes recombination. These modifications can induce different consequences: no change in terms of function or expression, a loss of function or expression or the acquisition of new functions. Functional changes are closely related to the 3D structure of proteins and RNA. If a mutation changes the structure of catalytic sites or the structure of domains linked to molecular interactions, the function of such a protein can be lost or modified. Consequently, comparing sequences with their homologs is key to understanding functional properties of genes and the molecular evolution of RNA/protein families (Rustici and Lesk, 1994). In this context, sequence alignments are fundamental tools to compare DNA, RNA and protein sequences in molecular or evolutionary biology. They are used to understand which key residues and sequence motifs have been conserved during evolution. On the other hand, the search for non conserved regions in sequences can highlight functional loss/gains or specific genetic events (Lecompte et al., 2001). More explicitly, multiple alignments are now used for various purposes:

- To analyze of protein organization: domains, insertions/deletions…
- To validate sequences: detection of sequencing errors, frame shifts, start/stop codon prediction errors, intron/exon prediction errors…
- To describe protein families and, by extension, their evolutionary context
- To distinguish orthologs and paralogs
- To analyze differential conservation of discriminating residues between sequence sub-families
- To predict 2D and 3D structures
- To predict functions
-

The alignment of sequences involves the creation of a matrix where the rows correspond to multiple sequences and conserved residues are placed in the same columns. When the sequences are of different length, insertion-deletion events (indels) are hypothesized to explain the variation and gaps are introduced into the alignment. Alignments can be produced by a wide variety of alignments that can be classified in four main categories (figure 2-3, adapted from (Lecompte et al., 2001)).Block alignments (figure 2-3A) represent only the most conserved motifs and do not contain any gaps. They

are used by the Probe program (Neuwald et al., 1997) and in the Blocks database (Henikoff et al., 2002). Segment alignments (figure 2-3B) are used by a number of database search programs such as BLAST (Altschul et al., 1990) and PSI-BLAST (Altschul and Koonin, 1998), and in pattern/motif databases such as Pfam (Punta et al., 2012b) or ProDom (Bru et al., 2005). They contain the most similar regions of the sequences and may contain short gaps representing indels. Local and global alignments (figure 2-3C,D) both contain the complete protein sequences and are typically produced by multiple alignment programs such as Dialign (Morgenstern, 2007; Morgenstern et al., 1998) or ClustalW (Larkin et al., 2007; Thompson et al., 1994). In local alignments, the conserved motifs are identified and the rest of the sequences are included for information only. Thus, only a subset of the residues is actually aligned. In global alignments, all the residues in both sequences participate in the alignment. More recently, programs combining local and global alignment have been developed. We can cite Probcons (Do et al., 2005), TCoffee (Wallace et al., 2006a) and the most recent version of Mafft (Katoh and Toh, 2008b) or Clustal Omega (Sievers et al., 2011). These are generally more accurate than older methods based on global or local algorithms alone (Thompson et al., 2011).

Figure 2-3. Four different types of multiple sequence alignment

2.3.2 Alignments based on higher level criteria

While useful for detecting homologies, multiple alignment techniques often fail when presented with a set of sequences sharing low identity (Thompson et al., 2011). To solve these drawbacks, the use of structural information, when available, has proved to be useful, because structures diverge at a lower rate compared to sequences (Abagyan and Batalov, 1997; Whisstock and Lesk, 2003). Several

aligners complete sequence homology with an additional comparison of 3D structures. This approach is exploited in the T-Coffee software suite (Di Tommaso et al., 2011) which includes two extensions, R-coffee (Wilm et al., 2008) and Expresso (Armougom et al., 2006) for the consideration of RNA and protein secondary structures respectively. Performing an alignment of 3D structures can require a considerable amount of CPU time. To address this, the authors of PROMALS3D make use of pre-computed structural alignment databases (Pei et al., 2008).

Another approach is not to complete classical sequence alignments with structural data but to directly consider the structural properties of sequence residues. Such alignment approaches can be roughly divided into three groups (Shealy and Valafar, 2012). Two groups use 3D data and optimize scores based on rigid-body superposition or tertiary interactions (distance matrices, contact maps). The last group includes 1D methods, which exploit the sequence itself by assigning residues to a vector of relevant properties and generally use faster string algorithms. For example, CLEMAPS (Friedberg et al., 2007) uses conformational letters coding probable conformational states of protein fragments (figure 2-4). Vorometric (Sacan et al., 2008) uses Voroni tessellations to determine the residue's environment. YAKUSA (Carpentier et al., 2005) and 3D-BLAST (Yang and Tung, 2006) use conformational angles.
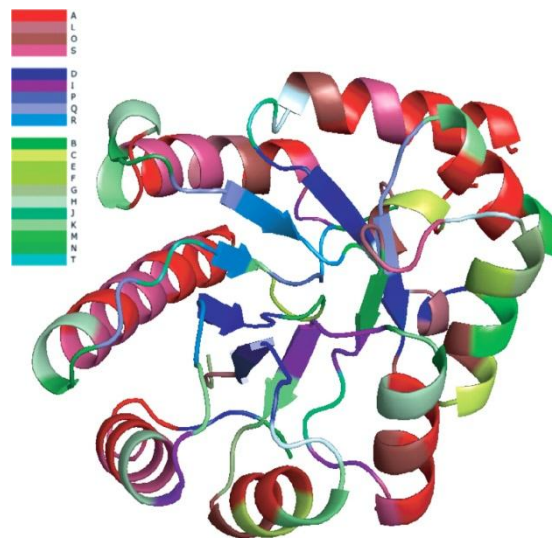
Figure 2-4. The 3D structure of eukaryotic ornithine decarboxylase, in which residues are colored by their structural properties. *Red hues designate fragments that have a high frequency in helices, blue hues those fragments that have a high frequency in strands. Adapted from Friedberg et al., 2007.*

Despite constant advances, 1D methods are mainly designed for fast database searching and have not compared favorably with 3D methods (figure 2-4) (Friedberg et al., 2007; Hasegawa and Holm, 2009). Consequently, use of structural alignments based on sequence to establish sequence homology remains anecdotic.

2.3.3 Phylogenetics

Phylogenetic-based approaches use alignment-based sequence similarity estimations in an evolutionary framework to predict the detailed evolutionary relationships between species or genes (Yang and Rannala, 2012). A phylogeny describes the ancestral states of a set of molecular sequences and the ancestral relations are generally represented by a phylogenetic tree. For example, the accepted universal tree of life, in which the living world is divided into three domains (bacteria, archaea, and eucaryota), was constructed from comparative analyses of ribosomal RNA sequences (Forterre and Philippe, 1999). Phylogenies can be used not only to describe molecular evolution but also complex evolutionary patterns. Some recent examples illustrated in figure 2-5 include the highly resolved tree of Life based on available complete genomes (Ciccarelli et al., 2006), the origin and spread of viral infection (Iyer et al., 2006) or the demographic changes and migration patterns of species (Grehan and Schwartz, 2009).
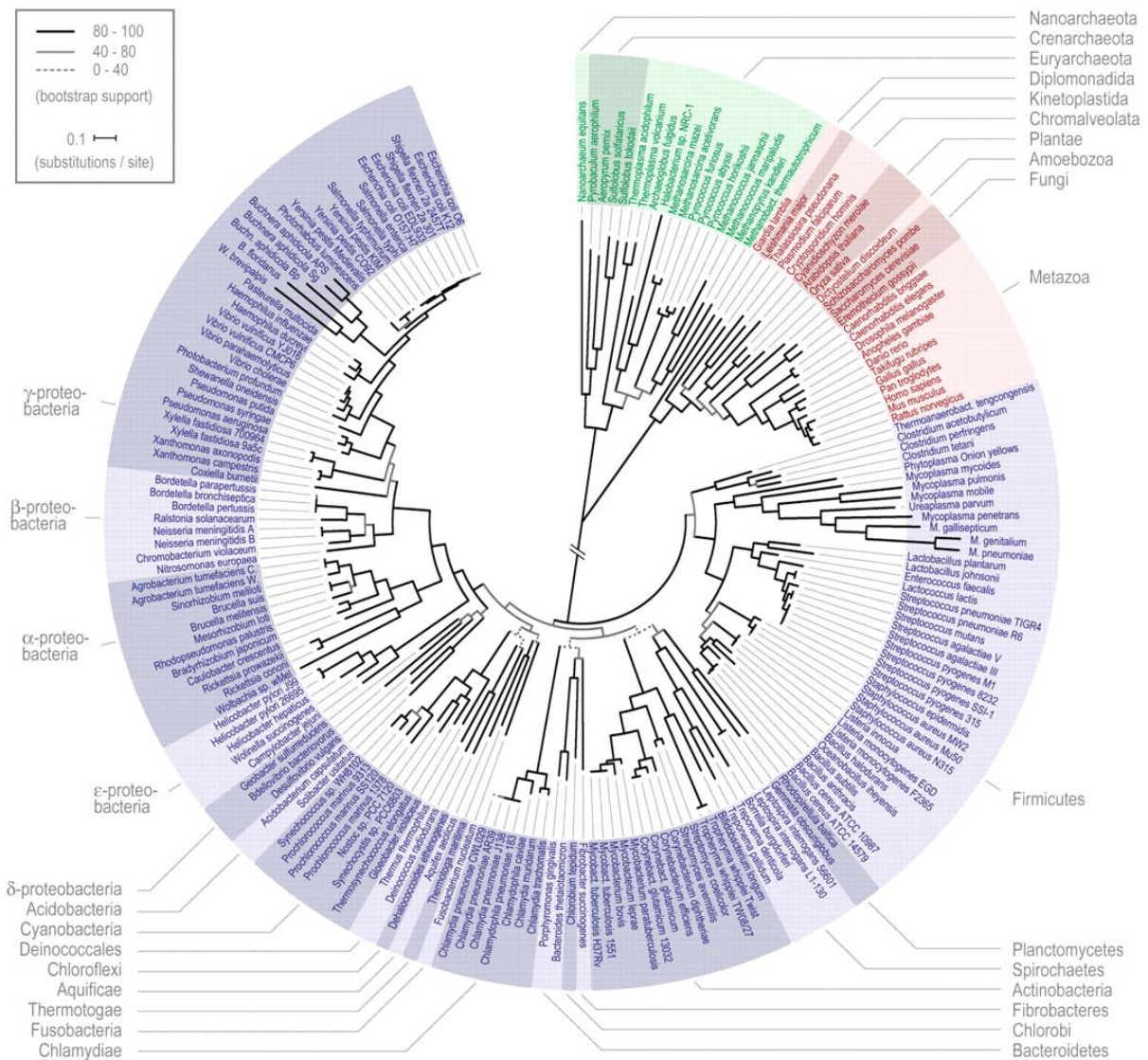
Figure 2-5. Global phylogeny of fully sequenced organisms in 2006. *The tree is based on a concatenated alignment of 31 universal protein families and covers 191 species. Green section, Archaea; red, Eukaryota; blue, Bacteria.*

Since all methods of phylogenetic tree reconstruction use distance measures based on multiple alignments, strategies used to construct alignments can have a large influence on the resulting phylogeny (Wang et al., 2011; Wu et al., 2012). The methods for calculating phylogenetic trees fall into two general categories (Pevsner, 2009b). These are distance-matrix methods, also known as clustering or algorithmic methods (e.g. UPGMA or neighbour-joining), and discrete data methods, also known as tree searching methods (e.g. parsimony, maximum likelihood, Bayesian methods). The use of phylogenetic trees for the detailed characterization of homology relationships will be discussed further in the next chapter.

3   ORTHOLOGY INFERENCE IN THE POST-GENOMIC ERA

In genomics studies, the classification of genes according to their evolutionary relationships is an essential step. In general, it is assumed that orthologous genes tend to conserve the same function, while paralogous genes can acquire new functions (the functional aspects of orthologs are discussed in paragraph 2.2.4.). The recent emergence of high-throughput sequencing and the dramatic increase of available complete genomes have increased the importance of orthology in functional annotation or comparative genomics. In this chapter we will review the current state of the art concerning computational methods for orthology inference and discuss future challenges. For readability purposes, we will refer to methods for detecting orthology relations, but obviously detecting orthology (resulting from speciation events) implicitly requires the inference of paralogy (resulting from duplication events).

3.1   A multitude of strategies

Automated detection of orthology relations between multiple organisms is a crucial issue in bioinformatics today, motivating numerous attempts to resolve methodological and computational issues. A profusion of methods has been developed during the last decade, most of them supported by pre-calculated orthology databases. A majority of these algorithms are based on protein sequences (single domain or complete transcripts) (described in 3.1.1). Other methods have also been developed that focus on protein domain-domain architecture (described in 3.1.2). More recently, the widespread use of omics approaches has introduced new potential characters for orthology inference, such as genomic context (described in 3.1.3) or conservation of molecular interactions (described in 3.1.4). Interestingly, the latter approaches rely strongly on the hypothesis of orthology functional conservation for their predictions, shifting the original definition from molecular evolution at the gene level to system evolution.

3.1.1   Sequence based inference

Today, dozens of orthology detection methodologies exist that are based on protein sequence homology. These approaches can be classified into four main classes, based on conceptual and practical differences (Altenhoff and Dessimoz, 2012; Chen et al., 2007; Kristensen et al., 2011; Kuzniar et al., 2008) : tree-based methods, graph-based methods, hybrid approaches and integrative approaches (table 3-1). Tree-based methods use multiple alignments followed by phylogenetic tree construction and infer orthology from the topology of the latter. Graph-based methods mainly use BLAST to estimate sequence similarities and estimate orthology with heuristics based on BLAST hit scores. Hybrid methods are a mix of tree-based and graph-based methods, generally confirming tree-based predictions with heuristic approaches. Finally, integrative approaches compile and score

results provided by several programs. Each approach has its own advantages and drawbacks, some of which are described in the following paragraphs.

Database name	Detection method	Covered phyla	Reference
COGs/TWOGa/KOGs	Graph	Bacteria, Eukaryota	(Tatusov et al., 2003; Tatusov et al., 1997)
COGs-COCO-CL	Tree	Bacteria	(Jothi et al., 2006)
COGs-LOFT	Tree	Bacteria	(van der Heijden et al., 2007)
eggNOG	Graph	Bacteria, Archaea, Eukaryota	(Powell et al., 2012)
EGO	Graph	Eukaryota	(Lee et al., 2002)
Ensembl Compara	Hybrid	Eukaryota	(Hubbard et al., 2007)
Gene-Oriented Ortholog Database		Vertebrates	(Ho et al., 2010)
GreenPhylDB	Tree	Plantae	(Rouard et al., 2011)
HCOP	Integrative	Eukaryota	(Seal et al., 2011)
HomoloGene	Hybrid	Eukaryota	(Wheeler et al., 2007)
HOGENOM	Tree	Bacteria, Archaea, Eukaryota	(Penel et al., 2009)
HOVERGEN	Tree	Vertebrates	(Dufayard et al., 2005)
HOMOLENS	Tree	Bacteria, Archaea, Eukaryota	(Penel et al., 2009)
HOPS	Tree	Eukaryota	(Storm and Sonnhammer, 2003)
INVHOGEN	Tree	Eukaryota (invertebrates)	(Paulsen and von Haeseler, 2006)
InParanoid	Graph	Eukaryota	(Ostlund et al., 2010)
KEGG Orthology	Graph	Bacteria, Eukaryota	(Kanehisa and Goto, 2000)
MBGD	Graph	Bacteria	(Uchiyama et al., 2010)
MGD	Graph	Mammalia	(Eppig et al., 2012)
OMA	Graph	Bacteria, Archaea, Eukaryota	(Altenhoff et al., 2011)
OrthoDB	Tree	Eukaryota	(Waterhouse et al., 2011)
OrthologID	Tree	Plantae	(Chiu et al., 2006)
OrthoInspector	Graph	Eukaryota	(Linard et al., 2011)
OrthoMCL	Graph	Eukaryota	(Chen et al., 2006)
Panther	Tree	Bacteria, Archaea, Eukaryota	(Mi et al., 2010)
PHOG	Hybrid	Bacteria, Archaea, Eukaryota	(Datta et al., 2009)
PhylomeDB	Tree	Eukaryota	(Huerta-Cepas et al., 2011)
PLAZA	Integrative	Plantae	(Van Bel et al., 2012)
P-POD	Integrative	Vertebrates	(Heinicke et al., 2007)
ProgGMap	Integrative	Eukaryota	(Kuzniar et al., 2009)
RoundUp	Graph	Bacteria, Archaea, Eukaryota	(DeLuca et al., 2012)
TreeFam	Hybrid	Eukaryota	(Ruan et al., 2008)
YOGY	Integrative	Eukaryota	(Penkett et al., 2006)

Table 3-1 Non-exhaustive list of current orthology databases. *Data compiled from (Kuzniar et al., 2008) and questfororthologs.org.*

3.1.1.1 Tree-based approaches

The standard definition for orthology was introduced during a phylogenetic tree analysis performed by Fitch W.M. (Fitch, 1970). It requires several steps. First, homologous sequences are collected and multiply aligned. Second, a gene or protein tree based on the alignment is constructed. Then, this tree is compared with a 'known' species tree to compare duplication nodes of the former with speciation nodes of the latter. This comparison is commonly called tree reconciliation (Page, 1994) and allows inference of speciation and duplication nodes. This standard protocol is widely used for low-throughput studies of gene family evolutionary histories, but requires manual intervention to check the reliability of the results. The post-genomic era has highlighted the need for automation of such analyses. As a consequence, several pipelines have been developed, focusing in particular on new algorithms for tree rooting and tree reconciliation, steps which have a large impact on the quality of orthology and paralogy prediction. The automated pipelines can take rooted or unrooted gene trees and species tree as inputs, or can use a set of sequences to create the multiple alignments and gene trees. Generally, such pipelines need the support of a computational grid.

Several well maintained databases exist that provide tree-based orthology predictions. HOGENOM and HOVERGEN databases contain tree-based predictions based on the RAP – Réconciliateur d'Arbres Phylogénétiques – tree-reconciliation program. RAP can handle unresolved trees and takes into account bootstraps and branch length parameters (Dufayard et al., 2005; Penel et al., 2009). SDI – Speciation Duplication Inference – (Zmasek and Eddy, 2001) and Orthostrapper (Storm and Sonnhammer, 2002) are two other tree-reconciliation programs focusing on bootstrap improvements and calculation of scores based on these bootstraps. They are used in the RIO (Resample Inference Ortholog) and HOPS (Hierarchical grouping of Orthologous and Paralogous Sequences) databases respectively. Both of these databases were constructed by applying tree reconciliation restricted to single Pfam domains and not complete transcripts (Storm and Sonnhammer, 2003; Zmasek and Eddy, 2002). More recently, alternative methods were developed that replace the tree reconciliation step with their own algorithms. Berkley-PHOG defines orthologs as sequences in different species that are each other's reciprocal nearest neighbour (RNN) in the tree. It introduces the concept of super-orthologs, based on a sub-tree containing only RNN orthologs (Datta et al., 2009). The PhylomeDB database predictions are based on Neighbour Joining trees that are optimized through the comparison of likelihoods calculated by seven models of branch-length optimization (Huerta-Cepas et al., 2007). Currently, PhylomeDB can be considered as the only tree-based database providing high-throughput predictions as it currently contains about 1,000 species and more than 2,000,000 trees, while new phylomes (complete collections of evolutionary histories of all genes in a genome) are regularly calculated.

Tree-based inference is particularly suitable for detecting specific evolutionary events such as gene loss or horizontal gene transfers (Dufayard et al., 2005), which are common events in molecular evolution (Blomme et al., 2006). However they present a number of major drawbacks, most of them being inherent to phylogenetic tree reconstruction:

- Need of a reliable species tree for tree-reconciliation, which can be difficult when studying unclearly defined phyla.
- Different algorithms for multiple alignment and phylogenetic tree reconstruction can produce heterogeneous results. Indeed, the quality of the multiple alignments has a large

impact on the phylogenetic tree and fast evolving genes can induce problems of long branch attraction (Bergsten, 2005).

- The root of the tree must be established, generally with the use of an outgroup. However, outgroups must be selected carefully, making the criterion less useful in automated large scale analysis, where some gene families may not be present in the outgroup species (Huelsenbeck et al., 2002).

- Tree reconciliation is a complicated task because it assumes that gene and species trees contain no errors (Goodman et al., 1979). As a gene tree can be inferred only from its family, reconciliation uses a limited amount of information, reducing the confidence of the reconciliated tree. Some authors introduced bootstrap values to support orthology levels (Yuan et al., 1998). Despite these efforts, some consistent bias in such tree reconciliation methods is still present, with an overestimation of the number of duplicates placed near the root and an overestimation of the number of losses across the tree (Hahn, 2007).

- Horizontal Gene Transfer (HGT) can be problematic for reliable phylogeny, in particular in prokaryote lineages. Despite some algorithmic corrections based on tree incongruence compared to the species tree, building phylogenetic trees taking into account HGT remains challenging (Philippe and Douady, 2003).

- All steps are computationally expensive for large gene families. Thus, tree-based inference can be difficult to apply in large-scale analysis.

Some authors have proposed alternative approaches to address these drawbacks. The need for a species tree can be avoided in two methods - COCO-CL: COrrelation COefficient based Clustering - (Jothi et al., 2006) and – LOFT: Levels of Orthology From Tree - (van der Heijden et al., 2007). Concerning the tree reconciliation steps, the bootstrap approach was complemented by a support value for all orthologous pairs by (Storm and Sonnhammer, 2002) and more recently tree reconciliation was adapted to Bayesian frameworks, improving correct assignments of horizontal gene transfers (Chung and Ane, 2011). Finally, some authors have focused on computational speed optimizations. For example, QuartetS decomposes polygenetic trees into quartet trees followed by the construction of a split network (Yu et al., 2011). The advantage of quartet trees is that they can adopt only 3 different topologies. Thus, decomposition of a tree into a set of quartets, estimation and filtering of the best bootstrap-supported quartet and finally reconstruction of a reconciliated tree are faster operations than the analysis of all possible topologies of a large phylogenetic tree.

3.1.1.2   *Graph-based approaches*

In principle, phylogenetic tree-based inference represents the most accurate way to determine orthology and paralogy (Brown and Sjolander, 2006; Gabaldon, 2008). However, it requires complex automated pipelines with extensive computational requirements, reducing their usability in genome-scale studies. In order to cope with the constant influx of new complete genomes, new heuristic methods have been developed. They rely on the general assumption that orthologs are more similar than paralogs. They are all based on pairwise sequence similarity searches, mainly using BLAST or Smith-Waterman alignments.

The first method using sequence pairwise comparison to detect orthologs is the Reciprocal Best Hit (RBH) (Tatusov et al., 1997). An RBH exists between two sequences belonging to two different

organisms if they are reciprocally the most similar sequences compared to all other sequences of both organisms. Consequently, this relation can only link 2 entities and cannot detect complex relations of co-orthology. Indeed, if a gene duplication event occurred after a speciation event, the orthology relation is complex and can link several inparalog genes (see section 2.2.2). In fact, for comparison of two organisms (O1 and O2), we have three types of pairwise entity relationships (figure 3-1). A 1-to-1 relationship is defined by a best hit between a protein of O1 and a protein of O2, complemented by a returning best hit from the protein of O2 to the protein of O1, known as a reciprocal best hit. A 1-to-many relationship is defined by a best hit from a given protein of O1 to any protein member of an inparalog group of O2, complemented by a returning best hit from any member of the inparalog group of O2 to the same protein of O1. Finally, a many-to-many relationship is defined by two best hits between proteins of two groups of inparalogs (a group in O1 and a group in O2).
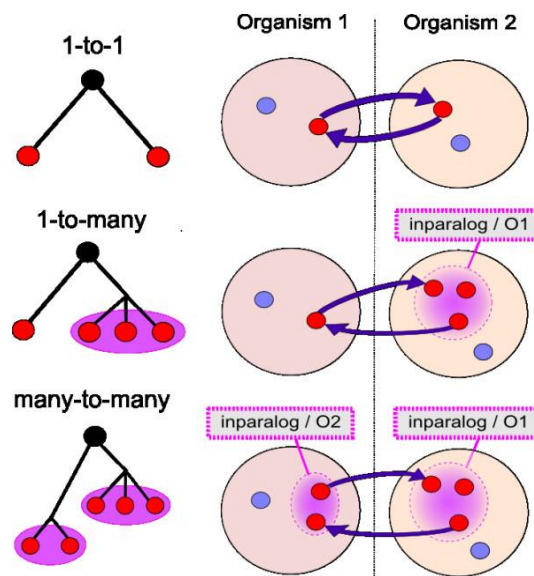
Figure 3-1. Comparison of inparalog groups. *BLAST best hits are used to define the potential relationships existing between inparalog groups.*

As for the tree-based methods, the availability of numerous genomes showed the limits of the RBH and its incapacity to predict complex co-orthology relations. Consequently, a large range of alternative heuristics has been developed to exploit pairwise gene similarities in large-scale studies. They use a dataset representing BLAST comparisons of all genes belonging to several species, commonly known as a BLAST 'all-versus-all'. This procedure assigns similarity values to all possible gene pairs and provides the basis to infer orthology in the so-called graph-based methods, where the graph is made up of nodes corresponding to genes and edges representing BLAST best hits or other sequence similarity measures. We can distinguish three types of graph-based methods: direct use of pairwise distance calculations, construction of 3-way best-hits and clustering approaches.

Inparanoid (Ostlund et al., 2010) was the first and only graph-based method based on pairwise distance calculations when we started development of OrthoInspector (see chapter 7). Inparanoid

predictions are based on a pairwise comparison between two organisms. The search for co-orthologs is centred on the RBH existing between two organisms. If a sequence is more similar to the sequences defining the RBH than to the similarity threshold defined by the RBH itself (in the respective organism), Inparanoid considers it to be an additional inparalog implicated in the orthologous relation. Consequently, Inparanoid can predict 1-to-1, 1-to-many or many-to-many relations between two organisms. This approach was later improved in the Ortholuge program to better handle gene-loss events by using phylogenetic distance ratios instead of BLAST similarities (Fulton et al., 2006).

COG/KOG, OrthoDB and eggNOG are three representatives of the 3-way best-hits approach, the former establishing the original method and the two latter extending it. In COGs (Cluster of Orthologous Groups) or KOGs (eukaryotic clusters of Orthologous proteins) the 3-way best hits are defined as the symmetrical best hit linking the proteins of three organisms (Tatusov et al., 2003; Tatusov et al., 1997). This triangle specifies the minimal COG (the minimal cluster). Then, all best hits linking these three proteins to new proteins of other species are used to extend the cluster of orthologous proteins. This approach is useful to rapidly group proteins belonging to the same gene family. However, a COG cluster can mix proteins with both orthology and paralogy relations because it does not consider any hierarchy between the species. To address this limitation, the eggNOG algorithm was developed (Jensen et al., 2008). Using COG clusters as a basis, the algorithm first defines in-paralogous proteins, then assigns orthology between proteins by joining triangles of reciprocal best hits at a given cutoff. Gene fusion cases are handled by avoiding fusion of clusters defining two homolog families. Then, a new triangulation is done by reducing the threshold to reach a wider taxonomic range. This process is repeated to create a hierarchical clustering of homologous proteins at different taxonomic levels. The OrthoDB database was constructed using the same protocol with some adjustments (Kriventseva et al., 2008). While eggNOG defines hierarchical groups corresponding to major taxonomic levels, OrthoDB defines hierarchical groups corresponding to any split in their species tree.

The third type of method uses clustering of the BLAST all-versus-all data to construct a best-hit graph. What differentiates clustering-based methods is the clustering algorithm they use. OrthoMCL (Li et al., 2003) applies a Markov Cluster algorithm on the graph to obtain clusters grouping orthologs and recent paralogs. Similarly, the DomClust algorithm (Uchiyama, 2006) performs a successive contraction of the graph using UPGMA clustering (Unweighted Pair-Group Method using arithmetic Averages). OMA uses a maximum-weight clique algorithm to cluster ortholog pairs into ortholous groups (Roth et al., 2008). SuperPartitions algorithm (Tekaia and Yeramian, 2012) first defines inparalogs using an in-house partitioning method and Markov Clustering of the intra-species reciprocal significant hits. Then, it links inparalog groups with classical inter-species RBH. The ReMark algorithm (Kim et al., 2011) first determines proto-clusters by a recursive method similar to Inparanoid, then confirms the clusters with Markov Clustering. Interestingly, most clustering-based approaches take into account the gene-fusion issues with heuristics, e.g. clusters should be fusioned if a fusion protein links them.

Finally, the RSD – Reciprocal Smallest Distance - algorithm (Wall et al., 2003) cannot be classified in any of the three previous categories. It combines local and global sequence alignments and maximum likelihood estimation of evolutionary rates to predict orthologous proteins. This algorithm is the basis for the RoundUp database (DeLuca et al., 2012).

3.1.1.3 Hybrid approaches

Hybrid methods use a mix of tree-based and graph-based methods at different stages in their pipelines. The tree concept is mainly used to refine ortholog groups or to guide ortholog clustering with the help of a phylogenetic tree. The goal of the method integration is to combine the phylogenetic resolution with the scalability of graph-based method (Kuzniar et al., 2008). Consequently, they can be applied to genome-wide analysis.

Ensembl Compara (Hubbard et al., 2007), HomoloGene (Wheeler et al., 2007), OrthoParaMap (Cannon and Young, 2003), PhIGs (Dehal and Boore, 2006), PHOGs (Datta et al., 2009), PhyOP (Goodstadt and Ponting, 2006), TreeFam (Ruan et al., 2008) are all examples of hybrid methods. They all create clusters of homologous sequences, then infer orthology/paralogy relations in these clusters using different criteria for refinement. Ensembl Compara builds clusters based on BLAST best-hits, constructs a phylogenetic tree for each cluster, labels ortholog relations as 1-to-1, 1-to-many, many-to-many between species pairs and completes predictions with new orthologs predicted from genome context through whole-genome alignments. Homologene, PhIGs and PHOGs use an incremental clustering of homologous sequences guided by a species phylogeny. OrthoParaMap integrates BLAST similarities, gene phylogenies and uses a tree-reconciliation step based on conserved gene neighbourhood and not species tree. PhyOP uses a clustering technique and a phylogenetic reconstruction that takes into account multiple transcripts of the same gene and can distinguish between functionally active orthologs and pseudogenes. TreeFam is a curated tree database whose originality, similarly to the Pfam database, comes from its split into automatically generated trees (TreeFam-B) and manually curated trees (TreeFam-A). In both databases, orthologs are inferred with a hierarchical clustering and phylogenetic tree reconstruction procedure.

3.1.2   Domain architecture based inference

It has been proposed that domain architecture composition is likely to be conserved during evolution due to functional constraints (Vogel et al., 2004). For this reason, while the similarity between primary sequences of orthologs may decrease dramatically in distantly related species, the domain composition is more likely to be conserved through evolution. Based on this assumption, some authors have used domain architecture to detect orthologous relationships between distantly related species. In particular, Chen at al. developed the DODO - DOmain based Detection of Orthologs – algorithm. It groups proteins with the same domain architecture, then refines the orthologous relationships within each group by identifying RBH in this smaller protein set. This strategy has proven to be efficient for the large-scale analysis, because the sequence clustering based on domain architecture eliminates the need for the classical BLAST all-vs-all construction. However, as the DODO algorithm needs the support of an annotation database (Pfam is used in the current version), it strongly depends on the amount of information that this database contains, thus restricting its applicability to a few species. Proteins without Pfam domain information are all grouped in an uncharacterized cluster and must be analyzed independently from DODO. Another criticism is that domain architecture can evolve faster than domain conservation, restricting the analysis to closely-related species.

3.1.3   Orthology and genomic context

An alternative approach has been to use conservation of gene neighbourhood to infer orthology relations, in particular when homologs share low similarity (Simillion et al., 2004) or in the case of single copy paralogs obtained after the loss of a member of the paralogy pair (Scannell et al., 2007). However, such a comparison is restricted to closely related species (Huynen and Bork, 1998). Several hybrid methods consider gene neighbourhood conservation to refine sequence similarity based predictions, although sequence similarity remains the first criteria to detect a set of orthologs before refinement. Several groups attempted to directly infer the orthology relationship by looking at conserved gene context (Dewey, 2011). These methods can be classified into three classes.

The first class is based on best hits between genomes to define 'clear' orthologs, i.e. only 1-to-1 orthologs in most methods. For example, the EGM - Encapsulated Gene-by-gene Matching - (Mahmood et al., 2010) and SYNERGY (Wapinski et al., 2007) algorithms search for optimal matching between gene sets of two genomes, based on an objective function that takes into account gene neighbourhood conservation. EGM constructs a bipartite graph based on one-to-one orthologs between two genomes, where edge weights represent both similarity and neighbourhood conservation. Orthologs are then inferred by finding a maximum matching in the graph. SYNERGY uses a protein-level evolutionary distance and a gene neighbourhood similarity score to produce both cluster-based and phylogeny-based orthology predictions. IONS - Identification of Orthologs by Neighbourhood and Similarity – (Seret and Baret, 2011) is similar to SYNERGY but does not require dataset-dependant parameters. The OrthoParaMap algorithm (Cannon and Young, 2003) predicts orthologs and paralogs mainly with blocks of conserved gene order but is able to analyze only one gene family as input and not an entire genome. This is the only method in this class not restricted to 1-to-1 relationships.

The second class includes methods that reconstruct parsimonious gene order evolutionary scenarios using models of evolution. A first approach is to construct a history of the events that induced the observed genomic position conservations (Sankoff, 1999). A gene pair linking two genomes is selected in each gene family such that if unselected genes of a family are removed, the distance between the resulting gene orders is minimized. These pairs are estimated to be the best candidates for positional orthology between genomes. MSOAR 2.0 (Shi et al., 2010) is an orthology prediction method based on reversal distance that extends the evolutionary model to take into account duplications, translocations and chromosomal fusions/fissions. MultiMSOAR 2.0 further extends the method of neighbourhood comparison to multiple genomes (Shi et al., 2011).

The third class of methods uses whole-genome alignments at the nucleotide level. They are generally limited to genomes with high linear colinearity, i.e. closely-related species (Blanchette, 2007). While they could in theory be used to estimate 1-to-many and many-to-many relations in the case of reference-based alignments, most of them are restricted to prediction of 1-to-1 orthologs and can be classified into hierarchical, local and hybrid approaches (Dewey, 2011). Hierarchical approaches first construct a high-level collinear orthology map between several genomes. This is a critical step referred to as the 'synteny block' finding problem that decomposes genomes into conserved blocks. Then, a nucleotide–level global alignment is computed for each block. Hierarchical methods thus combine two tools (conserved blocks search and aligner), for example: Mercator and MAVID (Dewey,

2007), Shuffle-LAGAN and LAGAN (Brudno et al., 2003) or Nucmer and SeAn::TCoffee (Angiuoli and Salzberg, 2011). Local methods follow the reverse sequence of operations; first small genome fragments are aligned, second a chaining of these fragments highlights longer collinear segments. Alignment of multiple genomes is obtained by a progressive merging of overlapping pairwise alignments. Some example methods use BLASTZ (Schwartz et al., 2003), MUMmer (Delcher et al., 2003) or CHAINNET (Kent et al., 2003). ProgressiveMauve (Darling et al., 2010) is a hybrid method that performs several rounds of finding local alignments, identifying sets of 1-to-1e collinear segments and filtering of paralogous segments.

3.1.4   Biological network-based inference

Recently, the availability of large-scale protein-protein interaction (PPI) networks has motivated the idea to exploit higher biological levels for orthology inference. This idea is controversial because it is focused on the functional conjecture of orthologs and not on their evolutionary definition. Thus, authors developing these approaches use the term 'functional ortholog' to differentiate their inferred relations from the classical orthology definition. This denomination still remains confusing and contributes to the general misunderstanding of the orthology concept. Here we will briefly describe these methods inferring 'functional orthologs'.

One of the first network alignments used for orthology detection was performed by Ogata et al., who predicted orthologous groups of proteins corresponding to conserved metabolic pathway motifs (Ogata et al., 2000). This study combined gene neighborhood conservation and metabolic pathway data for 10 bacterial genomes. In each organism, groups of functionally related enzymes were formed with the help of both criteria. Then, corresponding enzymes were mapped between the different species by an EC number matching. Finally, the alignment wasmanually refined based on sequence similarity.

Later, global interaction networks for model organisms became available, initiating algorithmic developments for a Global Network Alignment (GNA) of biological networks. The global scope of GNA enables species-level comparisons of biological networks. Bandyopadhyay et al. developed the PathBLAST algorithm to align conserved pathway segments in *Drosophila melanogaster* and *Saccharomyces cerevisiae* (Bandyopadhyay et al., 2006). More recently, Towfic et al. expanded such approaches by simultaneously integrating PPI and gene co-expression networks in fly, yeast, mouse and human (Towfic et al., 2010). Their alignment of networks and orthology assignation is supported by decision trees and probabilistic frameworks (Naïve-Bayes and Support Vector Machine). They found that such orthology prediction provided good results by comparing their data to the KEGG orthology database. This result can be criticized because they validated their data in a cyclic way. Indeed, the KEGG database orthology predictions are based on a combination of features from phylogenetic analysis, sequence similarity and similar pathway topology (Mao et al., 2005). Finally, Park et al. created Isobase, the first PPI network alignment-based orthology database (Park et al., 2011). However, this latter clearly mentioned the functional definition of its orthologs, which are described as 'isologs' (figure 3-2). These functional orthologs are derived from the multiple alignment based on the IsoRank algorithm (Singh et al., 2008) and performed for five major eukaryotic PPI

networks (yeast, fly, worm, mouse, and human). The latest version of the database also includes genetic interaction networks.
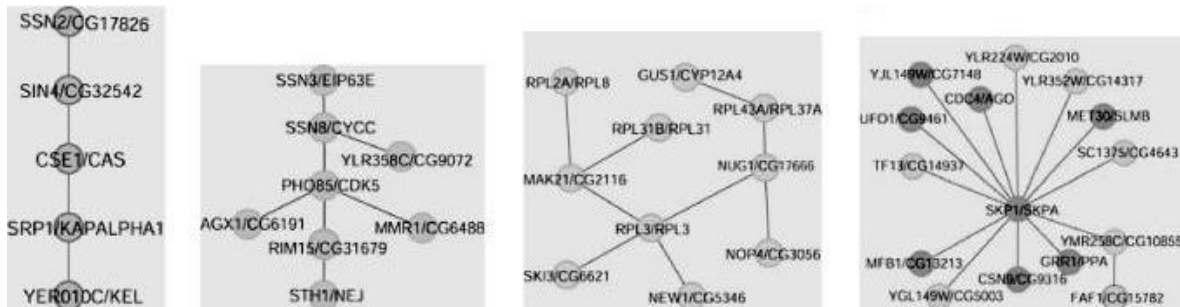
Figure 3-2. Some examples of conserved PPI subgraphs extracted from the yeast-fly Global Network Alignment (GNA). *The detection of conserved sub-graphs can be related to the detection of functional orthologs, also referred to as isologs. Node labels correspond to yeast/fly proteins. Adapted from Park et al., 2011.*

3.2   Limits of orthology inference

3.2.1   Coping with the increasing data influx

Orthology inference is essential for comparative and functional genomics. In the first decade of this century, the genomes of dozens of eukaryotic species and hundreds of prokaryotic species were completely sequenced. In this context, the development of graph-based methods for orthology inference allowed comparative analysis of these genomes. Today, technological advances in sequencing allow the publication of new complete genomes every day, providing incredible opportunities for comparative genomics and evolutionary studies at every taxonomic level. Despite being less computationally consuming than tree-based methods, graph-based methods are still based on sequence similarity tools such as BLAST. In particular, they require a BLAST all-versus-all as input, implying a quadratic number of sequence comparisons. Continuing to analyze all the available genomes with BLAST could rapidly become a critical limitation in terms of computational resources.

Recent methods are therefore focused on algorithmic optimization to further reduce the computational costs of orthology predictions. For example, Proteinortho (Lechner et al., 2011) uses heuristics similar to Inparanoid with algorithmic and multi-threading optimizations. It reduces the number of pairwise sequence comparisons required by concentrating on the most informative comparisons. xBASE-Orth (Halachev et al., 2011)) applies a "divide and conquer" algorithm to sequence similarity searches, avoiding the calculation of a BLAST all-against-all dataset. It creates pan-genomes - core genome shared by all species (Tettelin et al., 2005) - as proxies for the full collections of coding sequences at each taxonomic level and progressively climbs the taxonomic tree using the previously computed data. This interesting approach could be a first step towards the continuous incorporation of novel complete genomes that avoids the recalculation of all sequence

similarities. It should be noted that both of these approaches were used efficiently on thousands of genomes, but were restricted to bacterial and archaeal genomes. It would be interesting to apply such approaches to eukaryotic genomes where complex paralogy relations are more frequent. These optimizations could be complemented by BLAST sampling methods, such as that proposed by Friedrich & al., designed to significantly reduces the number of homologous sequences required for analysis and extraction of relevant information (Friedrich et al., 2007).

3.2.2 Domain recombinations, gene losses and horizontal gene transfers

Changes in domain architecture, produced by gene fusion/fission events and other evolutionary processes are a significant source of error in many genome annotation pipelines (Abascal and Valencia, 2003; Galperin and Koonin, 1998). In principle, all phylogenetic methods could handle such genetic events by resolving discrepancies in phylogenetic trees (Sjolander et al., 2011). However, the main limitation of domain-based phylogeny for orthology identification is the dependence of phylogenetic methods on sufficient site data, i.e. the length of the multiple sequence alignment. This drawback is particularly true for domain-based methods that exclude small domains with a length <50 residues (Moret et al., 2002). In the case of graph-based methods, in particular clustering-based methods, several algorithms have attempted to detect protein fusion/fission when detecting overlapping cluster of orthologous proteins (see 3.1.1.2). Nevertheless, for both approaches, multi-domain proteins still require manual refinements in general.

Horizontal gene transfer (HGT) is an important phenomenon, in particular in prokaryotes (Keeling and Palmer, 2008). It creates a particular homology relation known as xenology (see 2.2.3). Currently, few orthology inference methods explicitly take into account such relations because it requires a complex phylogenetic analysis, combining phylogenetic incongruence, atypical sequence composition and insertion/deletion patterns (Sundin, 2007). One solution may be to consider the phylogenetic distribution of predicted orthologs through visualization tools (Linard et al., 2011).

Gene loss remains the most important potential source of error for orthology predictions. In cases of multiple gene loss, most graph-based methods cannot differentiate between out-paralogs and orthologs (figure 3-3). Graph-based methods are particularly sensitive to gene loss because they consider only 'existing' relations and do not compare them to a species tree (Kuzniar et al., 2008). In theory, tree-based methods should be able to detect such genetic events, although only Ensembl Compara and TreeFam explicitly address this problem in their algorithm.
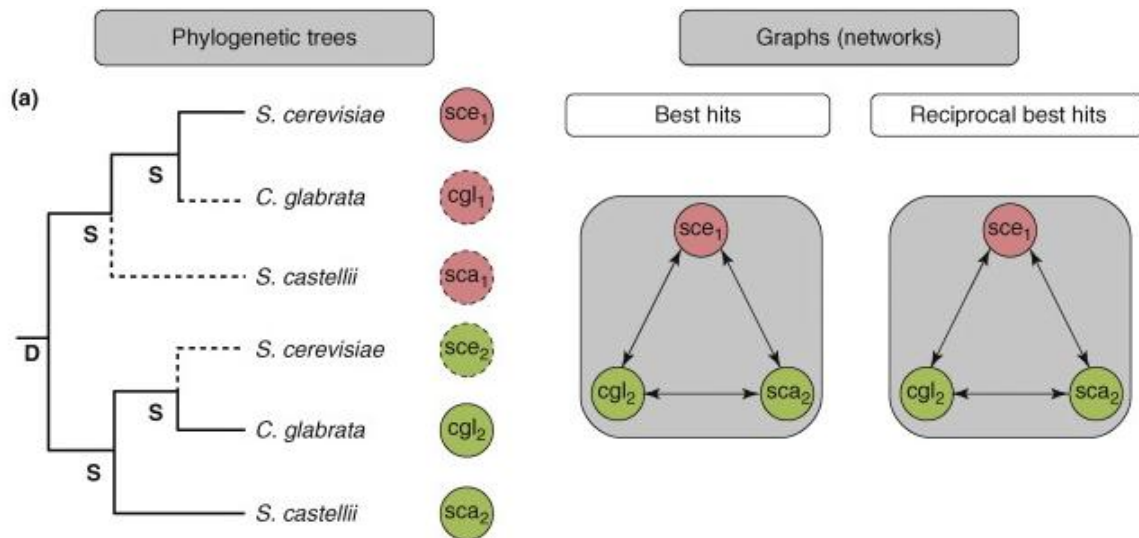
Figure 3-3. Homology relations in gene families where gene loss occurred as seen by tree-based and graph-based methods. *Contrary to graph-based methods, tree reconciliation between gene tree and species tree allows the detection of gene loss events in phylogenetic approaches. Sce1, cgl2 & sc2 are out-paralogs (duplicated prior to speciation), but are considered as orthologs by graph-based methods. D and S indicate duplication and speciation event respectively. Full lines represent existing genes, dotted lines represent lost genes. Adapted from Kuzniar & al.2008 , fungi data from Scannel & al.(2007) .*

3.2.3   The problem of alternative transcripts for ortholog prediction

Most orthology prediction methods use a single reference sequence for each protein-coding gene, avoiding the complex problem of alternative transcripts. However, most multi-exon genes can encode multiple protein isoforms, which often have different functions and may even be disease-related. In 2010, Jia et al. published a study extending orthologous groups predicted by Inparanoid to all alternate transcripts available in general sequence databases (Jia et al., 2010). Each ortholog group was divided into sub-clusters based on the sequence similarity of the isoforms. They observed that in considered species, functional similarity was higher within than between transcript-based sub-clusters for a single orthologous group. In other words, the function was more conserved between isoforms with the same exon structure in different species than isoforms with an alternative exon structure in the same species. This conclusion lead them to strongly recommend extension of the concept of orthology from the gene to the transcript level, by considering isoform sub-clusters of orthologous gene groups when available. Indeed, this would improve automatic propagation of functions from one isoform in a gene-based ortholog group to all equivalent isoforms in another species, thus limiting annotation and propagation errors. As an alternative approach, Fu and Lin produced a set of exon-level orthologous relationships from assigned gene ortholog pairs in the human and mouse genomes (Fu and Lin, 2012). First, they highlighted the current limits of orthology prediction at the exon level: about 26% of human and 11% of mouse genes in their dataset of 16545 1-to-1 orthologs had alternatively spliced transcripts. This observation indicates a substantial lack of potential exons supported by various splicing isoforms when considering transcript information in

current database. However, for genes where multiple transcripts are available, 92% of united exons were associated within an orthologous pair and several cases of 1-to-many exon pairs were observed, illustrating an interesting evolutionary behaviour for exon generation.

3.2.4 Performance of orthology predictions

Orthology prediction is a complicated task because the analysed subject, the gene, can be subject to heterogeneous and atypical evolutionary mechanisms (multiple duplications, rapid sequence divergence, domain organisation change, gene loss…). Graph-based and tree-based methods have their own advantages and drawbacks (figure 3-4). Several independent studies have been performed to compare the performance of these different approaches (Altenhoff and Dessimoz, 2009; Boeckmann et al., 2011; Chen et al., 2007; Creevey et al., 2011; Linard et al., 2011; Salichos and Rokas, 2011; Sjolander et al., 2011; Trachana et al., 2011). The conclusions reached by these authors are diverse and sometimes contradictory, the main difficulty being to obtain comparable sequence datasets and to define comparison criteria that can be applied to all methodologies. Nevertheless, some main trends can be compiled from these studies. Tree-based methods tend to be more sensitive than graph-based methods, but at the expense of some specificity loss. Their evolutionary framework permits to describe distant paralogy and co-orthology relations. However, searching for distant relations has a huge computational cost. Graph-based methods are a trade-off between sensitivity and computational time, with relatively similar predictions for all methods. Here, wrong predictions are mainly due to gene loss and domain recombination events. In a recent paper, Boeckmann & al. concluded that none of the databases provides a fully correct and comprehensive protein classification, but that tree-reconciliation and hierarchical clustering-based methods have the potential to correctly describe a gene phylogeny (Boeckmann et al., 2011). Another work from Hellmuth & al., focusing on a mathematical description of orthology in a graph theory framework, concluded that graph-based methods could improve their predictions with better mathematical modelling (Hellmuth et al., 2012).
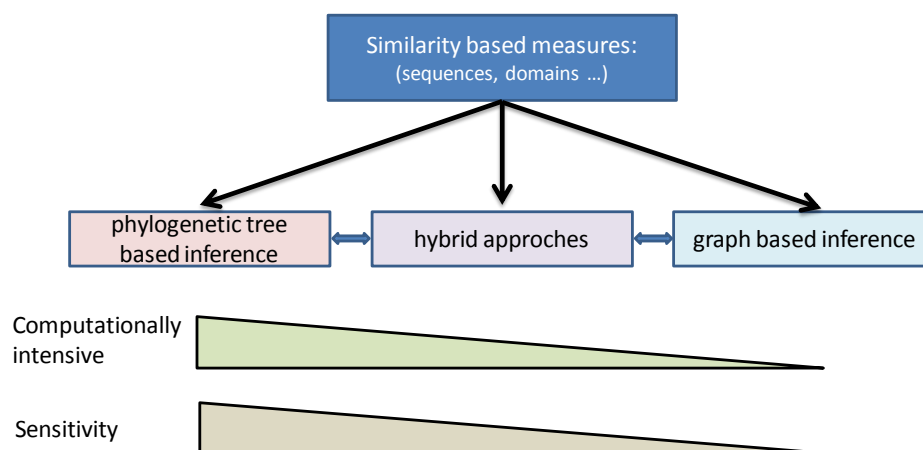
Figure 3-4. Main trends in performance of orthology prediction methods.
Phylogenetic-based approaches are generally considered to be more sensitive and can handle complex genetic events such as gene loss or domain recombination (if the phylogeny is domain-based). However, this choice is computationally intensive and is not suitable for very large-scale studies.

These general conclusions have been drawn from comparative studies based on heterogeneous datasets, representing heterogeneous species sets and protein families. There is clearly an urgent need for more standard evaluations using 'gold standard' community benchmarks of orthologs and for a clearer definition of orthology assessment criteria. Indeed, orthology being a homology relation, in general it is impossible to know *a priori* if the inferred evolutionary relation is correct or not. Therefore, such a relation can only be supported by multiple cross validations. The most reliable validation is to compare concordance of orthology predictions with reference trees of well-known protein families (phylogenetic assessment). However, the lack of such trees for hundreds of species, as well as their inconsistency for less well studied phyla, limits such a validation for genome-scale data. Chen et al. proposed two other comparison criteria based on the functional conjecture of orthologs: the consistency of protein function in orthologous groups and the consistency of domain architecture in orthologous groups (Chen et al., 2007). Again, these criteria (mainly based on GO annotations) can only be applied in well annotated gene families and assumes that the analyzed gene families have not experienced major genetic events, such as domain recombinations. Altenhoff et al. proposed three other criteria: similar Enzyme Classification (E.C.) annotations in orthologous groups, correlation in expression profiles and gene neighborhood conservation (Altenhoff and Dessimoz, 2009). Except for neighborhood conservation, these criteria are again strongly based on the functional conservation between orthologs. Judging orthology prediction reliability with E.C. numbers is dangerous, as co-orthologs may undergo subfunctionalisation or neofunctionalisation. It is just as dangerous to use expression profiles, because expression is also regulated by genomic context and epigenetic factors. The problems associated with the useful assessment of orthology predictions have been discussed recently, in particular in the context of the Quest for Orthologs Consortium (Gabaldon et al., 2009), and efforts are now underway to address these issues in the orthology community (see 3.4.2).

3.3   Integration Efforts

3.3.1   Combining different approaches

As described in the previous sections, each of the conceptual approaches for orthology inference has its advantages and drawbacks and the user is forced to choose between sensitivity and specificity. To exploit the advantages of the different methods, several authors have developed tools combining several alternative predictions. The idea is that the intersection of genes predicted as (co-)orthologs by several approaches can be used to reliably identify a set of 'true' orthologs or to describe a complex evolutionary scenario. Such comparisons highlight a need for efficient visualization tools, facilitating the analysis of the intersections/exclusions of multiple ortholog datasets.

In 2005, the HGNC Comparison of Orthology Predictions (HCOP) (Seal et al., 2011) was the first database to integrate orthology predictions from different orthology-dedicated resources. Today, it integrates human gene orthology predictions from EnsemblCompara, Homologene, Inparanoid, OMA, Treefam and the model organism databases, Evola (Matsuya et al., 2008), MGI (Eppig et al., 2012), OPTIC (Heger and Ponting, 2007), UCSC (Dreszer et al., 2012) and ZFIN (Bradford et al., 2011). In HCOP, the reliability of an orthologous relation is based on the number of occurrences in different

databases and synteny information, but this information is not supported by any visualization tool or informative scoring scheme. More recently, the authors of the ProGMap website (Kuzniar et al., 2009) integrated predictions from COG, OrthoMCL and HomoloGene databases. They developed a graph-based tool to describe the correspondences and discrepancies between the predictions made by these methods (Figure 3-5). By constructing a network of links using a fast hashing/mapping method originally developed for document classification, they constructed a graph for each protein family describing the relationships existing between the various data sets.
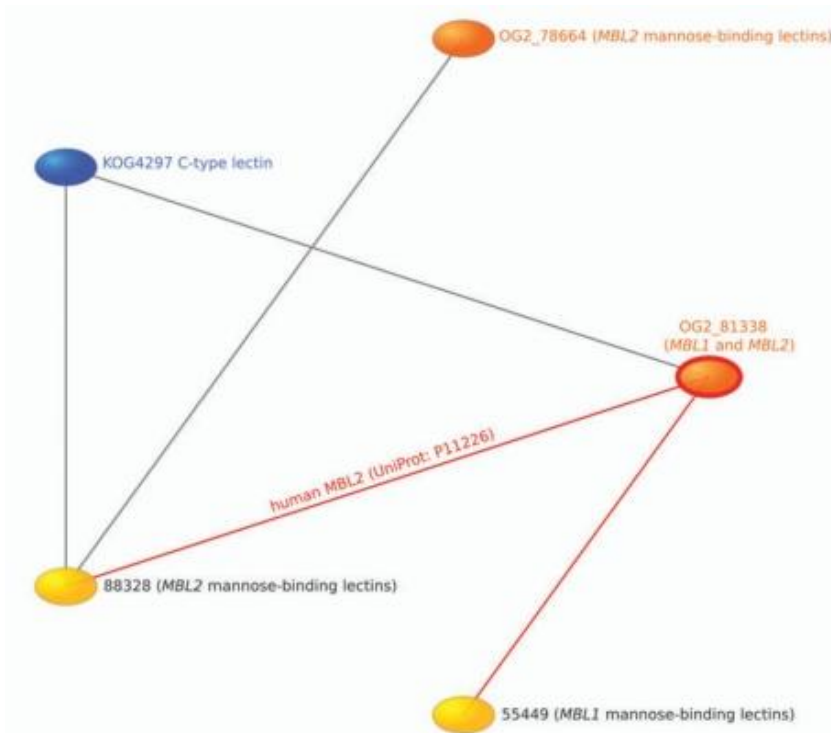
Figure 3-5. Comparing protein ortholog groups using the ProGMap network visualization tool. *This screenshot describes the orthologous relationships between five mannose-binding lectins. Groups sharing at least one protein are connected with an edge. In this example, the HomoloGene database (yellow) divides the lectins into two groups. KOG (blue) regroups them into one group and OrthoMCL (orange) separates them in a different way. Adapted from Kuzniar et al., 2009.*

A second example of integration can be found in PLAZA (Van Bel et al., 2012), a platform dedicated to comparative genomics in plants. As for ProGMap, PLAZA combines a graph-based method with a phylogeny based method. It uses orthologous groups based on OrthoMCL and compiles them with in-house phylogenetic predictions, gene neighbourhood conservation and inparalogy predictions using an OrthoInspector-like method. Then, a weighted voting scheme based on the sensitivity of individual tools is used to estimate orthology predictions. An online tool allows the visualization of paralogous sequence groups and describes which relations of a group are supported by several methodologies (figure 3-6).
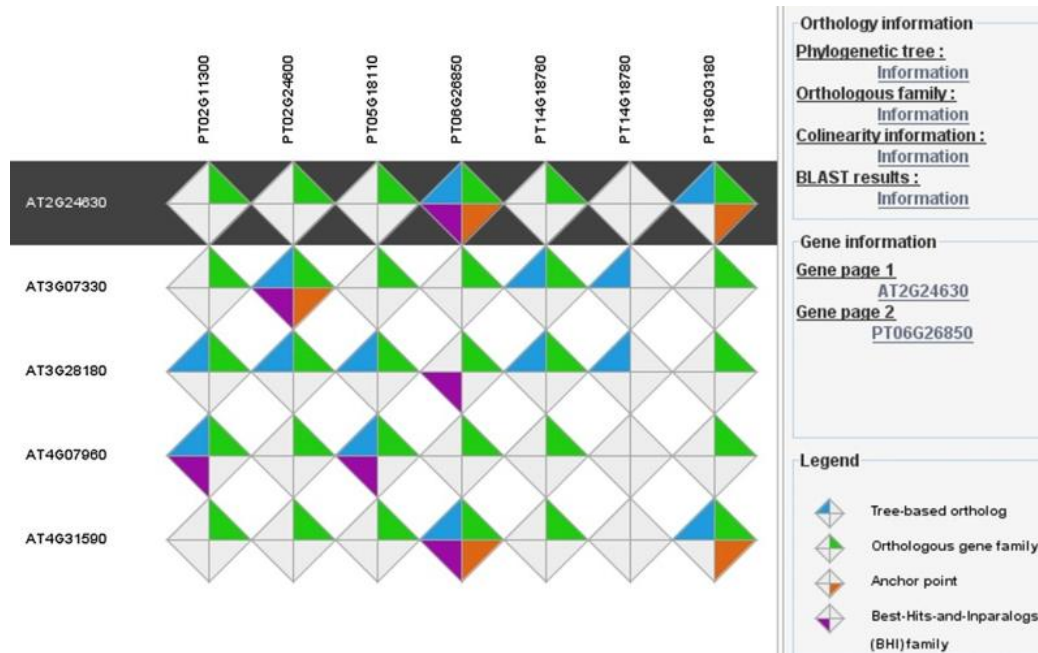
Figure 3-6. Integrative Orthology Viewer of the PLAZA platform. *The platform provides orthology relations between plant genomes based on OrthoMCL (green), phylogenetic inference (blue), gene neighbourhood (orange) and OrthoInspector co-orthologs (purple). Adapted from Van Bel & al. 2012.*

Another database integrating multiple orthology predictions is MetaPhOrs (Pryszcz et al., 2011), a repository of phylogeny-based orthologs, combining resources from 7 phylogenetic orthology databases (PhylomeDB, EnsemblCompara, EggNOG, OrthoMCL, COG, Fungal Orthogroups, and TreeFam). MetaPhOrs predictions are based on a pipeline that uses the phylogenetic trees of the 7 source databases, but not their final orthology predictions. For any given pair of sequences, all phylogenetic trees containing these sequences are retrieved. Then a species overlap algorithm is used on each tree to predict the type of homology relationship existing between this sequence pair. The MetaPhOrs authors performed an interesting comparison of their results with several tree-based databases, estimating specificity and sensitivity using a fungal ortholog benchmark. Currently, this is the only study demonstrating that method integration could provide better sensitivity and specificity for orthology predictions.

3.4 Unifying orthology research efforts: achievements and perspectives

3.4.1 Quest for Ortholog Consortium: a recent community initiative

In 2009, some authors of the most perennial orthology databases decided to organize a 'Quest for Orthologs' meeting to discuss and address major limitations and the perspectives for orthology inference (Gabaldon et al., 2009). This meeting included experts in genome evolution, developers of orthology prediction methods and curators of general databases with a common goal of discussing

current orthology inference issues and main challenges. This initiative gave rise to the Quest for Ortholog Consortium and motivated a second meeting that gathered numerous research teams interested in orthology research (Dessimoz et al., 2012). Several orthology inference challenges highlighted by this consortium have been discussed in the previous sections and the two first results of this community effort are described in the following chapters.

3.4.2   Benchmarking

In computing, a benchmark is the act of performing operations, in order to assess the relative performance of a program by running a number of standard tests and trials against it. In the case of orthology benchmarking, the goal is to construct a dataset of reliable orthology relations and to test whether the predictions produced by an algorithm correspond to this 'gold' standard. Until recently, no specific benchmark existed for orthology relations; most authors produced their own benchmark with the publication of their methods, which meant that global comparisons between methods were very difficult.

Thanks to the discussions initiated during the Quest for Orthologs meetings, significant efforts have been made during the last year to produce standard community benchmarks accepted by most research teams (publication n°5). An orthology benchmark should contain gene families representing heterogeneous evolutionary scenarios. Trachana & al. published a comprehensive orthology benchmark, representing 70 protein families, classified by biological characteristics (speed of Evolution, low complexity regions/repeats, domain shuffling/evolution, multigene families/paralogy) and technical characteristics (low/high alignment quality) (Trachana et al., 2011). A useful benchmark should also contain standard reference proteomes, with one representative sequence for each gene. This was previously done independently by each author. The EBI and in particular, the Uniprot database, now provide 'complete proteomes' (all non-redundant transcripts for the protein-coding genes) and 'reference proteomes' (one representative transcript per protein-coding gene) for all completely sequence organisms, offering a common dataset for all orthology databases. A specific sub-set of species based on these reference proteomes has been constructed to test orthology methods (http://www.ebi.ac.uk/reference_proteomes/) and an online benchmarking service for the Quest for Orthologs Consortium has been developed (http://linneus54.inf.ethz.ch:8080/cgi-bin/gateway.pl).

3.4.3   OrthoXML, an orthology ontology

The large number of diverse algorithms and databases focusing on orthology relations has led to an increasing number of data formats, hindering efforts towards their integration. One solution to increase the interoperability of the data was to create an ontology dedicated to orthology relations. In computational sciences, an ontology is a data model, i.e. a formal, structured representation of the knowledge in a particular domain. Two years ago, Ostlund & al. introduced OrthoXML, a standardized data exchange format dedicated to the representation of orthologous relations (Ostlund et al., 2010). This standard was designed to support both graph-based and tree-based

definitions of orthology in a consistent way (figure 3-7). The OrthoXML format is now integrated in several major orthology databases, and is supported by several libraries and regular updates (http://orthoxml.org/).

```xml
<?xml version="1.0" encoding="utf-8"?>
<orthoXML xmlns="http://orthoXML.org/2011/" version="0.3" origin="orthoXML.org" originVersion="1">
  <species name="Homo sapiens" NCBITaxId="9606">
    <database name="someDB" version="42">
      <genes>
        <gene id="1" geneId="hsa1"/>
      </genes>
    </database>
  </species>
  <species name="Pan troglodytes" NCBITaxId="9598">
    <database name="someDB" version="42">
      <genes>
        <gene id="2" geneId="ptr1"/>
      </genes>
    </database>
  </species>
  <species name="Mus musculus" NCBITaxId="10090">
    <database name="someDB" version="42">
      <genes>
        <gene id="3" geneId="mmu1"/>
        <gene id="4" geneId="mmu2"/>
      </genes>
    </database>
  </species>
  <groups>
    <orthologGroup>
      <orthologGroup>
        <geneRef id="1" />
        <geneRef id="2" />
      </orthologGroup>
      <paralogGroup>
        <geneRef id="3" />
        <geneRef id="4" />
      </paralogGroup>
    </orthologGroup>
  </groups>
</orthoXML>
```


Figure 3-7. Example of an orthologous relationship in the OrthoXML format. *The format identifies the 'group' level, including all co-orthologs relative to a specific taxonomic level and the orthologs/paralogs composing this group.*

4   FROM GENE CENTRIC BIOLOGY TO SYSTEMS BIOLOGY

A biological system is no longer considered as a simple collection of components, but as a whole. Thus, holistic thinking is expanding in biology and is replacing the reductionist perspective, motivated by the idea that the functions of a given system can only be understood by considering the interplay between its components. This idea is quite old, since the term 'systems biology' was introduced by Ludwig von Bertalanffy at the beginning of the 20th century (von Bertalanffy and Woodger, 1933). Despite its age, the concept was only integrated in the wider biological community from the year 2000 onwards. Today, the majority of biological studies include a system biology approach.

Recent technological breakthroughs have made the production of genome-scale datasets more accessible for many laboratories. Such datasets can help to provide a more global picture of biological phenomena and the so-called 'omics' approaches finally provide us with the opportunity to consider biological systems from a systemic point of view. Biological studies now often compile genome-scale datasets representing several biological levels (genomics, transcriptomics, proteomics, interactomics, etc.). This multi-scale approach is contributing to the socio-scientific movement referred to as systems biology. We can compare this movement to the discovery of the structure of DNA and the subsequent breakthroughs in molecular techniques that emerged (PCR, Sanger sequencing, etc.). Biology is thus shifting from the gene-centric to the systemic point of view.

4.1   Defining systems biology

4.1.1   A philosophy more than a research field

It is currently difficult to provide a clear definition of 'systems biology'. Looking at the scientific literature, various definitions have been proposed depending on the author's sensitivity. The four following examples illustrate these differences:

- *"Systems biology studies biological systems by systematically perturbing them (biologically, genetically, or chemically); monitoring the gene, protein, and informational pathway responses; integrating these data; and ultimately, formulating mathematical models that describe the structure of the system and its response to individual perturbations."* (Ideker et al., 2001)
- *"To understand complex biological systems requires the integration of experimental and computational research — in other words a systems biology approach."* (Kitano, 2002)
- *" […] the objective of systems biology [can be] defined as the understanding of network behavior, and in particular their dynamic aspects, which requires the utilization of mathematical modeling tightly linked to experiment."* (Cassman and Center, 2007)
- *"By discovering how function arises in dynamic interactions, systems biology addresses the missing links between molecules and physiology. Top-down systems biology identifies molecular interaction networks on the basis of correlated molecular behavior observed in*

genome-wide "omics" studies. Bottom-up systems biology examines the mechanisms through which functional properties arise in the interactions of known components." (Cassman and Center, 2007)

Finally, a more general definition is given by Wikipedia (en.wikipedia.org/wiki/Systems_theory):

- Systems biology is a term often used to describe a number of trends in bioscience research, and a movement which draws on those trends. Proponents describe systems biology as a biology-based inter-disciplinary study field that focuses on complex interactions in biological systems, claiming that it uses a new perspective (integration instead of reduction). [...] Systems biology refers to a cluster of peripherally overlapping concepts rather than a single well-delineated field. However the term has widespread currency and popularity as of 2007, with chairs and institutes of systems biology proliferating worldwide.

Given such heterogeneous (but overlapping) ways of thinking, one cannot hope to define a consensual definition for systems biology. Nevertheless, a major aspect is conserved in all these definitions despite the diversity of opinions: a system-level study considers all the components of a system by integrating information about all biological levels. The systems biology related concepts discussed in the following sections will focus on three specific aspects relevant to the results described in chapter 8 and 9. First, the main omics approaches that recent technological breakthroughs have made available for current systems biology studies are resumed. Then, the subject of multi-scale data integration is discussed. Finally, a large part of this chapter focuses on the importance of biological network construction and representation in systems biology, including the current bioinformatics resources available to analyse biological pathways. All these subjects highlight how current biology is moving from a bottom-up to a top-down strategy for resolving systems architecture, functional properties and dynamics.

4.1.2 Systems biology and systems sciences

In systems biology, the large amount of data and their integration repesents an opportunity for the modelling of biological systems properties. The detailed knowledge produced for some particular biological networks, mainly metabolic pathways and signaling networks, has motivated the idea that biological systems can be represented mathematically (Feist et al., 2009).

Such studies are generally based on a cyclic protocol of three steps: data integration, construction of a model and *in vivo* quantitative measures, followed by model refinement based on the newly collected data (figure 4-1). The studies are therefore multi-disciplinary and require different expertise such as chemical kinetics, control theory and mathematical modelling. Moreover, cellular networks are generally characterized by a large number of parameters and constraints. Consequently, network modelling is supported by computational techniques from graph theory to describe the behaviour of the systems. The computationally modelled systems can be used for numerous purposes such as bioengineering. For example, a biological pathway such as the biosynthesis of valine and leucine can be mathematically modelled to predict flux patterns of this pathway (Dreger et al., 2009; Lee et al., 2012). The modelling aspect inherent to numerous systems biology studies, in particular the

mathematical modelling aspects, is not the focus of this thesis and will only be briefly mentioned in the following paragraphs.
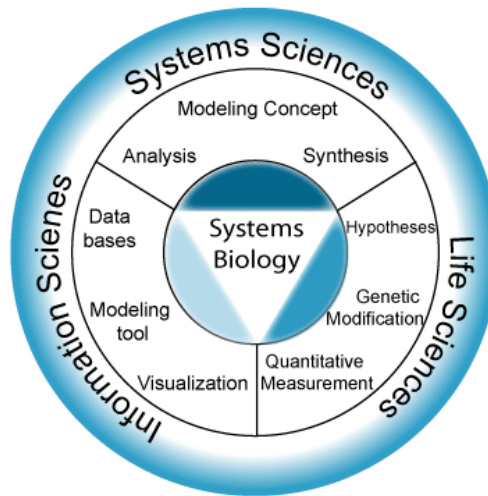
Figure 4-1. The 'modelling' view of systems biology. *The use of dynamic systems theory is applied to molecular biology and supported by computer sciences. In this view, dynamics is the conceptual difference between 'systems biology' and classical bioinformatics.*

4.2   'Omics' and multi-scale perspectives

4.2.1   Producing global pictures of biological systems

The 'omics' suffix is generally employed to describe the generation or study of genome-scale datasets. These datasets represent the high-throughput identification of large sets of representatives of a particular group of molecules or other biological objects. Omics approaches are now applied to all biological levels, ranging from the genetic to the ecological level. Omics studies can be divided into three main categories: the analysis of molecular entities (DNA, RNA, histone methylation states, proteins, molecular networks…), the study of biological communities (microbiomics, biomics…) and the medical sciences (pharmacogenomics, toxicogenomics…). In the following paragraph, only the molecular omics will be considered.

Molecular omics mainly focuses on the inventory of all representatives of biological molecules. This inventory depends on the biological sample and its spatio-temporal characteristics (cell type, tissue, culture condition, developmental state…). Recent technological advances, such as Next Generation Sequencing (NGS), Mass Spectrometry (MS) or Nuclear Magnetic Resonance (NMR) facilitate the deployment of molecular omics strategies and provide many new opportunities for building an integrated description of complex biological systems. Figure 4-2 shows a non-exhaustive list of molecular omics, ranging from the molecular to the biological system level.
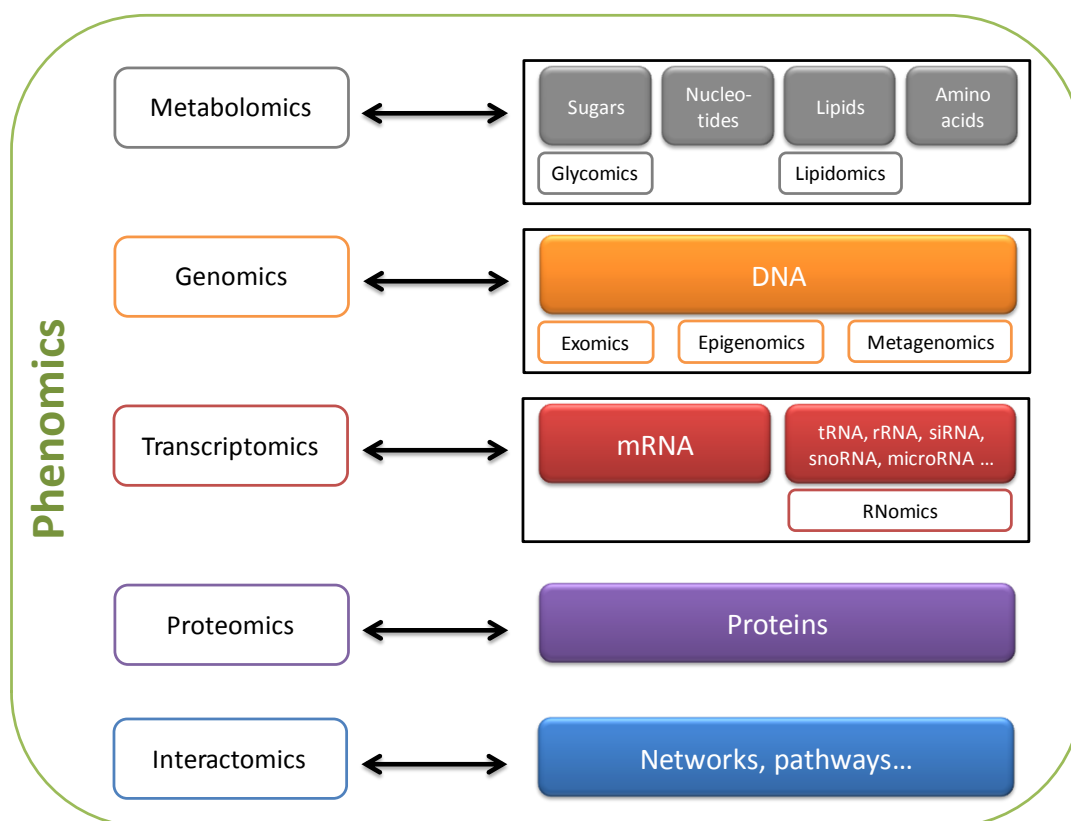
Figure 4-2. Overview of main molecular omics in the systems biology era.

The system level, represented by interactomics, is extensively described in paragraph 4.3 through the example of biological networks, one the focuses of this thesis. In the following sections, other omics are briefly described by regrouping them into small molecule, DNA, RNA and protein-level categories:

4.2.1.1   small molecule omics:

Several omics approaches have been developed to study the chemical compounds that are used as substrates, products or cofactors by amino acid or nucleotide polymers. For example, metabolomics focus on the cellular metabolites that represent the chemical fingerprints of a specific cellular process. Metabolomics is still an emerging field as current technologies cannot provide a full collection of all the metabolites of a cell. The chemical complexity of biological samples limits metabolomic analyses to the quantification of a portion of the metabolome by mass spectrometry (MS) or the overview of metabolite compositions by nuclear magnetic resonance (NMR) (Patti et al., 2012). However, major advances have been made in two specialized metabolomic fields, namely glycomics and lipidomics. A glycome describes a collection of sugars in a cell or an organism for example, whether they are free or present in more complex molecules. A lipidome describes an equivalent collection of lipids. Both fields have experienced a large expansion due to technological advances in mass spectrometry (MS), high performance liquid chromatography (HPLC) and nuclear magnetic resonance (NMR) (Aoki-Kinoshita, 2008; Wenk, 2005).

4.2.1.2   DNA-related omics:

Genomics includes a large set of techniques to study the genomes of organisms. Genomics studies are based on DNA sequences and their by-products: genes, regulatory elements, protein products, etc. This vast field can be decomposed into three categories. First, comparative genomics focuses on the relationship between genome structure and function across different biological species or strains. It is generally supported by high-throughput sequencing techniques and well-implemented bioinformatics applications such as sequence alignment or phylogeny construction (Xie et al., 2005). Second, functional genomics studies the functions and interactions of genes or proteins. It is generally supported by high-throughput experimental techniques such as microarrays, siRNA screening, large-scale mutagenesis or SAGE (Pevsner, 2009a). The last genomics sub-field is metagenomics. This domain focuses on the environmental aspects of genetic material. Its goal is to understand how genes can be linked to the physico-chemical and biological properties of an environment and how genes can be viewed as fingerprints of biological diversity (Marco, 2011). Functional genomics and metagenomics are also well supported by bioinformatics techniques, such as homology predictions, phylogenetic profiling or protein domain searches to predict biological functions (Parks and Beiko, 2010).

While genomic sequences provide the basis for many molecular studies, they cannot represent the dynamic DNA modifications that occur in biological systems. The analysis of these modifications has recently led to the creation of specialized genomics fields. For example, epigenomics attempts to identify the complete set of epigenetic modifications of specific genetic material, mainly DNA methylation and histone modifications in eukaryotes. The epigenome datasets are mainly produced using ChIP-Chip and ChIP-Seq technologies (Laird, 2010).

4.2.1.3   RNA-related omics:

Transcriptomics concerns the analysis of large sets of RNA molecules, including mRNA, rRNA, tRNA, and non-coding RNA. It should be noted that cellular transcriptomes are highly heterogeneous since they depend on developmental, environmental and physiological conditions. Individual transcriptomes were previously produced by hybridation-based or tag-based sequencing approaches. Today, the high-throughput RNA-seq sequencing technique is preponderant (Wang et al., 2009). A specific field called RNomics is dedicated to the analysis of small non-mRNAs (snmRNAs), also called non-coding RNA (ncRNA), especially small nucleolar RNAs (snoRNAs), microRNAs and small interfering RNAs. The field is still expanding as the cellular role of ncRNA is only beginning to be understood (Huttenhofer et al., 2002). In the case of mRNA, transcriptomics can be used to determine the transcriptional structure of genes (start sites, 5′ and 3′ ends, splicing, post-transcriptional modifications) and quantify their expression. Recently, the developments in high-throughput sequencing have also provided new opportunities for efficient analysis of the exon composition of a transcriptome. The corresponding dataset, an 'exome', can be used for example, for the analysis of intra-species genetic variation and its development is mainly motivated by promises of new 'personalized' medicine (Ng et al., 2008).

4.2.1.4   protein-related omics:

Proteomics approaches are used to perform large-scale studies of protein structures and functions. The field actually covers very different approaches. For example, proteomes are used to study post-

transcriptional modifications such as ubiquitination, phosphorylation, methylation, acetylation, etc. For this purpose, protein sequence data are generated by mass spectrometry and modifications are highlighted by a bioinformatics comparison with genomic data (Reumann, 2011). Structural genomics is more focused on the structural properties and the molecular dynamics that characterize proteins and their interactions. This domain combines biophysical approaches, bioinformatics and molecular modeling to achieve this goal (Ghersi and Sanchez, 2011). Proteomics approaches can also be combined with genomics and bioinformatics for the simultaneous refinement of multiple genome annotations (Gallien et al., 2009).

4.2.1.5   Phenomics

This list of molecular omics can be complemented by 'phenomics'. Phenomics is used to describe the molecular, physiological and physical traits of an organism and their variations after perturbations. Consequently, a phenome can be built by combining the different omics approaches and grouping the molecular information. RNA interference is also used to produce libraries of phenotypes for the same species. Nevertheless, the main challenge of phenomics is the development of high-throughput measurement systems and the automatic integration of many omics parameters (Vankadavath et al., 2009).

4.2.2   Multi-level data integration

After a decade of developments, several omics fields are now shifting their main focus from technological development to exploitation of the datasets produced. For example, automatic transcriptomic protocols are now routinely used in the medical field for molecular or genetic diagnosis (Lamberts and Uitterlinden, 2009). The ever cheaper cost of omics is even opening the way for a direct-to-consumer diagnosis market, giving rise to new ethical and political debates (Caulfield and McGuire, 2012). The variety and complementary of different omics allow more extensive descriptions of biological systems. Nevertheless, despite the mature state of omics, the direct integration of multiple omic datasets representing different biological levels remains challenging. Strategies for integration are not keeping up with the technological developments. In some fields such as genomics, online databases contain so much data and are being updated so fast, that they can no longer be managed by current methodologies. With the rapid progress observed in all omics fields, this situation will soon become a reality in most biological domains.

A major problem concerns the quality of omics data. Their high-throughput nature means that they generally produce a non-negligible quantity of noise. For the older high-throughput techniques such as micro-arrays, various routine approaches are now used to limit this drawback (Raffelsberger et al., 2008; Zhang et al., 2009). For more recent technologies such as high-throughput sequencing, approaches for data quality assessment are still being developed (Bravo and Irizarry, 2010; Nothnagel et al., 2011). The dangers of producing more data to the detriment of data quality have been evoked on many occasions. For example, several studies showed the impact of low coverage genome sequencing on subsequent analyses (Milinkovitch et al., 2010; Prosdocimi et al., 2012).

Data management is a second major challenge in the omics era. The sharing of large datasets requires that similar data definitions should be used by the different data owners. To resolve this problem, several community efforts were initiated to create biological and biomedical ontologies (Smith et al., 2007; Thompson et al., 2005a). The data hierarchies and standard definitions defined by ontologies facilitate the integration of omics datasets. Based on these ontologies, several database engines or data warehouses have been developed for efficient retrieval of biological information (Kasprzyk, 2011). However, ontologies are often limited to the description of one type of entity at a time (DNA, multiple alignment, protein network…). Handling the wide variety of biological data in a single management system require lots of computational development. A few alternative approaches have been developed in an attempt to address this drawback. For instance, the database design of a high-throughput microscopy approach was adapted to biological data, allowing day-to-day evolution with no need for rigid standards (Millard et al., 2011). Another idea is to pre-compute database content graphs representing different views upon the data. The views can then be selected to rapidly regroup new datasets and to more efficiently answer new biological questions (Bard et al., 2010; Boyle et al., 2009).
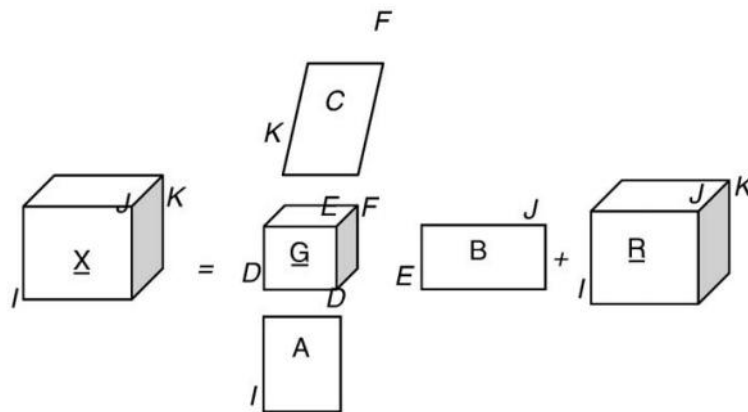
Figure 4-3: The 'multi-way' framework, an example of a multidimensional data integration and correlation analysis. *A 3-way data structure X (I×J×K), is decomposed in several matrices. A (I×D), B (J×E) and C (K×F) represents the ways in which the information related to the samples can be compressed; G (D×E×F) is the core matrix, which explains how components of the different modes are related; R (I×J×K) is the residual matrix. G can be assumed to be an approximation of X, summarising the variation in X in terms of matrices A, B and C (adapted from* Conesa et al., 2010*).*

Nevertheless, the main challenge for systems biology is to develop efficient methodologies for the integration and selection of the most relevant data for a specific biological question. Many databases integrate multiple sources of data, but few authors provide an integration framework that facilitates the analysis of potential relationships between the data, such as correlations or divergences. The literature describes only a few developments addressing large-scale integration. For example, Hwang et al. developed a methodology that can handle multiple data sets differing in type, size, and coverage for biological networks. This methodology minimizes the number of false positives and false negatives (Hwang et al., 2005). Another interesting approach was proposed by Conesa et al. (figure 4-3). They proposed a 'multi-way' approach, a technique performing a dimensional reduction of

multidimensional data structures in such a way that relationships between and within dimensions can be extracted and analysed (Conesa et al., 2010). Similarly, statistical component analysis can be used to describe global correlations of biological datasets (Wolf et al., 2006). Another multivariate strategy was developed by Bylesjo et al. for the integration of plant transcriptomics, metabolomics and proteomics data through the O2PLS methodology, a regression method separating predictive variation from the variation that is unique to each platform as well as residual variation (Bylesjo et al., 2007).

Data integration and summarization is often supported by powerful visualization tools. For example, Srinivasan et al. developed a Bayesian approach to integrate co-expression, co-inheritance, co-location and co-evolution data and to build the statistical interactomes of 11 species (Srinivasan et al., 2006). Their pipeline is complemented by a 3D representation of statistical correlations between the datasets (figure 4-4). Similarly, Secrier et al. developed a 3D visualization and integration package for the exploitation of temporal and tissue-related patterns (Secrier et al., 2012). Finally, Nguyen et al. developed a semantic map projection method to summarize biological annotation information in large datasets (Nguyen et al., 2009).
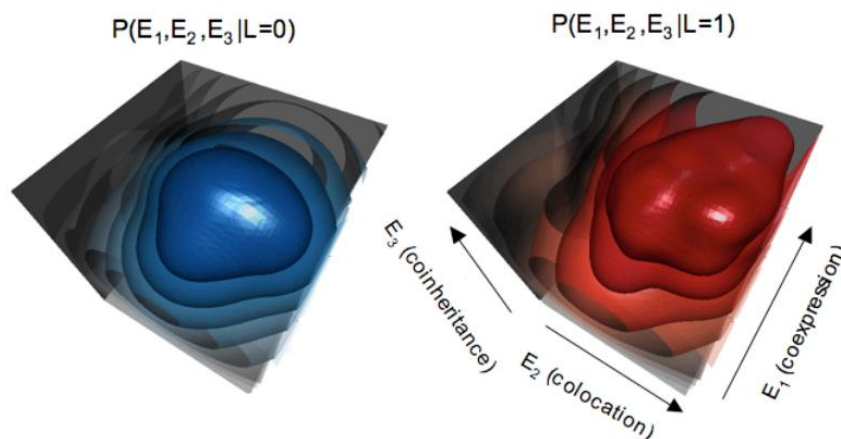
Figure 4-4. An example of innovative visualization for the integration of 11 microbial omics datasets. *When comparing protein interactions over several species, protein pairs can be functionally unrelated (L=0, left 3D surface) or share the same functional pathway (L=1, right 3D surface). Transparent to opaque level sets describe posterior probabilities of similar functional interaction and are spaced at even volumetric increments, so that the inner most shell encloses 20% of the volume, the second shell encloses 40% and so forth. 3D surfaces reveals that functionally linked pairs (red, L = 1) tend to have higher coexpression and coinheritance than pairs that participate in separate pathways (blue, L = 0). Adapted from Srinivasan et al., 2006.*

These examples reflect the need for efficient tools to filter, store and integrate very large heterogeneous biological datasets. Despite being an advanced subject of research in fields such as geospatial imagery (Chen et al., 2003) or human imagery (Keator et al., 2008), developments are still in their infancy in biology. This assessment can be linked to the trends in data growth in these fields.

Survey telescopes now produce petabytes (PB) of data, while genome-sequencing machines can currently produce 1 TB per week of information. However, this latter will probably grow to terabytes per day soon and simulations could easily produce petabyte-scaled information (Council, 2010).

In contrast to global integration, an alternative philosophy involves targeting the more pertinent information in a dataset when dealing with a particular biological question. This approach is generally referred to as 'gene prioritisation' (Tranchevent et al., 2011). The objective is to identify the best candidate genes that could be implicated in a particular biological process or a disease. To handle the current amount of biological data, developments have been performed to automate this task. Gene prioritisation can be resumed in four steps: (i) the creation of a training set of known genes and the extraction of gene characteristics to establish a typical gene 'profile' that describes a particular biological process or a disease, (ii) the extraction of the same characteristics for the candidate genes, (iii) the comparison of the candidate characteristics with the training set profile, (iv) the classification of the candidates, according to a score estimating the consistency of the candidate genes with the profile. The software developed for gene prioritisation now use multiple criteria to characterise genes, including gene annotations, interaction data, evolutionary frameworks, text mining, etc. (Tranchevent et al., 2008).

4.3   A focus on biological networks

4.3.1   Representing life with networks

One of the first tasks requiring large scale integration of data from multiple biological levels was the description of biological networks. Network description started with the biochemical description of canonical metabolic networks. For instance, in 1957 the Krebs cycle was described and was later completed by the identification of implicated enzymes (Kornberg, 1987). Today, biological networks are used to represent many kinds of biological interaction, ranging from biochemical interactions in molecular biology to food webs in ecology. Networks have multiple advantages. They can represent the genome-scale complexity of interactions occurring in biological systems (Yu et al., 2008). Changes in their topology and parameters can be used to model the dynamics of biological systems (Rohwer, 2012). A comparative analysis of network topology can decipher the constraints and flexibility with which biological systems respond to their evolution (Babu, 2010b). They can even model higher biological perspectives such as ecosystem interactions. For example, food webs are used to model food competitions in ecosystems (Guimera et al., 2010), while neuroendocrinologic networks proved to be useful in understanding the evolution of social behaviour (O'Connell and Hofmann, 2011).

In molecular biology, genome-scale interactions can be generated for many cellular compounds (figure 4-5) (Zhu et al., 2007). Using elementary biological components and interactions, the bottom-up approach of network reconstruction is now widely used to represent systems behaviour and dynamics (Barabasi and Oltvai, 2004) and hundreds of biological networks have been described in model organisms (Hyduke and Palsson, 2010; Yu et al., 2008). Moreover, networks are a way to categorize systems of very different origin within a single framework and many comparative and topological studies have investigated the general properties observed in all biological networks (Zhu

et al., 2007). Network studies are now performed in various domains based on molecular data. In biomedical research, they can be used to describe how a network perturbation can induce disease-related responses (Vidal et al., 2011). In evolutionary studies, networks demonstrate the importance of the cellular context in protein evolution (Nehrt et al., 2011). Hundreds of databases have been developed to regroup different kinds of biological network data. A rapid overview on the website www.pathguide.org gives an idea of the diversity of these databases. Many of these databases focus on one model organism or on a specific kind of network (PPI network, regulatory networks…) or compile chemical dynamics data.
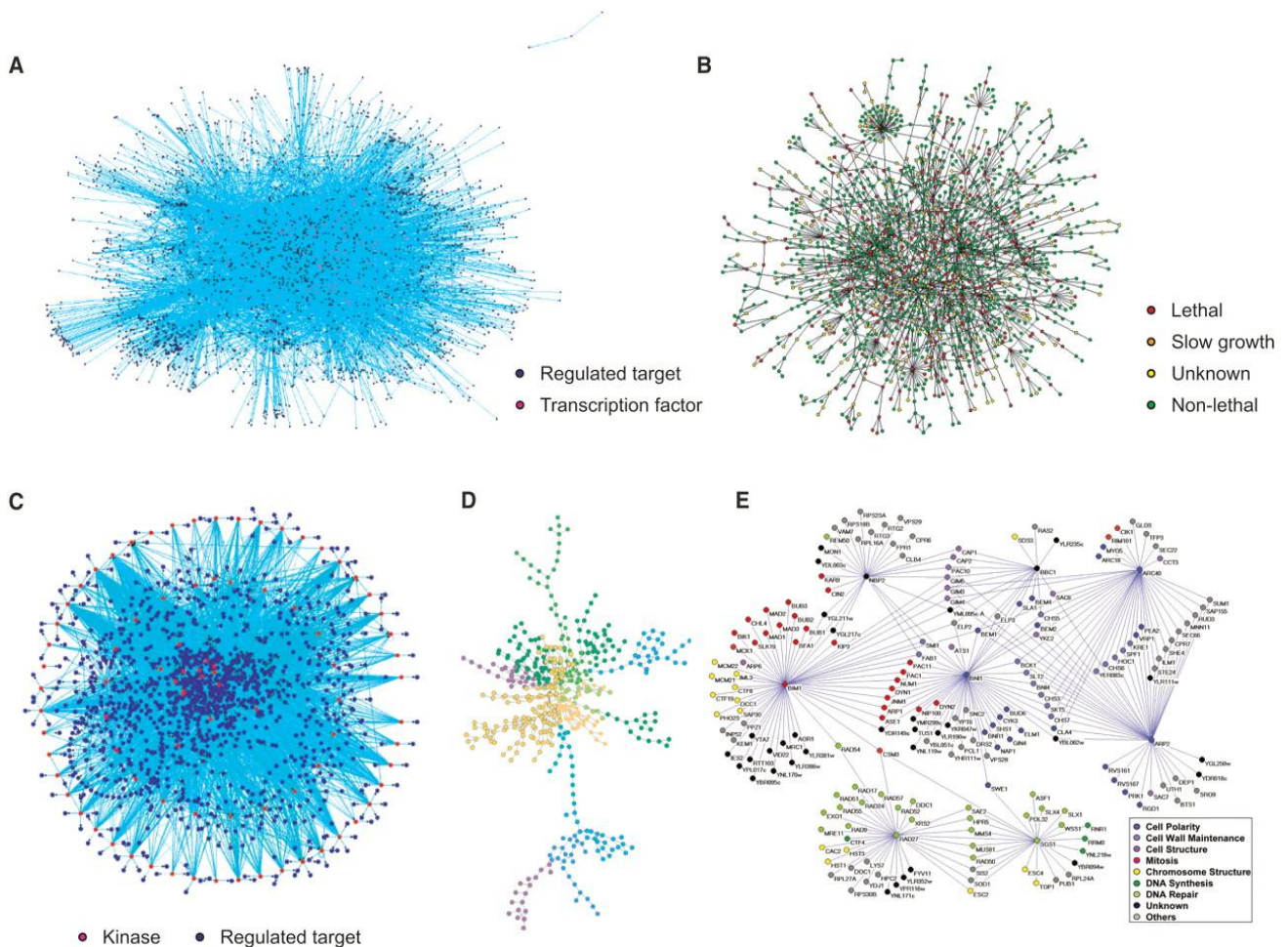
Figure 4-5. Overview of major networks in molecular biology. *(A) A yeast transcription factor-binding network built with large-scale ChIP–chip and small-scale experiments. (B) A yeast protein–protein interaction network identified by yeast two-hybrid, affinity purification and mass spectrometry approaches. (C) A yeast phosphorylation network identified using protein microarrays (D) An E. coli metabolic network with 574 reactions and 473 metabolites colored according to their modules. (E) A yeast genetic network constructed with synthetic lethal interactions. Adapted from Zu et al., 2007.*

4.3.2 The concept of biological pathways

Biological networks are now a routine tool and are used as often as biological sequence data. Any kind of interaction can be represented with a network, explaining their diversity. A specific concept found in many studies is the biological pathway, particularly what is known as the canonical pathway. A pathway description is generally considered as canonical when it corresponds to a succession of biomolecular events that are observed in many biological conditions. In other words, a canonical pathway can be considered as a central, often essential, network module of biological systems. Such a path represents a succession of interactions leading to the control of a particular biological phenomenon. For instance, metabolic pathways correspond to a succession of catalytic reactions ending in the production of a particular metabolite. Similarly, a pathway can correspond to the cascade of cellular interactions that controls the initiation of a specific event (mitosis, apoptosis, etc.). A canonical pathway can also correspond to a kernel of interactions that are implicated in many biological events. For example, the canonical MAPK signalling pathway describes a succession of kinase phorsphorylations influencing many other pathways and biological processes. It should be noted that this definition of canonical pathway is subjective as many metabolites or proteins of a pathway be implicated in many biological processes. Nevertheless, this concept helps the scientific community to consider a biological system as a patchwork of defined biological modules, one module being active under a specific condition.

The results presented in chapter IX take advantage of this modular view and are based on the KEGG PATHWAY database. This database includes many different concepts, such as metabolic reactions, gene regulation and protein-protein interactions. The pathways defined in KEGG are consequently a compilation of several types of interactions, reunited in a single map to describe all the aspects of particular biological phenomena (energy production, kinase signalling pathway, endocytose, etc.). We can distinguish interaction data corresponding to three kinds of biological networks in KEGG: metabolic pathways, regulatory networks and protein-protein interaction networks. KEGG integrates principles of these three biological networks in a single framework (Ooi et al., 2010).

4.3.2.1 Metabolic pathways

A metabolic pathway describes a succession of chemical transformations that occur in biological systems. This chain of reactions is catalyzed by a cascade of biological enzymes that transforms dietary minerals (glucide, lipids, protids), vitamins, and other cofactors into metabolites that can be used as cellular material or converted to chemical energy. Many metabolic pathways are constructed *de novo* by the compilation of experimental data and genomic annotations in a step-by-step process (Covert et al., 2001). The description of a metabolic network generally regroups genomic, biochemical and physiological data. Metabolic networks can use different representations depending on their topological or dynamic exploitation (Larhlimi et al., 2011). Figure 4-6 shows the 4 most common metabolic network representations. All representations correspond to the following set of chemical equations: $1B \rightarrow 1C$; $1D \rightarrow 2B$; $2B \rightarrow 1A + 1E$; $1A + 1E \rightarrow 1F$; $1F \rightarrow 2B$; $1B + 1C \rightarrow 1D$.

- The graph of complexes is the simplest representation (Deville et al., 2003). Reactions (edges) connect the substrate metabolite and the product metabolite complex (nodes).

- In a directed weighted hypergraph (Klamt et al., 2009) nodes represent the metabolites and the directed hyperedges represent the reactions. A weight representing stochiometry is generally assigned to the edges.
- A bipartite graph is a simplification of a hypergraph, where only physical interactions are shown without stochiometry. Two types of nodes are differentiated in two partitions: the set of metabolites and the set of reactions. Edges are directed depending on whether metabolites are substrates (metabolite→reaction) or products (reaction→metabolite).
- A metabolic network can also be coded in a stoichiometric matrix, representing the structure of a metabolic network in terms of relationships between metabolites and reactions with metabolites as rows and reactions as columns. Given a reaction, the corresponding column contains the negative of the stoichiometric coefficients of the substrates, the stoichiometric coefficients of the products and zero values for the remaining metabolites.
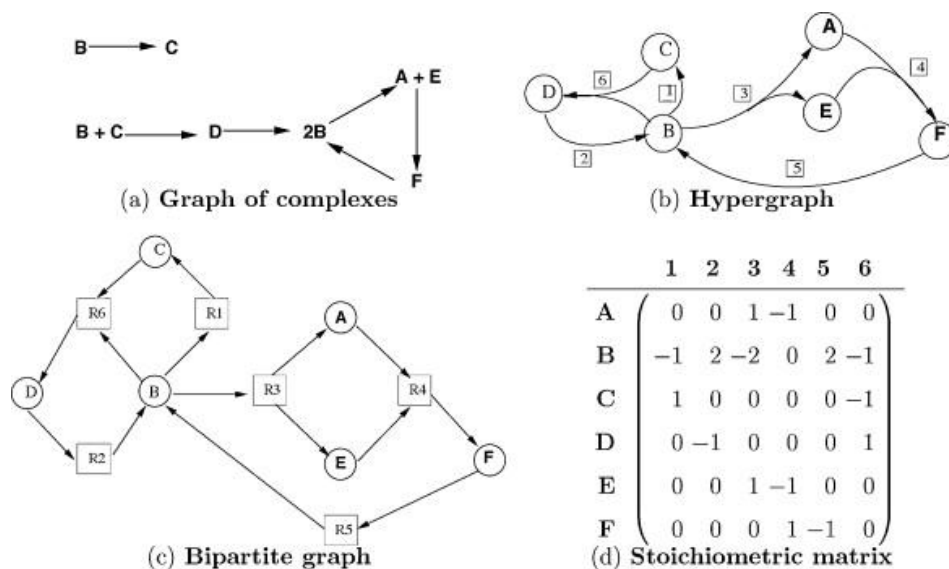
Figure 4-6. Several examples of metabolic network representations. *The biochemical reactions of a metabolic network can be represented in different ways. (a) graph of complexes. Each side of a metabolic reaction defines a complex; (b) hypergraph where hyperedges correspond to reactions and nodes represent metabolites; (c) bipartite graph where reactions are indicated by rectangles and metabolites correspond to circles; and (d) stoichiometric matrix. Adapted from Larhlimi et al., 2011.*

These representations (except bipartite graphs) consider the dynamic properties of metabolic networks. Indeed, the modelling of metabolic pathway inputs and outputs is one of their main applications. Once the metabolic network model is build, its dynamic properties can be simulated. There are two main classes of metabolic network analysis. First, flux balance analysis and metabolic flux analysis rely on experimentally quantifying a set of metabolic fluxes in the network by analyzing substrate consumption or product excretion (Dal'Molin et al., 2010). These approaches complement the topological model with a complete dynamic model by resolving regions for which stochiometry is unknown (Rohwer, 2012). Subsequently, enzyme kinetics can be introduced to model concentrations of substrate and product metabolites. Metabolic pathways can then be used as predictive models in several fields such as nutrigenomics or biological engineering (see chapter 4.4).

4.3.2.2 Regulatory networks

Gene regulatory networks (GRN) are network models where the inter-gene dependencies are described in a directed graph. Nodes represent genes and edges point from a regulator gene to its targets (Kim and Park, 2011). The edges ideally represent dependencies at the transcriptional level, generally a transcription factor (TF) and its influence on regulatory cis-element of the genes, this information being generally verified in wet-lab experiments. Consequently, the construction of a GRN requires three kinds of information: the spatio-temporal expression pattern of the TFs, the cis-regulatory modules they bind to and the causal link between the TF activity and the target genes' expression. Thus, reuniting experimental knowledge to construct edges and nodes in a reliable GRN is laborious and most GRN are small scale with a tradeoff between accuracy and completeness (Wilczynski and Furlong, 2010).

Small GRN are particularly used in developmental biology, which benefits from the spatio-temporal resolution of GRNs (Levine and Davidson, 2005). However, complete GRNs at the cellular level have been constructed, mainly for *Escherichia coli* and *Saccharomyces cerevisiae*, for which regulatory data from the literature and large-scale DNA-binding data from chromatin immunoprecipitation experiments are abundant (Gama-Castro et al., 2011). Several authors have also proposed algorithmic approaches to automatically generate GRNs from time series of expression profiles (Li et al., 2005). Large-scale GRNs can shed light on the organization of transcriptional regulation. For example, Babu showed that the basic unit formed by a TF and its target gene is organized into a limited number of motifs (figure 4-7) (Babu, 2010b). Linking the motifs highlights larger modules of interactions that are nested and interconnected through local regulatory hubs. The sum of all the modules composes the transcriptional regulatory network of a cell.
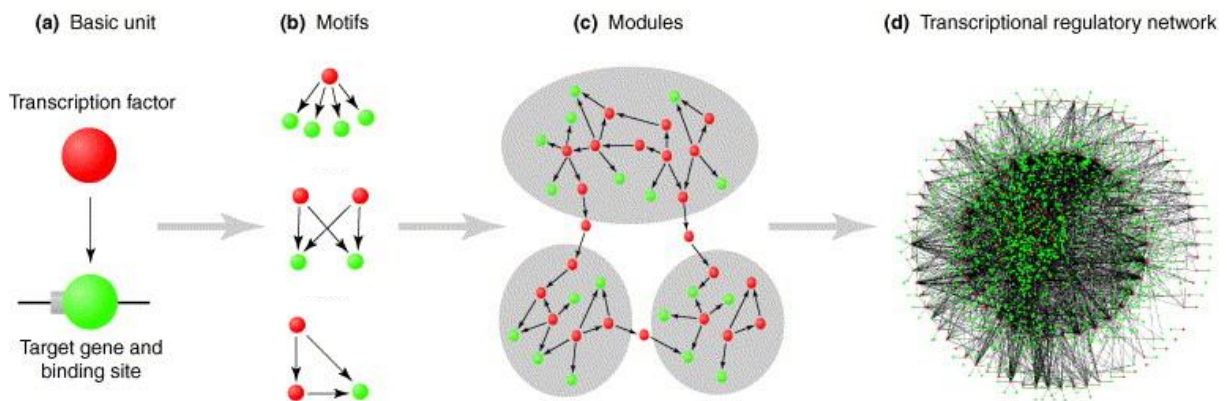
Figure 4-7. Structural organisation of transcriptional regulatory networks. *(a) The 'basic unit' comprises the transcription factor and its target gene. (b) Units are organised into network 'motifs', which comprise specific patterns of inter-regulation. (c) Network motifs can be interconnected to form semi-independent 'modules'. (d) The entire assembly of regulatory interactions constitutes the 'transcriptional regulatory network'. Adapted from Babu et al., 2010.*

Depending on environmental conditions, only a part of the complete GRN network is activated. GRNs are exploited by mathematical models capturing system behaviour. The input can be viewed as a set of TFs, while outputs are expression levels. Such modelling generally associates Boolean (AND,OR,NOT) or discrete functions with the gene nodes. This mathematical representation allows the simulation of cell behaviour and can be used to predict cellular response to environmental changes (if TFs associated to the response are known).

4.3.2.3   *Protein-protein interaction networks*

Protein–protein interactions (PPI) networks compile the physical protein interactions of biological systems, regardless of whether this interaction occurs in a stable protein complex or a transient interaction. PPI networks are essential in biological systems as they are the basis for several signal transduction and transcriptional regulatory networks. They were the first networks to be massively produced at the genome-scale, with the first complete interaction map being achieved in yeast (Gavin et al., 2002; Schwikowski et al., 2000), followed by *C. elegans* (Li et al., 2004), *D. melanogaster* (Giot et al., 2003) and human (Rual et al., 2005). Construction of PPI networks can be done through many computational and experimental approaches. Large datasets of new protein interactions can be discovered in yeast-two hybrid, affinity purification/mass spectrometry and protein microarrays experiments (Kaake et al., 2010; Panchenko and Przytycka, 2008). However, they present a low reproducibility and while identifying some PPIs with a high confidence, many false positives can be generated. Computational approaches can complement experimental approaches or discover new PPI at a lower rate (Raman, 2010). Genomic methods for the detection of gene fusions, conserved gene neighbourhood and phylogenetic conservation profiles can highlight new PPIs. For example, protein co-evolution was used to detect potential new PPIs in a phylogenetic framework (Juan et al., 2008). An original method of PPI inference based on large-scale text mining (He et al., 2009) was also developed. Despite this large choice of experimental and computational techniques, it is considered that our current knowledge covers only 30% of the yeast interactome and 10% of the human interactome, a fact mainly due to the time and context-dependent nature of interactions (Baker, 2012). A recent article highlights this lack and claims that only a few studies have proposed new approaches for differential mappings in biological networks (Ideker and Krogan, 2012). Figure 4-8 shows an example of differential mapping applied on an interaction network (Bisson et al., 2011).

PPI networks can be used for many purposes. First, the exploitation of the context of a protein in cellular networks allows new functional predictions. For example, the interolog approach helps to understand the function of uncharacterized proteins, such as their moonlighting aspects (Janga et al., 2011; Yu et al., 2004). Yet, this approach can be considered as dangerous because PPIs are thought to be more conserved within species than across species (Mika and Rost, 2006).

The topological aspect of networks is also actively exploited. The analysis of the impact of deletion of gene nodes or edges in PPI networks is useful for gene prioritization (Chang, 2009). Network hubs can be used for the detection and the understanding of a particular class of proteins: the transient proteins i.e. protein with low binding affinity, for which behaviour and dynamics are relatively unknown (Perkins et al., 2010). Computational methods have also been developed to decipher how biological information is transmitted between different protein network modules by information flow analysis (Missiuro et al., 2009).
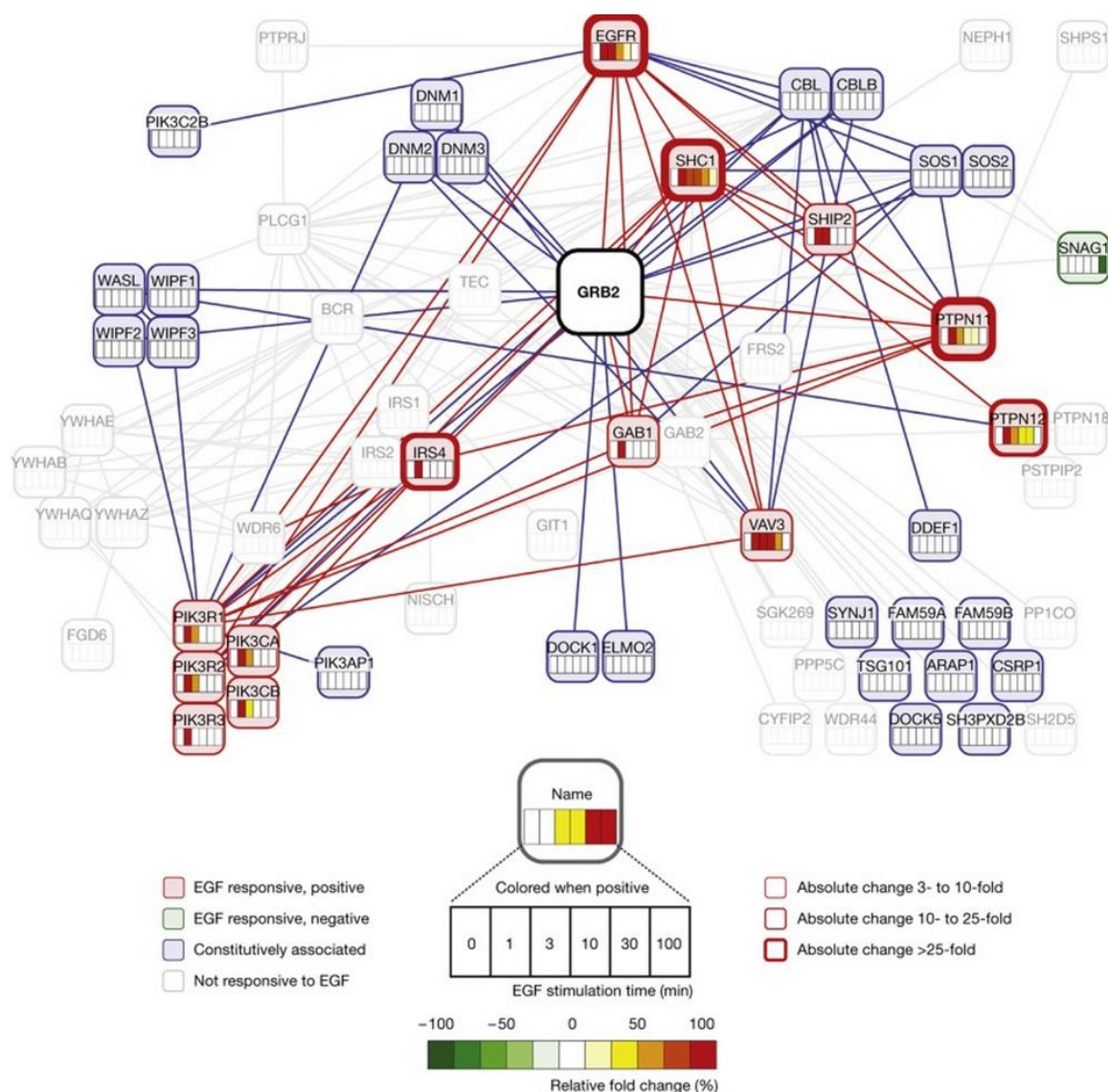
Figure 4-8. Example of a time-scaled differential mapping in a PPI network. *The diagram represents the dynamic protein interaction network involving GRB2, an adaptor protein involved in multiple aspects of cellular function. Red-shaded nodes represent proteins that are recruited to GRB2 complexes after epithelial growth factor (EGF) stimulation irrespective of time, green-shaded nodes those that are decreased and blue-shaded nodes those present in GRB2 complexes in nonstimulated cells. The thickness of the node border is proportional to the intensity of the change compared with control levels. Rectangles inside the nodes show the relative fold change for each time point. Adapted from Bisson et al., 2011.*

4.3.3 Biological network characterization

As briefly mentioned in the overview of metabolic, regulatory and PPI networks, topology is an essential aspect of network analysis. Topological properties of biological networks were mainly inferred on PPI networks because they were the first networks available at the genome scale. Many parameters can be used to describe network characteristics (Ideker and Krogan, 2012). We can

differentiate two topological levels of descriptions: a low-level topology, describing local network characteristics and a high-level topology, related to the global structure of the network. Generally, high-level topologies are characterised by their distinctive low-level parameters. Supplementary topological characteristics linked to network evolution (network modules, robustness, evolvability, etc.) are discussed in chapter 5.

4.3.3.1 Low-level topology: connectivity, redundancy and hubs

First, network nodes can be associated with connectivity measures. The connectivity (or degree) of a node represents the number of links it has with other nodes of the network. This measure is the basis for two other connectivity measures: the 'degree distribution' and the 'clustering coefficient' (Watts and Strogatz, 1998b). The degree distribution is obtained by dividing the node connectivity by the total number of nodes and helps to differentiate between the different global network topologies (see 4.3.3.2). The clustering coefficient is a measure that characterizes the tendency of nodes to forms clusters.

The second type of local descriptors concerns the different 'paths' that exist in the network. The 'shortest path' (or characteristic path length) is defined as the minimum number of edges between two nodes and represents overall network navigability (Barabasi and Oltvai, 2004). In contrast, the 'network diameter' is the greatest distance between two nodes of a network (the longer 'shortest path'). Finally, the 'betweenness' measure represents the centrality of a node in a network and nodes with a high betweenness connect a large number of shortest paths in the network.

4.3.3.2 High-level topology: hubs, random, small-world and free-scale networks

Biological networks have been characterized by several high-level characteristics. One of the first observations was the presence of 'hubs', i.e. highly connected nodes in biological networks. Their role was determined in several model organisms. Hub proteins were found to produce larger phenotypic outcomes than less-connected proteins when deleted (Yu et al., 2008) and generally correspond to essential and more abundant genes (Ivanic et al., 2009; Jeong et al., 2001). These characteristics reinforce the role of hub proteins as potential candidates when searching for disease related genes (this aspect is discussed in paragraph 4.4.1).

At a higher level, the profile of connections characterizing a network can be associated with different network models. First, random networks contain N nodes and connect each pair of nodes with a probability p. The node degrees follow a Poisson distribution indicating that most nodes have approximately the same number of links. Another topology corresponds to small-world networks, characterized by two properties: (i) individual nodes have few neighbours; (ii) most nodes can be reached from one another through few steps. Their characteristic path length is similar to random networks in that they have a higher clustering coefficient. The third topology observed in biological networks corresponds to scale-free networks. These are characterized by a power-law degree distribution (Barabasi and Albert, 1999). Scale-free networks have a high degree of robustness. The properties of a scale-free network are often determined by a relatively small number of highly connected hubs. The corresponding biological networks are considered as resistant against random

node failures, but are sensitive to the failure of hubs. The probability that a node is highly connected is statistically more significant than in a random graph.

4.4 Practical exploitation of biological networks in systems biology

Although biological networks are an active field of research, much work still needs to be done to reach a comprehensive picture of all the interactions and regulations that occur in biological systems. In particular, the topological analysis of biological networks is still ongoing. New interactions are regularly identified, allowing new insights on biological system structures and behaviours. However, biological network studies are not only restricted to network construction and modelling. Despite their incomplete nature, the current networks are used in many biological fields. Many newly generated omics data are now cross-linked with network knowledge to provide a system-level picture of biological phenomena. Referencing all the applications would be a colossal work. In the following paragraphs, some examples of the most recent applications of biological networks in biomedical research and bioengineering are discussed.

4.4.1 The emerging concept of 'network medicine'

A disease is rarely linked to the abnormality of a single gene. In many cases, no clear link can be directly drawn between genotype and phenotype. The decreasing cost of genome-wide association studies and full genome sequencing is now motivating the consideration of biological networks for many biomedical applications. In this context, systems biology is now expanding from theoretical research to a new field known as 'personalized medicine' (Chen and Snyder, 2012; Weston and Hood, 2004). The idea that analysing biological network perturbations can help to understand disease is now clearly established (Barabasi et al., 2011). Specific sub-structures of a network are more prone to favour the appearance of disease if they undergo perturbations (figure 4-9). For example, several studies found that genes linked to diseases with similar phenotypes have a significantly increased tendency to interact directly with each other, forming 'disease network modules' (Gandhi et al., 2006; Goh et al., 2007; Xu and Li, 2006).

Several studies have focused on the effect that the perturbation of these disease modules could induce. For example, Engin et al. simulated the impact of drugs at the systems level by "attacking" network nodes or edges of a PPI network with a drug (Engin et al., 2012). Similarly, Wang et al. integrated 3D protein structure information with high-quality large-scale PPI data to examine the relationships between human diseases and synonymous/non-synonymous single nucleotide polymorphisms (SNPs) (Wang et al., 2012). Their analysis produced a prediction model where the mutation in protein interfaces can be related to a specific group of diseases. Network perturbations can also be induced by external pathogens, particularly viruses. Virus-host interactions evolved to rewire host cellular pathways to the advantage of the viruses (Tafforeau et al., 2012). These rewirings are mainly operated in PPI networks, sometimes called 'virhostomes': for example, virhostomes have been constructed for the influenza virus (Shapira et al., 2009) and HIV (Jager et al., 2011). These

networks showed that viral proteins preferentially target hubs in host interactome networks (Shapira et al., 2009). Interestingly, this systemic view of virus interactomes initiated the idea of a new virus classification mixing phylogeny and viral-host PPI profiling (Xu et al., 2011).
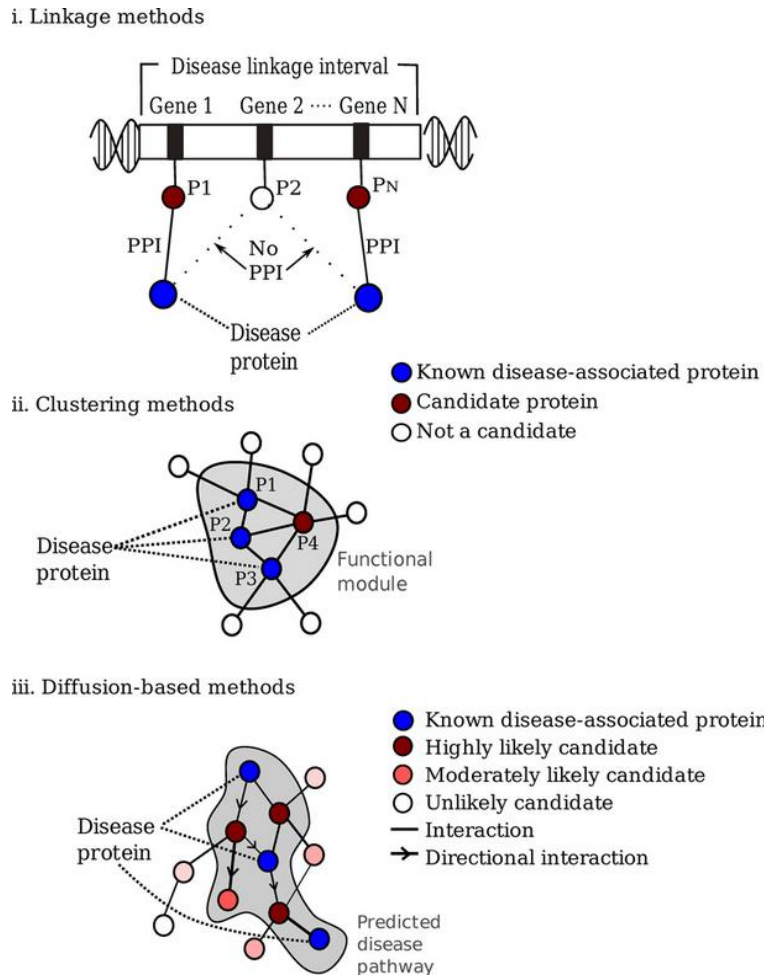
Figure 4-9. Example prediction of disease-associated proteins based on a PPI network. *(i) Proteins linked to genes known as disease-related are flagged as candidate genes. (ii) A graph clustering technique defines the disease modules of the interactomes, adding new potential candidates to disease-related proteins. (iii) Diffusion-based methods use a random walker that visits each node in the interactome with a certain probability. The outcome of the algorithm is a score assigned to each protein representing the likelihood that a particular protein is associated with the disease. Adapted from Barabasi et al. 2011.*

But one of the major current technological advances is the feasibility of time scale analysis of the regulatory networks that are dysfunctional in a given disease state (de la Fuente, 2010). One of the first experimental proofs of concept of personalized medicine was published recently by Chen and co-workers (Chen et al., 2012). They performed an analysis that combines genomic, transcriptomic, proteomic, metabolomic, and autoantibody profiles from a single individual over a 14 month period. Their analysis demonstrated the biological network changes that were the consequence of two consecutive viral infections. Moreover, this analysis showed how these perturbations favoured

unstable glucose blood levels and a type 2 diabetes predisposition. This striking result shows promise for a future personalized omics monitoring and a personalized medicine based on systems biology knowledge.

4.4.2 Pathway engineering, a step towards synthetic biology

The study of the structures and dynamics of hundreds of biological networks provides the basis for artificially modifying network topology or influencing network dynamics for a particular purpose. The large field of bioengineering is focused on this purpose and is strongly motivated by a technology transfer to biotechnology industries. Bioengineering requires a comprehensive knowledge of the biological systems involved. At the molecular level, protein expression can be modified by promoter enhancement or by RNA synthetic devices, and protein 2D or 3D structure can be modified to enhance its activity (Krivoruchko et al., 2011). At the network level, genetic modifications can be operated to introduce new protein activators or repressors to modify the regulatory network, or whole synthetic metabolic pathways can be engineered to enhance metabolic flux or produce new biochemical or chemical products (Felnagle et al., 2012; Krivoruchko et al., 2011). *Saccharomyces cerevisiae* and *Escherichia coli* are the most common organisms in bioengineering because their biological systems have currently the most complete description. However, recent developments show that algae and plants are good candidates for specific compound production (Larkum et al., 2012; Shin et al., 2012). Bioengineering has been successful in several domains. Industrial cell reactors can now be used to produce pharmaceutical compounds (Shin et al., 2012), while many synthetic isoprenoid pathways are used in perfumes and food aromas (Misawa, 2011). Other developments are related to important societal problems such heavy metal bioremediation (Soares and Soares, 2012) or plastic polymer production (Penloglou et al., 2012).

A decade of bioengineering development has created a specialized field that describes biological systems from an engineering perspective. Engineering-driven approaches of modularization, rationalization and modelling, have been slowly transferred to biological networks, creating the field of synthetic biology. This field uses frameworks established in electrical engineering and biological networks are "wired" to manifest logical forms of cellular control (Khalil and Collins, 2010). In the near future, synthetic biology aims to complement traditional genetic engineering (mutations which eliminate network nodes or chemical inhibition) with a set of research tools for a fully controlled investigation of biological networks (Weber and Fussenegger, 2012). Figure 4-10 illustrates such a device: a tunable controller (dial) can be turned on and off (switch) and is activated by a protein input (light sensor) (Bashor et al., 2010). An application of a similar circuit was designed by Levskaya and coworkers: the output of the module is connected to an actin polymerization function, allowing a light control of cellular cytoskeleton (Levskaya et al., 2009).
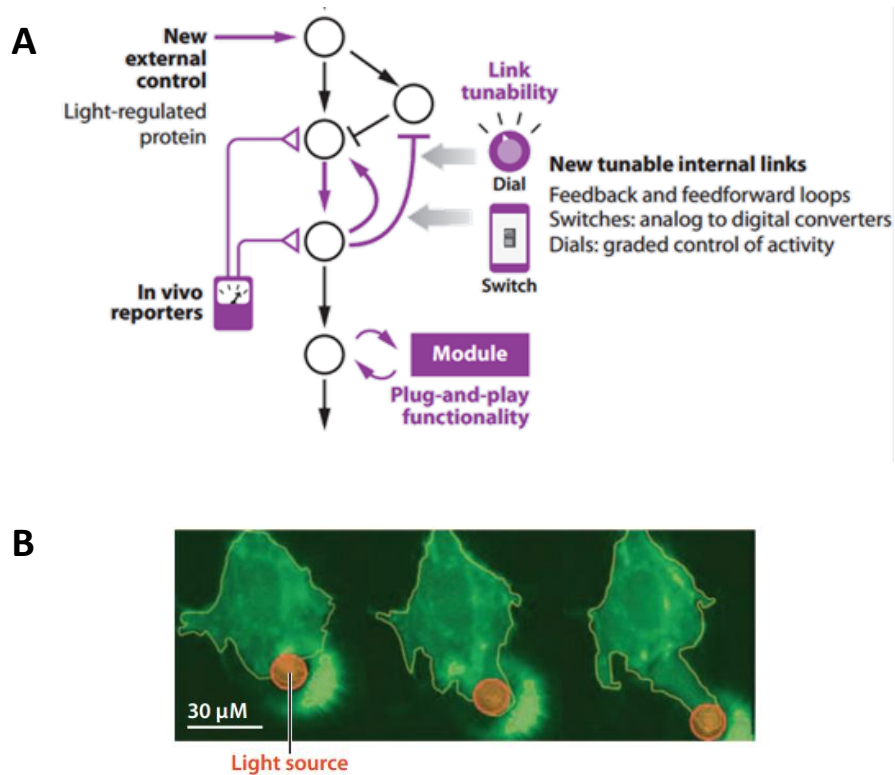
Figure 4-10. A synthetic biology pathway module. *(A) The module contains switch and dial functionality. The circuit is activated by a light sensitive protein and can be connected to a specific biological function. (B) The circuit was adapted in a cellular system for the activation of actin polymerisation, allowing a light controlled polymerisation of the cytoskeleton. Adapted from Bash et al., 2010 and Levskaya et al., 2009.*

4.5 Bioinformatics resources for biological pathways

4.5.1 An overview of pathway databases

Pathway databases are heterogeneous in terms of organism content, pathway type content, data format and even in the conceptualization of pathways. One of the first pathway databases was the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa et al., 2012). KEGG is a fully manually constructed database compiling chemical and genomic data from many databases. It contains many types of pathways (metabolic, signaling, cell cycle, cell communication, etc.) and provides eukaryotic, archaeal and bacterial pathways through an orthology-based pipeline. Another example database is Reactome, a knowledgebase that describes human biological processes and specifically developed tools for computational analyses (Croft et al., 2011). Similarly to KEGG, Reactome contains manually constructed pathways, available for different species through the orthology predictions of Ensembl Compara. However, these maps can be easily extended by the biological community with user-friendly tools. Moreover, the different pathways are organized in a hierarchical fashion, a large pathway being considered as a patchwork of other pathway modules. The Reactome authors paid attention to the interfacing of the database with other computational

resources: the data structure is based on OWL standards (http://www.w3.org/standards/techs/owl) and the biomart database engine (www.biomart.org). A third general pathway resource is wikipathways (Kelder et al., 2012). This website provides a wiki-like interface to allow the community to create and edit biological pathways. Content is similar to KEGG and Reactome, but the open-access and user-friendly interface currently make it the most complete pathway database with more than 1700 pathways, 21 species and more than 300 hundred active editors. Its main drawback is that the pathway querying tools are more limited than KEGG and Reactome. Finally, we can cite the BioCarta database developed by BioCarta LLC (http://www.biocarta.com). The content of BioCarta is similar to KEGG and Reactome, pathways can be edited or submitted after registration and validated by the BioCarta team. However, there is no way to directly access BioCarta data without a clear request to the company.

KEGG, Reactome, Wikipathway and BioCarta compile pathways representing different types of biological processes. They are complemented by several pathway databases specialized in metabolic and signaling pathways. The NCI-Nature pathway database groups human regulatory and signaling networks available in different flat file formats and users can submit new pathway schemas (Schaefer et al., 2009). Similarly, the INOH database provides files describing metabolic and signal transduction pathways (Schaefer et al., 2009). INOH data can be visualized with in-house software. Some metabolic databases are not only based on a manual construction but use computational pipelines to build pathway models. For example, the Homo sapiens Recon 1 database constructs a genome-scale model connecting all human metabolic pathways by mixing manually curated pathways and in silico predictions based on literature mining and provides tools to model human metabolic phenotypes (Duarte et al., 2007). Similarly, the Edinburgh Human Metabolic Network (EHMN) mixes KEGG and Uniprot genome annotations with literature data to construct a high-quality genome-scale metabolic network in human (Ma et al., 2007). The most complete collection of metabolic networks is provided by the collection of Pathway/Genome Databases (PGDBs) (Caspi et al., 2012). This collection, known as Biocyc, groups several databases dedicated to human and animal models, plants and MetaCyc, a database compiling metabolism pathways for more than 2200 eukaryotes and bacteria. Biocyc databases contain metabolic pathways that are based on scientific literature. But in MetaCyc these high-quality pathways are automatically transferred by an automated electronic referencing of newly sequenced organisms. This pipeline uses several biological criteria such as phylogenetic profiling and the consideration of metabolic operons. All pathways can be manually edited by the scientific community.

4.5.2 Computational representation of pathways

The abundance of various pathway databases can be a technical problem for users because they use different pathway definitions, heterogeneous data structures and exchange formats. Two ontologies have been developed to describe biological pathway related content. The INOH ontology was developed synchronously with the INOH metabolic pathway database and is designed to annotate molecules in the scientific literature on signal transduction pathways (Yamamoto et al., 2004). The BioPAX - Biological Pathway Exchange – ontology is more general and offers a well-defined semantics for pathway representation (Demir et al., 2010). This standard is now integrated in most pathway

databases and the latest version officially supports metabolic pathways, signaling pathways, gene regulatory networks, molecular interactions and genetic interactions. The SMBL (System Biology Markup Language) is also used in metabolic pathway databases but is more adapted to systems modeling (Gauges et al., 2006).

4.5.3 Integrating multiple pathway databases

With the abundance of pathway databases, several authors have developed data warehouse systems to integrate pathway data from different sources. HiPathDB (Yu et al., 2012), HPD (Chowbina et al., 2009) and integromeDB (Baitaluk et al., 2012) were developed for this purpose. IntegromeDB groups all pathway data from Reactome, KEGG, BioCarta, NCI-Nature pathways, WikiPathways and HumanCyc with gene-based searches and provides a dedicated visualization tool. HiPathDB and HPD provide more complex tools that decipher the overlap existing between the different database pathways. Moreover, HiPathDB infers a new network representation corresponding to a synthesis of all network databases.

4.5.4 Consistency of pathway databases

An interesting question in biological pathway analysis is which database provides the most exhaustive description of a particular process. Despite the availability of many reviews referencing the different databases, only one group has performed a critical assessment of pathway database content. Stobbe et al. compared the human metabolic pathway of EHMN, HumanCyc, KEGG, and Reactome (Stobbe et al., 2011). They compared several criteria such as genes, metabolites, EC numbers and reactions. Reactions were considered to be the same if all substrates and products matched. The global overlap of the databases is resumed in the following table:

Criteria	Gene	EC number	Metabolite	Reaction	Reaction without considering e^-, H^+, H_2O
# of common entities	3858	164	4679	7758	6968
Similar description coverage	13%	51%	9%	1%	3%

Table 4-1. Coverage of four metabolic databases for 4 different entities.

These results illustrate the very low coverage that exists between pathway databases for metabolic pathways for all components of the pathway. A second publication of the same group focused on an in-depth analysis of the well-known TCA cycle to understand why database coverage is so limited (Stobbe et al., 2012). The analysis was extended to 10 metabolic databases. Surprisingly, only 3 of the

9 main reactions of the cycle were concordant over the 3 databases (figure 4-11). The succinate to fumarate reaction was the same in only half of the databases and secondary reactions catalysed by other enzymes were associated only in 1 to 4 databases, depending on the reaction. The TCA cycle has been well established for decades and a vast literature is available to support its reaction steps. Stobbe et al. highlighted several reasons explaining this result. First, several genes, found in one or more pathway databases, were suggested to be involved in the TCA cycle, but with no evidence in the literature. Second, several reactions of the TCA cycle are still actively studied and some parts of the cycle are still debated (unidirectional or bi-directional reactions, cofactors, etc.). These results demonstrate that (i) we still do not have a complete picture of the dynamics of biological networks, even for a metabolic network considered as a standard and (ii) the development of community standards for the description of biological pathways is now urgent.
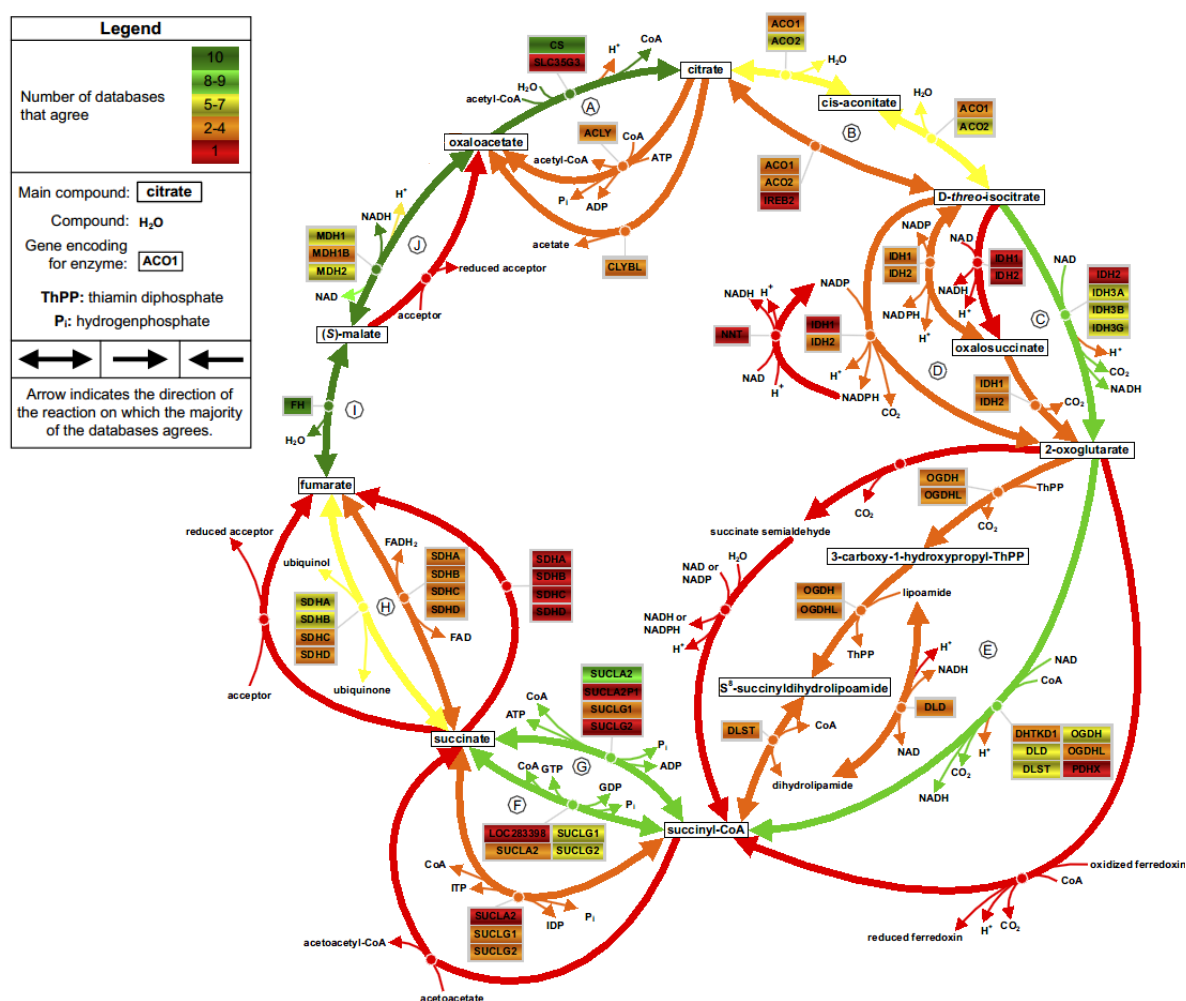
Figure 4-11. Pathway database consistency for the TCA cycle. *The inventory of all genes, metabolites and reactions was done in 10 different metabolic pathway databases. Agreement between databases is represented by a colour gradient from red (no agreement between database) to green (full agreement in all databases). Adapted from Stobbe et al., 2012.*

5 EVOLUTION AND SYSTEMS BIOLOGY: BIDIRECTIONAL BENEFITS

The systems biology era is an incredible opportunity for evolutionary studies: contributing multi-level approaches to complement the analysis of genomic and phenotypic variation. System-level variations can involve both the structure of a network and its dynamics, opening to way to finally linking genotypic and phenotypic evolution. This idea is not new. For example, the intra-species variation of endocrinal levels has already been observed in primates (Coe et al., 1992) and the evolution of corresponding steroid receptors was related to the origin of vertebrates (Baker, 1997). However, this example describes evolutionary innovations of an organism-level network of interactions. It is only recently, thanks to the technological omics developments, that biological network variations can be observed at the molecular level and in multiple organisms. In contrast to genomics, which is now supported by a complete evolutionary theory and numerous analysis tools, characterising higher biological levels in an evolutionary framework remains a challenge. In this chapter, I will briefly describe how the technological breakthroughs in genome sequencing have significantly modified evolutionary theories. Then, I will present the first ideas that have been proposed to exploit the currently available systems level data in an evolutionary framework. Such integrated analyses will not only extend our knowledge of evolutionary mechanisms, but also provide new evolutionary tools that will be useful in many systems-related fields.

5.1 Recent evolutionary discoveries

The huge volume of genomic data resulting from recent high throughput sequencing is a good example of how technological breakthroughs can revolutionise our view of Evolution. For example, during the last decade, many independent studies have been performed on hundreds of protist genomes, revealing a complex evolutionary history of unicellular eukaryotes and highlighting the fact that embryophytes, fungi and metazoan phyla do not represent the full eukaryotic diversity (figure 5-1) (Brinkmann and Philippe, 2007).

Another striking result discovered recently is the importance of horizontal gene transfers (HGT). HGT are preponderant in most bacterial phyla (Gupta and Griffiths, 2002) and also play a major role in eukaryotes (Bock, 2010; Keeling, 2009). For example, three independent HGT participated in the emergence of plant parasitism in the nematode lineage (Haegeman et al., 2011). An HGT between an algae and a sea slug even allowed the latter to perform photosynthesis (Rumpho et al., 2008). Nevertheless, eukaryotic HGT remains relatively rare. In contrast, HGT seems to be a major driver of evolution in bacteria and some authors have proposed that HGT have an impact similar to gene duplications in bacterial genome evolution (Boto, 2010). The consideration of these mechanisms has initiated new philosophical debates since classical phylogenies cannot accurately represent this 'bush' description of bacterial evolutionary history. The classical tree of inheritance described by Darwin is not representative of their evolution and as a consequence, some authors are now developing concepts of 3D phylogenomic networks (figure 5-2) (Dagan, 2011). Interestingly, these observations are based on the currently available bacterial genomes, of which 60% are human

pathogens (www.genomesonline.org, menu 'statistics'). Thus, massive genome sequencing has shed light on new evolutionary processes, and in turn, the elucidation of these processes contributes to our understanding of systems-level functions, such as the implication of HGT in bacterial pathogenicity and drug resistance (Dzidic and Bedekovic, 2003).
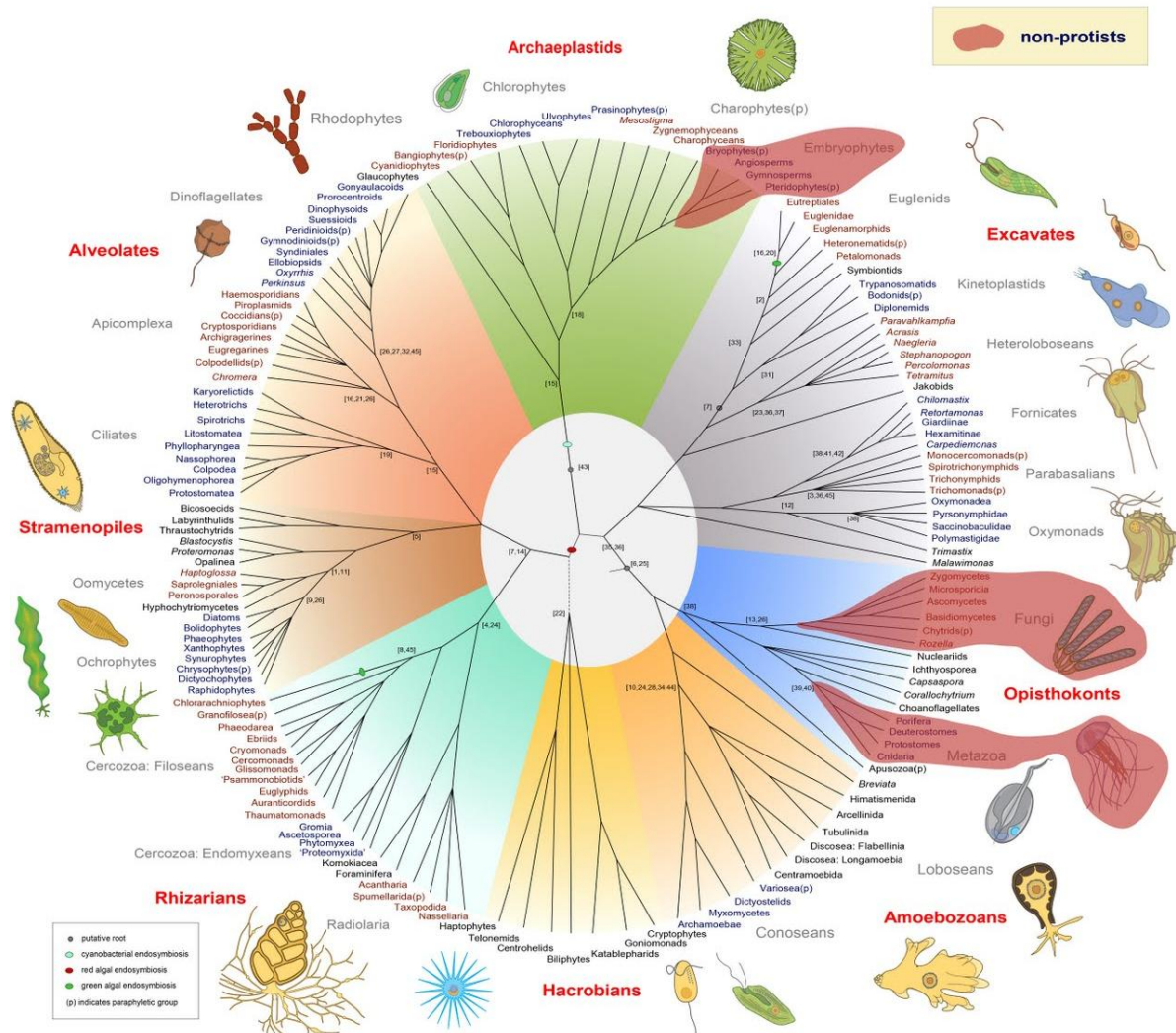
Figure 5-1. A recent tree of eukaryotes updated with protists genomes. *Adapted from kepticwonder.fieldofscience.com.*

These examples illustrate the bidirectional contributions of developments in sequencing technologies and new evolutionary theories. However, genomic data has been available for 50 years and many bioinformatics approaches have been developed during this time to extract the knowledge hidden in the sequences. In contrast, systems biology is based on recent methodologies that are only beginning to achieve a certain amount of maturity and system-level evolutionary analyses are still in their infancy. Such an analysis framework is sometimes referred to as 'evolutionary systems biology', an emerging field studying evolution and innovation at all levels of biological organisation, from genes and genomes, to biological networks and whole organisms as well as their communities.
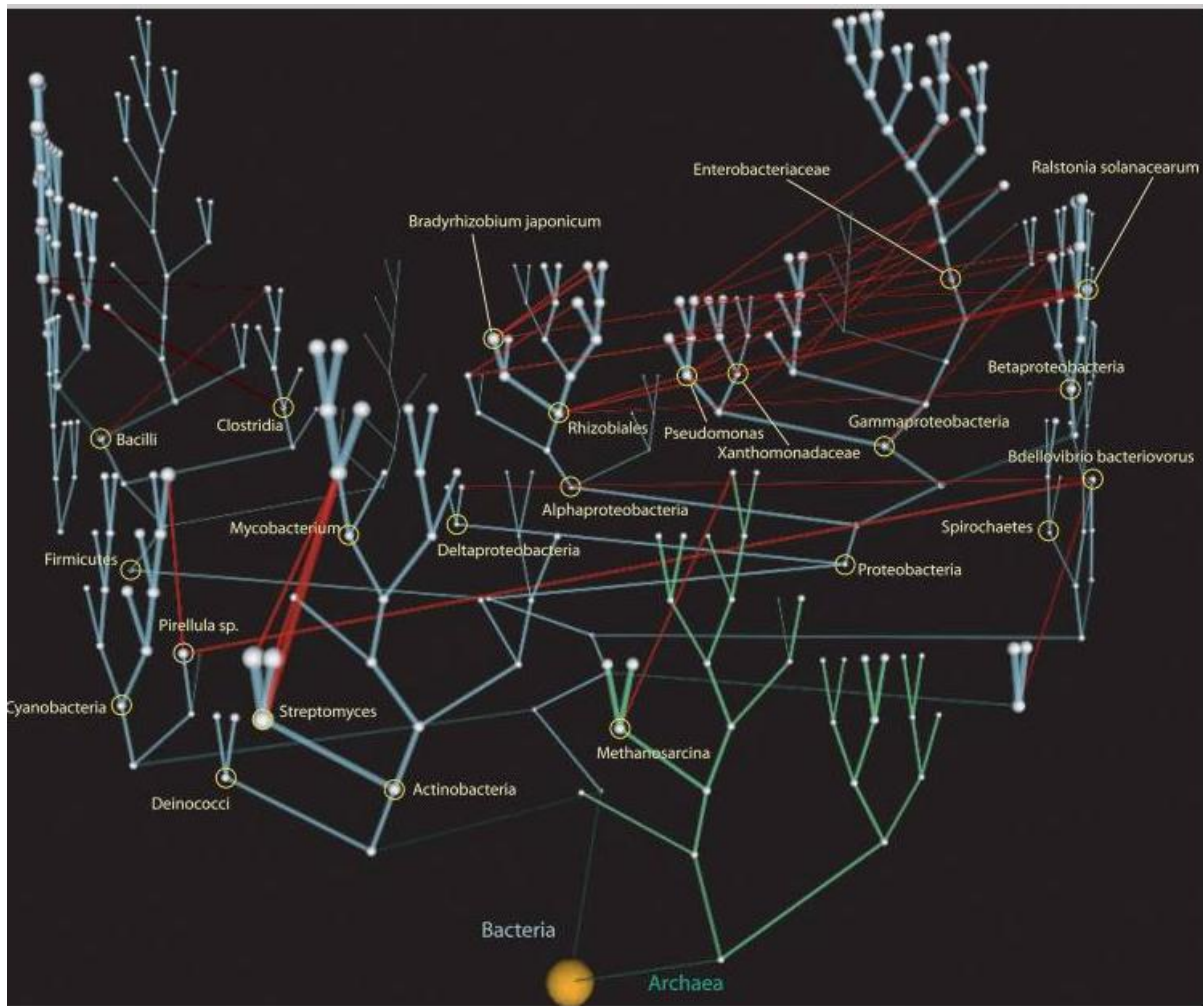
Figure 5-2. A 3D phylogenomic network. *This network combines classical phylogenetic tree representations (blue and clear green edges) with HGTs (red and dark green edges) for different bacterial and archaeal phyla. HGT are preponderant in bacteria, resulting in global pictures of gene evolutionary histories that are more complex than the classical gene-inheritance based phylogenetic representation. Adapted from Dagan et al, 2011.*

5.2   Discovering evolutionary knowledge at multiple biological levels

The modern evolutionary synthesis reconciled genetics with the Darwinian principles of Evolution, but so far few studies have characterized the evolution of higher biological levels. Today, advances in systems biology are opening the way to an assessment of the variation inherent to biological systems at multiple levels among different organisms. Here, we describe some recent results related to this multi-level variation and the corresponding debates that are on-going. The particular case of biological network and pathway evolution is more extensively described in paragraph 5.3.

5.2.1 Evolutionary role of non-coding RNAs

Non-coding RNAs (ncRNAs) can interact with DNA, RNA and protein molecules and are involved in diverse structural, functional and regulatory activities. They play roles in nuclear organization and transcriptional, post-transcriptional and epigenetic processes (Morris, 2012). Compared with protein-coding sequence, microRNA sequence tends to be weakly constrained and understanding their role in biological systems evolution is challenging. Nevertheless, new technologies are now available to study ncRNA in multiple species, providing new opportunities to elucidate their evolutionary roles. Current transcriptomics studies are producing a vast amount of data concerning ncRNAs, although it is difficult to assess whether they are functional or whether they represent noisy transcription. Initially, ncRNA evolutionary studies based on genome comparisons showed a poor conservation of the corresponding intergenic sequences and were not able to find evidence of their function (Wang et al., 2004). More recently, some conserved ncRNA subfamilies, such as long intervening noncoding RNAs (lincRNAs) have been described in mammals (Guttman et al., 2010) or fishes (Ponting et al., 2009) with the help of RNA-seq technologies. Despite a rapid sequence evolution, a conservation of the functional role of lincRNAs was demonstrated, in particular for embryonic development (Ulitsky et al., 2011). A significant fraction of another ncRNA class, the miRNAs, is known to be conserved over many species (Hertel et al., 2006; Hoeppner et al., 2009). Interestingly, microRNAs have been used several times as phylogenetic markers to study the emergence of vertebrates (Heimberg et al., 2010).

5.2.2 Evolution of gene expression

Phenotypic diversity can be partially understood by genome and protein-coding gene analysis. However, regulatory mutations affecting gene expression have been considered essential for understanding phenotypes for a long time (King and Wilson, 1975). Several studies have attempted to identify the differences in expression levels between closely related species, mainly human and apes (Khaitovich et al., 2006) and some authors have successfully extracted general evolutionary trends, such as the strong inverse correlation between a gene's sequence evolutionary rate and expression level (Koonin and Wolf, 2006). Until recently, transcriptomes were mainly produced with microarray techniques, a technique requiring tissue specific probes, which made it difficult to apply to multiple organisms. The development of RNA sequencing (RNA-seq) now facilitates the determination of expression levels in multiple tissues. One of the first multi-species comparisons of expression levels was performed by Brawand and co-workers, who analysed the expression levels in brain, cerebellum, heart, kidney, liver and testis from nine mammalian species (human, apes, marsupials, monotremes and a bird as outgroup) (Brawand et al., 2011). Their results show many interesting evolutionary trends. For example, they observed large differences of expression patterns in primate brains, much larger than those observed in the corresponding genomes. In a separate study, they also highlighted the dosage compensation, i.e. the X-linked gene expression level, that characterizes the evolution of sex chromosomes in amniotes (Julien et al., 2012).

5.2.3   Evolutionary role of epigenetics

An epigenetic effect is defined as one or several factors altering a phenotype that is heritable but not solely due to changes in DNA (Goldberg et al., 2007). In cells, chromatin modelling, such as DNA methylation or histone modification, is the main mechanisms of epigenetics and its role was confirmed as essential in cell differentiation and tissue development (Mohn and Schubeler, 2009). However, epigenetic inheritance can also be transmitted during gamete production, initiating a trans-generational inheritance. Conserved stability of chromatin modification has been described in many organisms. Bacteria can present an epigenetic based resistance to antibiotics (Adam et al., 2008). In cultures of palm oil, somaclonal phenotypic variation was observed between clones, related to their retroelements, transposons and methylation status (Kalendar et al., 2011). In mice, environmental changes influence the locus responsible for the Agouti fur colour, a methylation state that can be inherited (figure 5-3) (Wolstenholme et al., 2011). Different versions of a heritable epigenetic modification can even be conserved in a population, a phenomenon similar to gene alleles and referred to as 'epialleles' (Maury et al., 2012). The inventory of human epialleles and the transgenerational inheritance of an aberrant epigenetic state could be a new key to understanding some diseases (Morgan and Whitelaw, 2008). Epigenetically mediated transgenerational inheritance is now considered as an important mechanism in Evolution. However, the exact evolutionary implications of epigenetic inheritance are unclear. Current debates include the role of epigenetic effects in the generation of novel phenotypes, the mediation of transgenerational adaptive plasticity or the formation of additional inheritance channels in parallel with DNA (Jablonka and Raz, 2009; Shea et al., 2011).
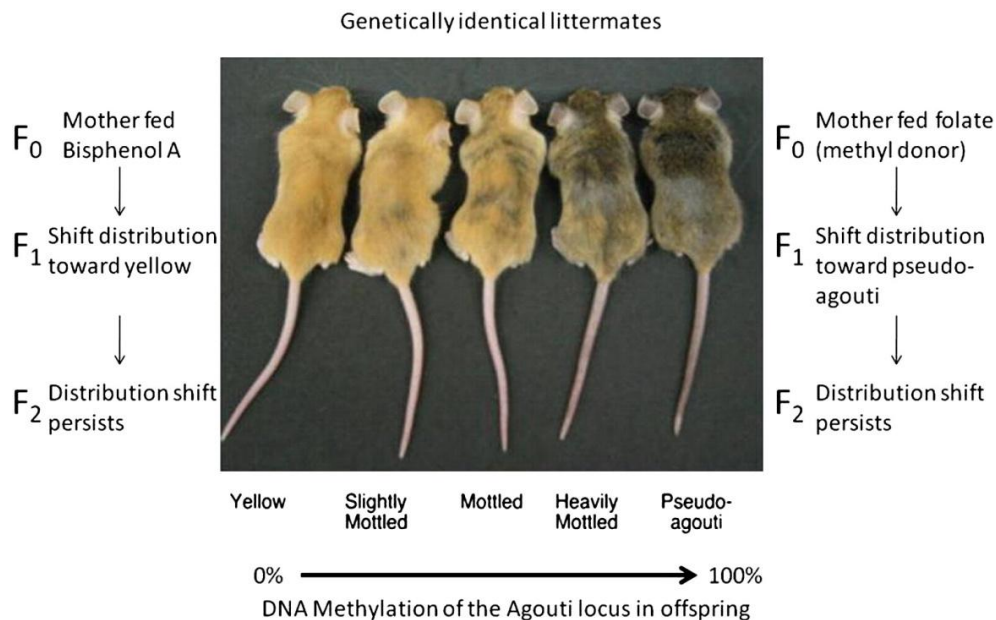
Figure 5-3. An example of epigenetic inheritance: DNA methylation changes at the Agouti locus. *Environmental changes can change the DNA methylation at the mouse Agouti locus. Mice with a low level of Agouti methylation are yellow, while mice with a high level of DNA methylation are agouti. Treatment with bisphenol A or folic acid can shift coat colour distribution towards yellow or agouti, respectively. This coat color shift persists to the next generation indicating that the epigenetic change that occurs at the agouti locus is heritable. Adapted from Wolstenholm et al., 2011.*

5.2.4 Towards an extended evolutionary synthesis

The growth of systems biology and the multi-level view in evolutionary studies suggest that a new consensus is now necessary in order to construct an updated evolutionary synthesis. Conrad Waddington suggested the consideration of the epigenetic landscape in evolutionary studies over 5 years ago (Goldberg et al., 2007). His concept of epigenetics was however only a metaphor for how gene regulation modulates development. Today, biology systems are beginning to be characterised at many levels and modern biology describes the coordinated activity of proteins within networks, rather than individual or even a few interacting gene products. The gap between the widely accepted first evolutionary synthesis, a synthesis mainly integrating genetics with evolutionary theories, and the current multi-level description of biological systems is growing. Consequently, more and more authors are highlighting the need for a modernized evolutionary synthesis integrating the increasing knowledge generated by systems biology (Bard, 2010; Brower, 2010; Chen and Wu, 2007). The integration of multiple genome-scale datasets has begun to uncover general evolutionary trends at the system level, studies that have been coined by several authors as 'evolutionary systems biology' (Gu, 2011; Koonin and Wolf, 2006; Loewe, 2009). Interestingly, similarly to the term 'system biology' 10 years ago, many groups and institutes are now qualifying their teams as specialized in 'evolutionary systems biology', these teams being originally specialized in comparative genomics, evolutionary bioinformatics and molecular biology.

This modernization is still actively discussed in the evolutionary field, but some guiding threads are emerging. A first agreement is that the role of epigenetic inheritance should be integrated in classical evolutionary theories (Dickins and Rahman, 2012; Pigliucci, 2007). Second, the field of evo-devo (see chapter 1) and the particular evolutionary constraints that it describes in metazoa should also be considered (Pennisi, 2008). Another problem is that some recent discoveries contest some well-established principles of the current evolutionary theory. For example, there is evidence that the LUCA (Last Universal Common Ancestor) might have been dramatically different from modern cells: possibly, a loose collection of virus-like genetic elements that could be denoted as LUCAS, a Last Universal Common Ancestral State (Koonin, 2009b). The 'bush' description of bacterial evolution also contests the linear properties of classical phylogenetic trees, one of the pillars of Darwinian theories. Indeed, in bacteria, gene inheritance is not the only basis for molecular evolution and HGT plays a major role. These new hypotheses are initiating an interesting debate about Life and Evolution: are they really 'law-like' and can we hope to describe a single unified model of Evolution? (Weiss and Buchanan, 2011).

5.3 Evolution of biological networks

One of the greatest successes of systems biology has been the development of a large knowledge base describing biological networks and their dynamics (see chapter 4). Omics approaches are now facilitating the production of network data and several genome-scale networks have been constructed in model organisms (figure 5-4). These networks provide an opportunity to decipher how network architecture evolves over time.
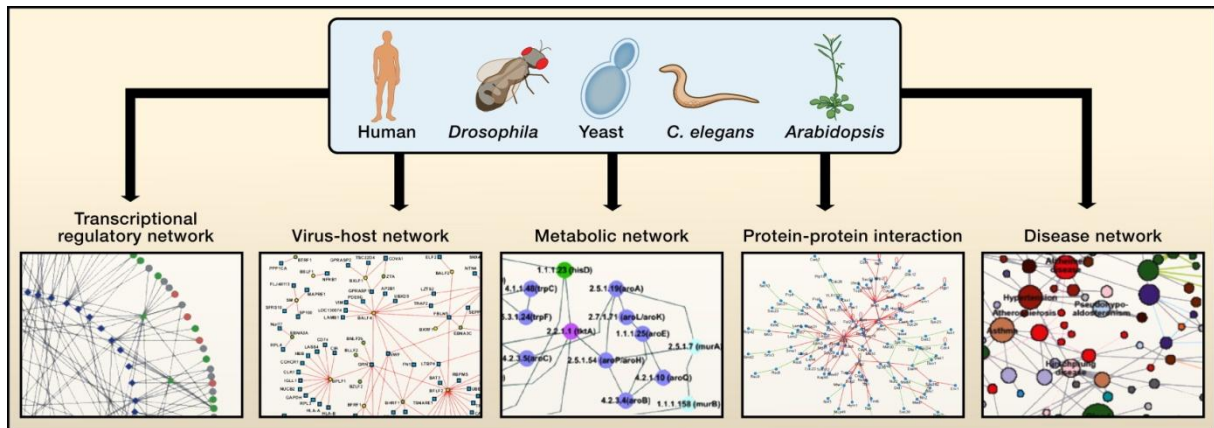
Figure 5-4. The available interactome networks in model organisms. Adapted from Pennisi, 2008.

The existing studies of biological network evolution can be divided into two main approaches. First, many authors are interested in a mathematical description of networks and how their structure changes and grows over time (Gibson and Goldberg, 2011). Such work is generally based on simulated networks that are submitted to external changes. For example, mathematical models were developed to explain the modular aspects of networks (Pfeiffer et al., 2005). The concept of network evolvability has also been proposed (Crombach and Hogeweg, 2008), stating that networks are evolvable while their robustness to mutations is maintained and delimitating a landscape of network evolution. A second type of study concerns the inference of evolutionary knowledge from real biological network datasets. In th rest of the chapter, I will focus on these approaches. These include bottom-up approaches, to investigate how the nodes and edges of a specific network are modified through time, and top-down approaches to study the global network properties that characterize network evolution.

5.3.1   Mechanisms of network evolution

Although the mechanisms of biological network evolution are clearly complex, network modifications can be resumed in three basic operations: node gain/loss, link gain/loss. These events are often coupled to genetic events such as deleterious mutations, domain recombinations, 3D structure modifications, gene duplications/deletions or HGTs. But they can also be linked to modifications of regulatory elements or epigenetic modifications (Koonin and Wolf, 2010). We can illustrate all these events by looking at the example of the Notch signalling pathway, for which a complete evolutionary analysis has been performed (Aravind et al., 2009) (figure 5-5).
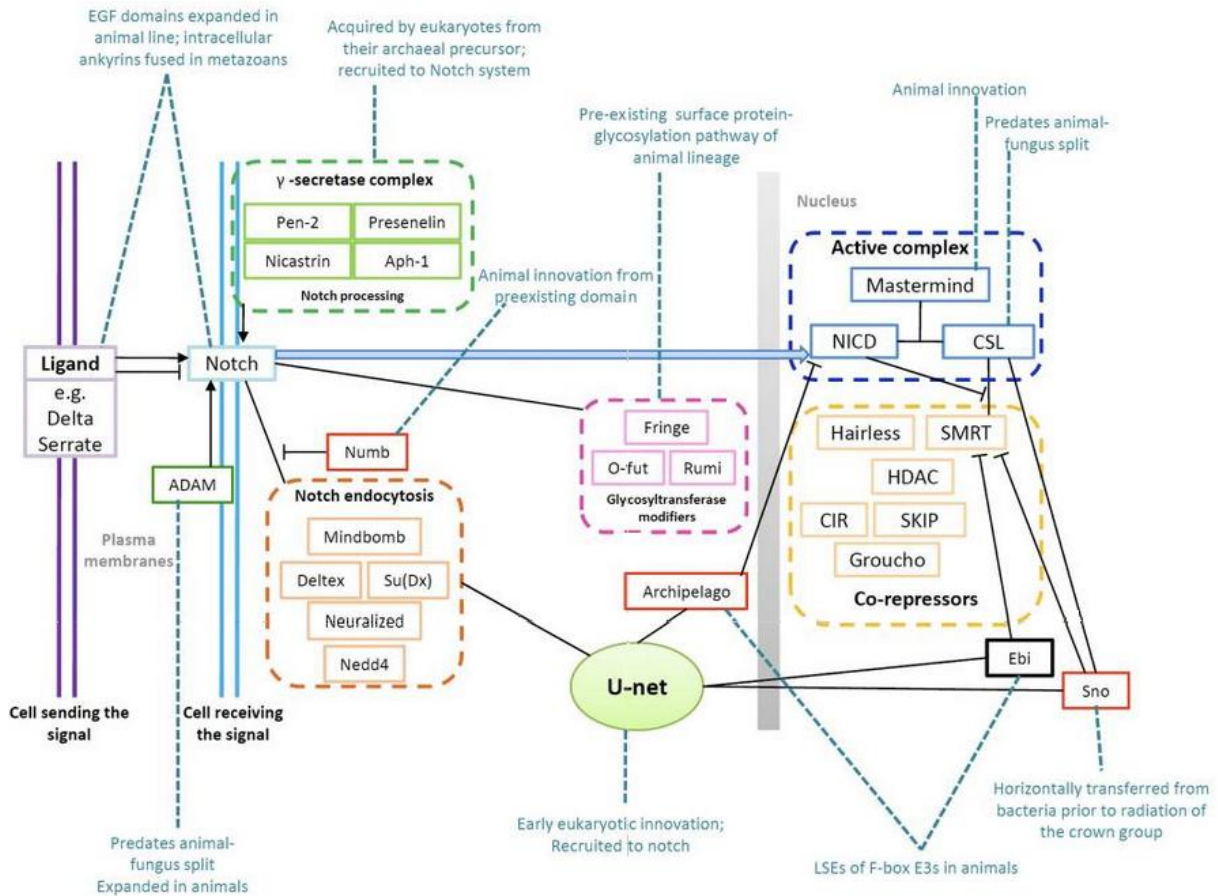
Figure 5-5. A representation of the Notch signalling network and the evolutionary events that shaped it. Boxes indicate gene co-functional linkages. The network is widely annotated with labels indicating the evolutionary history of different components. Adapted from Aravind et al., 2009.

The network includes a gene horizontally transferred from bacteria (Sno), the recruitment of a complete network module (the secretase complex), some innovations specific to animals (Numb) and gene family expansions. The functional repurposing of cellular machines seems quite important for network plasticity and the use of conserved machines in different pathways with different input/output has already been observed *in vivo* in fungi and mammals (Frost et al., 2012).The complexity of these network evolutionary scenarios is complemented by the complexity of the gene expression patterns. For example, a study of transcript and isoform diversity showed that nearly 20% of the human genes possess alternative interaction potential, suggesting that transcriptional variation can significantly rewire human interactomes (Davis et al., 2012).

5.3.2   Comparing biological networks from multiple species

The comparison (or alignment) of biological networks from multiple species can be performed to analyze network evolution, using a philosophy similar to protein or genome alignments. Network alignment is however a more complicated task. Several methods have been developed that focus on

finding similarities between the structure or topology of two or more networks, including both local and global network alignments. Interestingly, most of these methods are based on sequence homology (mainly orthology) to make a correspondence between nodes in networks from different species. Local methods are generally based on similar path scores, conserved protein clusters or conserved subnetworks defined by arbitrary structures (Flannick et al., 2006; Kalaev et al., 2008; Kelley et al., 2004). Global network alignments provide a unique correspondence between every node in the smaller network to exactly one node in the larger network. They are based on heuristics aimed at maximizing the overall match between the two networks or on learning algorithms that use a training set of known network alignments coming from closely-related species (Flannick et al., 2009; Shih and Parthasarathy, 2012; Singh et al., 2007).
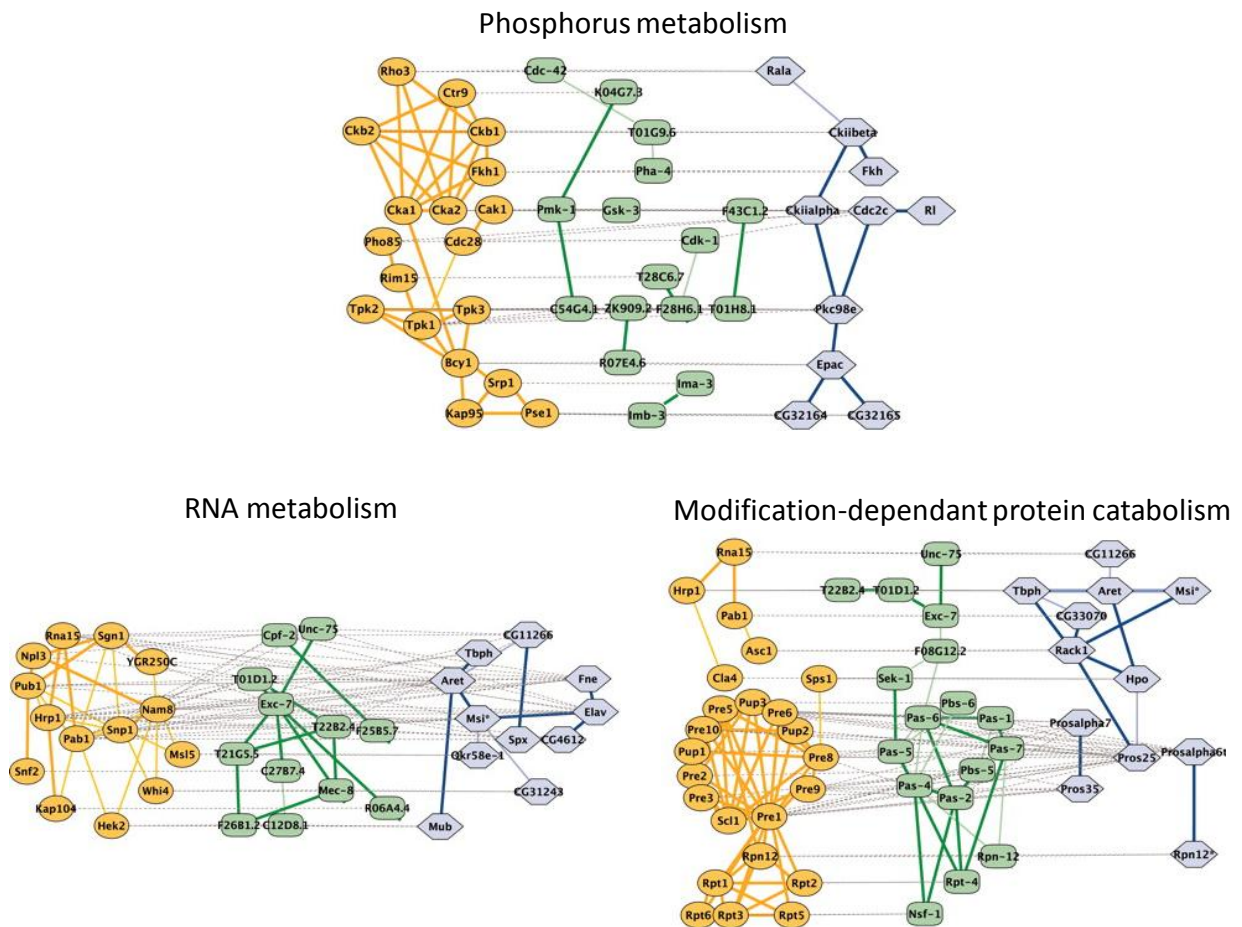
Figure 5-6. Some examples of conserved network modules in yeast, worm, and fly. *The PATHBLAST tool was used to discover conserved topologies over the three species PPI networks. Yeast proteins are represented by orange ovals, worm proteins by green rectangles and fly proteins by blue hexagons. They are connected by direct (thick line) or indirect (connection via a common network neighbour; thin line) protein interactions. Dotted lines link sequence homologs which are horizontally aligned. Adapted from Sharon et al., 2005.*

Network alignments have produced many interesting results. For example, it was observed that some large modules of interactions are conserved over multiple species (figure 5-6) (Sharan et al., 2005). Network alignments have also been used to predict functional orthologs (Bandyopadhyay et al., 2006). Regulatory networks have been observed to be more plastic and evolve more rapidly than PPI networks, probably linked to a faster evolution of transcription factors compared to their target genes (Babu, 2010a). This picture seems to be reinforced in bacteria, which present an extreme plasticity in their transcriptional regulatory networks (Lozada-Chavez et al., 2006). Finally, global alignments have been used to produce distance matrices representing the differences between species networks for the reconstruction of a network-based phylogenetic tree (Kuchaiev et al., 2010).

5.3.3   Discovering global network properties

The multi-species comparison of biological networks is not only used to highlight which components of the system are conserved or to construct system-based trees. The availability of genome-scale networks in multiple species has led some authors to search for general properties governing biological networks and several models have been proposed. For example, the networks are qualified as 'scale-free' and can be characterized by a 'small world' structure (Watts and Strogatz, 1998)(Watts and Strogatz, 1998a), meaning that they can be considered as a collection of interconnected hubs and each node can be reached from any other by a short path. Furthermore, the number of links to each node in biological networks has been associated with a power law distribution (Barabasi and Albert, 1999). Such studies have motivated some authors to propose that 'cellular networks are governed by universal laws' (Barabasi and Oltvai, 2004). Two other parameters have been used to characterize network evolutionary changes: their robustness (Zhang and Zhang, 2009) and evolvability (Chen and Lin, 2011). Robustness represents the conservative forces of the networks between generations under the influence of random perturbations. In contrast, evolvability represents the tolerated perturbations induced by genetic and epigenetic modifications, perturbations that could eventually be positively selected during evolution. However, the properties (power-law, small-world) used in network modelling have been contested recently, with the availability of more and more networks. The corresponding data do not always fit the expected theoretical models, and the cases where a good fit has been observed may often result from sampling artefacts or improper data representation (Lima-Mendez and van Helden, 2009).

5.4   Limits of current methodologies

5.4.1   How to integrate multiple biological levels in an evolutionary framework?

Currently, the main challenge in evolutionary studies is to develop an efficient framework for integrating data from multiple biological levels at the same time. As discussed in the previous paragraphs, this is essential because system-level phenomena have an important influence on evolutionary outcomes. Most evolutionary theories and principles are however related to a single level (e.g. genome or network topology evolution). Some general system-level trends have been

discovered, such as the correlation between expression levels, protein abundance, network centrality of genes (figure 5-7), but these conclusions were mainly extracted from a manual compilation of several independent studies (Koonin and Wolf, 2006). Classical evolutionary tools are not adapted to studying such large-scale correlations.

—	NP	PPI	GI	EL	CAI	PA	KE	PGL	ER
NP	*								
PPI	++	*							
GI	++	+	*						
EL	+++	+++	–	*					
CAI	ND	+++	ND	+++	*				
PA	ND	+++	ND	+++	+++	*			
KE	+	+++	–	+++	+++	+++	*		
PGL	NS	– –	NS	– – –	ND	ND	– –	*	
ER	– –	– – – –	– – –	– – – –	– – – –	– –	+++	*	

Figure 5-7. Evolutionary correlations between multiple biological parameters. The plus signs indicate positive correlations, and the minus signs indicate negative correlations. CAI, codon adaptation index; EL, expression level; ER, evolutionary rate; GI, number of genetic interactions; KE, lethal effect of gene knockout; NP, number of paralogs; PA, protein abundance; PGL, propensity for gene loss; PPI, number of physical protein–protein interaction partners. ND, not determined; NS, not significant. Adapted from Koonin and Wolf, 2006.

In fact, to our knowledge, no standardized approaches exist for integrating heterogeneous multi-scale evolutionary data in a single framework. Moreover, no tools exist to describe gene or network evolutionary histories based on genome-scale multilevel datasets.

5.4.2 How to formalize evolutionary variation in biological networks?

Systems biology constantly producies new genome-scale datasets and managing this flux is a major challenge for evolutionary biology, like other domains. Some attempts have been made to manage the data with high-throughput genome-scale phylogenies (Levasseur et al., 2012a; Ruan et al., 2008; Wapinski et al., 2007). However, these trees only consider genomic data. Moreover, they are binary trees, based on genomic distances and the problem of estimating systems-level distances between multiple organisms remains. The classical tree representation can also be a limiting factor in itself, as demonstrated by the upgrade of phylogenetic trees to phylogenomic networks in bacterial studies. Indeed, the high plasticity of biological networks is difficult to formalize. Some inspiration could come from the systems biology attempts to represent the spatiotemporal dynamics of interactomic networks (Goel et al., 2011), but this representation is still based on static networks and is generally manually constructed. New tools are clearly needed to automatically generate system-level evolutionary histories.

6   MATERIAL AND METHODS

The results described in chapters VII to IX were obtained using numerous methods and algorithms for data retrieval, software development, sequence alignment and biological knowledge extraction. This work was performed using the existing infrastructures and computer resources of the Laboratoire de Bioinformatique et Génomique Intégratives (LBGI), the Plate-forme de Bioinformatique de Strasbourg (BIPS) and the Décrypthon computing grid (Bard et al., 2010). The BIPS is a high-throughput platform for comparative and structural genomics, member of the Réseau National des plates-formes Bioinformatiques (ReNaBi).

6.1   Computing resources

6.1.1   Servers

Computational power was provided by the central servers of the Institut de Génétique, Biologie Moléculaire et Cellulaire (IGBMC). In 2010, the institute renewed its computational infrastructure, implementing a 'blade centre / master server / storage server' architecture (figure 6-1).
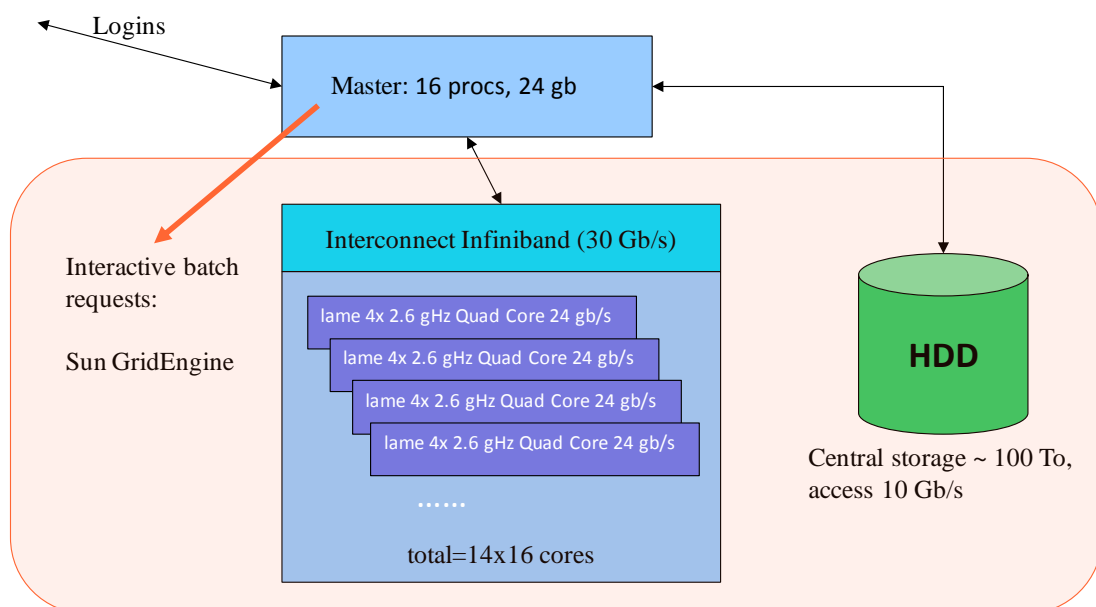
Figure 6-1. The "blade centre / master server / storage server" architecture at the IGBMC.

In this architecture, a master server controls all input/output user connections and allows interactions between the different elements of the architecture. Storage is centralized in a dedicated server. Computational power is provided by a set of 'blades', i.e. a set of physical cards dedicated to CPU computational power (4x quad core processors per blade). An interactive batch job management system, the Sun GridEngine, runs on the master server and is accessible for registered users. Using this queuing system, jobs can be sent directly to the server and the master server automatically shares available CPUs depending on job requirements. This allows a balanced sharing of CPU resources between users.

6.1.2 Décrypthon Grid

The Décrypthon is a computational grid supported by the French Muscular Dystrophy Association (Association Française contre les Myopathies: AFM), allowing seamless access to computation and storage resources for application developers and scientists (Bard et al., 2010). The project is a partnership between 3 organisms: the AFM and IBM, who have been partners since 2001, and the CNRS, who joined the project in 2005. Currently, the grid computational power is based on several supercomputers installed by IBM in 6 French universities providing a total of 500 Gflop (Bordeaux 1, Lille 1, Paris 6 Jussieu, ENS Lyon, Crihan in Rouen, Orsay) and individual personal computers via a world community grid. The first ortholog dataset built using OrthoInspector, presented in Chapter VII, is based on a BLAST all-against-all of 60 eukaryotic species calculated on this grid. We also used the Décrypthon grid services during the multiple alignment construction for the EvoluCodes (Chapter VIII), since they required data from 20 vertebrate proteomes (> 500 000 sequences).

6.1.3 Database systems

A database can be defined as a structured collection of data. The data are organized according to a data model describing reality concepts, objects and relations. The data model is thus an abstract structure that provides the means to effectively describe the specific data structures required by an application. A database is generally supported by a DataBase Management System (DBMS). The data collection together with the DBMS is called a database system. In computer sciences, the relational model is now the most widely used data model and is supported by a standard querying language, called SQL (Structured Query Language), facilitating database interoperability.

6.1.3.1 General database systems

Three relational database systems were used in the different software and methods developed during this thesis: MySQL (www.mysql.com), postgreSQL (www.postgresql.org) and IBM DB2 (http://www-01.ibm.com/software/data/db2/). Both MySQL and postgreSQL are open source systems, widely used in computer engineering and web development. IBM DB2 is proprietary software developed by IBM Corporation and is business oriented. OrthoInspector implements an intrinsic support for PostgreSQL and MySQL, allowing automated installation of a new orthology database with these systems. The pipeline used to produce the EvoluCodes is supported by a

PostgreSQL database. The IBM DB2 system was used for EvoluCode exploitation and analysis, due to the fact that his system includes a wide set of tools for data mining, modeling, scoring and visualization: the Intelligent Miner software suite.

6.1.3.2   BIRD database system

The BIRD system was developed in the laboratory by Ngoc Hoan Nguyen and is designed to manage and integrate heterogeneous biological data. BIRD is a software layer, creating and managing genomic, transcriptomic and proteomic resources with the help of a configurable data model. It integrates an ontology driven API and a set of database rule analysers. The integration rules allow the user to easily create a database based on these semantics and his specific needs. BIRD is based on the initial ideas and concepts developed in the Saada project (http://www.projet-plume.org/fiche/saada) devoted to the management of astronomical data. The system has been developed using Java technology. In its current version, BIRD uses IBM DB2 to store data and operate with a powerful full-text and XML data model. The web application can be hosted by a Tomcat server or an IBM WebSphere Application Server. In addition, BIRD is supported by a biology oriented search engine called BIRD-QL that was developed to facilitate access to databases and the extraction of relevant information. It allows the biologist to easily express queries and to extract knowledge by classical constraints and scientific functions. Many routine scripts developed during this thesis use the BIRD system for data retrieval.

6.2   Bioinformatics resources

6.2.1   Biological databases

6.2.1.1   General sequence databases

We used three well-known biological databases to extract sequence data: Uniprot (Apweiler and constortium, 2012), RefseqP (Sayers et al., 2012) and Ensembl (version 54) (Hubbard et al., 2007).

The Universal Protein Resource (UniProt) is a comprehensive resource for protein sequence and annotation data. It is composed of several databases, mainly the UniProt Knowledgebase (UniProtKB), the UniProt Reference Clusters (UniRef), and the UniProt Archive (UniParc). UniProt is the result of a collaboration between the European Bioinformatics Institute (EBI), the SIB Swiss Institute of Bioinformatics and the Protein Information Resource (PIR).

The National Center for Biotechnology Information (NCBI) provides the Reference Sequence (RefSeq) database, a collection of taxonomically diverse, non-redundant and annotated sequences. Each RefSeq release is constructed from sequence data submitted to the International Nucleotide Sequence Database Collaboration (INSDC). A portion of the RefSeq dataset is then curated by NCBI staff and collaborating groups. The RefSeq section dedicated to protein sequences in commonly referred to as RefSeqP.

The goal of the Ensembl database is to automatically annotate genomes, integrate the annotation with other available biological data and make all this publicly available via the web. Originally,

Ensembl focused on vertebrate genomes with 64 consecutive releases of the database. Today, Ensembl has expanded to include all phyla, with 5 new databases (EnsemblBacteria, EnsemblProtists, EnsemblPlants, EnsemblFungi, EnsemblMetazoa) accessible via a common entry point called EnsemblGenomes (http://www.ensemblgenomes.org/). Ensembl is a joint project between the European Bioinformatics Institute (EBI) and the Wellcome Trust Sanger Institute (WTSI).

The first and second versions of the OrthoInspector database are composed of eukaryotic proteomes downloaded from Uniprot, RefseqP and Ensembl (version 54 and 64) databases. In the first version, we avoided multiple transcript issues in two different ways. For Ensembl data, the longest protein sequence was selected for each predicted gene annotated as 'protein-coding'. For Uniprot and RefseqP data, each sequence was compared to all others from the same organism using BLAST and excluded if a longer sequence was found sharing more than 99% identity. Manually-annotated entries from Swissprot were preferred over TrEMBL and RefseqP entries. The second version of OrthoInspector uses the same protocol for data from Ensembl. However, for other available genomes, we selected complete genomes referenced in the new 'reference_proteome' dataset provided by Uniprot. The 'reference_proteome' database references organisms with completely sequenced genomes and provides curated reference proteomes, i.e. proteomes with one representative transcript for each protein-coding gene.

The EvoluCodes are based on a human protein set retrieved from the Human Protein Initiative (HPI) project (O'Donovan et al., 2001). This project defined a master human proteome set with quality standards based on UniprotKB/Swiss-Prot databases. To construct the EvoluCodes, we chose 16 vertebrate proteomes, by selecting species that best represent major vertebrate phyla, i.e., fish, batracia, sauropsida and mammals. The complete proteomes for these organisms were downloaded from Ensembl (version 51) and a local database was created with more than 500,000 vertebrate sequences.

6.2.1.2   Annotation database

The MACSIMS annotation process (see 6.2.3.2) used in the construction of the EvoluCodes incorporates Gene Ontology (Ashburner et al., 2000) annotations and protein domain definitions from the Pfam database (Punta et al., 2012a). Functional enrichment analysis of EvoluCodes was also based on GO annotations.

The Gene Ontology (GO) project aims to create consistent descriptions of gene products in different databases. The project has developed three structured controlled vocabularies (ontologies) that describe genes in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner.

The Pfam database is a collection of protein domain families. Each family is represented by multiple sequence alignments and hidden Markov models (HMMs). The Pfam database is divided into two parts: Pfam-A and Pfam-B. Pfam-B families are generated automatically with sequences from most major databases and are un-annotated. Pfam-A entries contain computational predictions inferred from the most recent release of UniProtKB at a given time-point. Each Pfam-A family is seeded on a curated alignment containing a small set of representative members of the family, from which a Hidden Markov Model profile (profile HMMs) is built. This profile is then used to aggregate new sequences and to build a new, more comprehensive alignment.

6.2.1.3 Pathway database

The system-level analysis presented in chapter IX explores human pathway maps defined by the Kyoto Encyclopedia of Genes and Genomes (KEGG) knowledge base (Kanehisa et al., 2012). KEGG aims to build computer representation of biological systems, consisting of molecular building blocks of genes and proteins (genomic information) and chemical substances (chemical information). These blocks are combined with biological knowledge of molecular wiring from the literature to form diagrams of interaction, reaction and relation networks (systems information, figure 6-2).
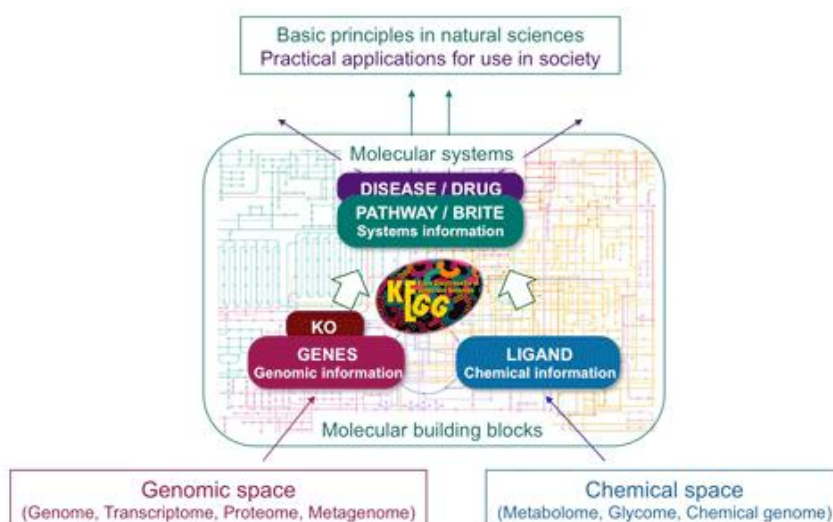
Figure 6-2. The different types of biological information integrated in the KEGG database. *Adapted from http://www.kegg.jp/.*

Today, the knowledge base is composed of seventeen databases categorized into systems information, genomic information and chemical information. The EvoluCode analysis was performed in the context of the KEGG PATHWAY database. This is a collection of manually drawn pathway maps, representing molecular interactions and reaction networks for different biological processes, such as metabolism, genetic information processing, environmental information processing, cellular processes or organismal systems. KEGG data were retrieved with the help of the KEGG SOAP (Simple Object Access Protocol) server (http://www.kegg.jp/kegg/soap/).

6.2.2 Sequence aligners

The BLAST (Basic Local Alignment Search Tool) includes a sequence comparison algorithm providing local pairwise alignment for biological sequences (McGinnis and Madden, 2004). There exist specialized versions of the algorithm dedicated to protein or nucleotide query sequences and sequence libraries (BLASTn, BLASTp, tBLASTn …). The complete software suite of BLAST tools has been recently rewritten in C++ for better performance and compatibility with state-of-the-art software systems. This update is named BLAST+ and will eventually replace the traditional BLAST

tools that were written in C (Camacho et al., 2009). We exploited this new version for the creation of the BLAST all-against-all used for OrthoInspector predictions.

6.2.3   Expert systems

6.2.3.1   PipeAlign: a protein family analysis tool

PipeAlign is a tool developed in the LBGI for the automated analysis of protein families (Plewniak et al., 2003) through the construction of multiple alignment of complete sequences (MACS). The pipeline integrates a six step process, corresponding to six different sequence analysis programs, ranging from the search for homolog sequences in protein and 3D structure databases, to the construction of hierarchical relationships within and between families (figure 6-3).
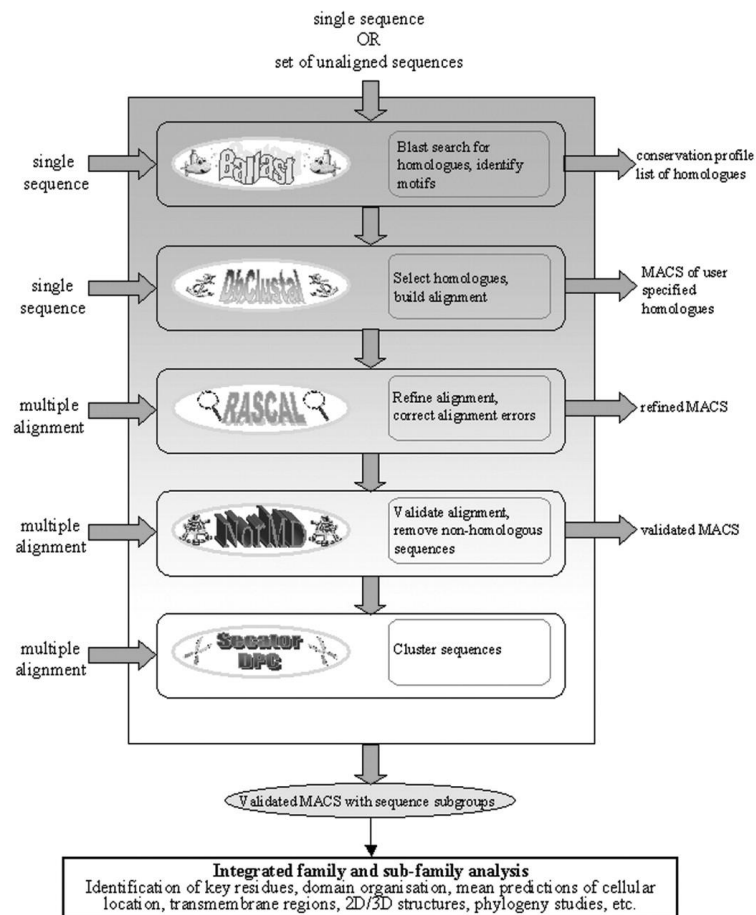
Figure 6-3. Overview of PipeAlign pipeline. *Adapted from Plewniak et al., 2003.*

To produce the MACS (Multiple Alignment of Complete Sequences) used in the construction of the EvoluCodes, we used a modified version of the published PipeAlign, where DbClustal was replaced by the MAFFT program (Katoh et al., 2002) since the computational speed of MAFFT is better suited to high throughput projects. Thus, the pipeline can be summarized as follows:

1. **Database searches with BLASTp and post-processing of results with Ballast:**

   Given an input sequence, a BLASTp search of the Uniprot database is performed. Ballast then builds a conservation profile of the database sequences detected by BLASTp with an E-value <10. The contribution of each database hit is proportional to its E-value. Then, Local Maximum segments (LMSs) are identified, corresponding to sequence segments that are more conserved than their neighbouring regions. The position of the LMSs in the query and database sequences are identified and are stored in a file as a list of anchors for input to MAFFT.

2. **Construction of the MACS with MAFFT:**

   Mafft is a suite of programs offering various multiple alignment strategies, of which two complementary versions were tested: a rapid, less accurate version (fftns2) and an iterative refinement (linsi) (figure 6-4). See publication n°4 for more details concerning MAFFT performance.
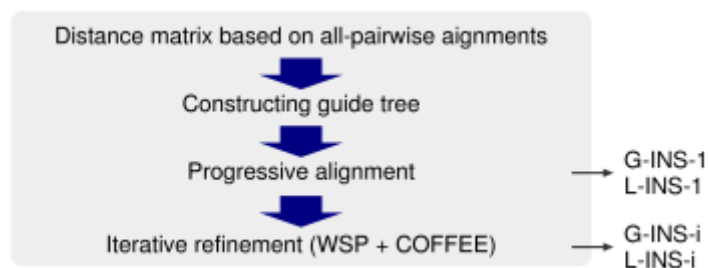
Figure 6-4. Principle of the linsi alignment strategy of MAFFT.

In linsi, alignments are scored using an objective evaluation function combining a weighted sum-of-pairs (WSP) score (Katoh and Toh, 2008a) and a COFFEE-like score (Wallace et al., 2006b), which evaluates the consistency between a multiple alignment and a predefined set of pairwise alignments (Katoh et al., 2005). The alignment is progressively refined until the score reaches a defined threshold. The linsi method is particularly suitable for aligning sets of sequences that contain large non-homologous regions, such as multi-domain protein families.

3. **Correction of errors with RASCAL:**

   The RASCAL program is designed to detect and correct local errors introduced in the MACS (Thompson et al., 2003). The multiple alignment is divided horizontally and vertically to form the blocks of a lattice in which the well aligned regions are differentiated. For an efficient refinement strategy, alignment correction is limited to the less reliable regions. Thus, potential alignment errors are detected by comparing profiles of the different blocks. RASCAL then performs a single re-alignment of each badly aligned region using an algorithm similar to that implemented in ClustalW (Larkin et al., 2007).

4. **Alignment-based homology evaluation with LEON:**

   The LEON program (Thompson et al., 2004)detects sequences that are unrelated to the query sequence. This step is necessary because we include all sequences detected by BLASTp with E-value <10. Although this allows us to incorporate very divergent sequences, it also

introduces some unrelated ones. LEON uses the conserved blocks detected by RASCAL and chains them into longer conserved regions that may correspond to homologous structural/functional domains. Then, sequences with no homologous regions are removed from the MACS.

5. **Quality evaluation with NorMD:**
   NorMD is an objective function for MACS quality evaluation (Thompson et al., 2001). It combines the advantages of both column-scoring techniques and residue similarity scores and is based on the Mean Distance (MD) scores introduced in ClustalX (Thompson et al., 2002). The score is normalised for various factors, including the number of sequences, the alignment length and the presence of gaps, and allows us to define a cutoff above which a MACS is considered of high quality.

6. **Sequence clustering with Secator:**
   The Secator algorithm clusters sequences in a MACS into potential functional subgroups (Wicker et al., 2001). First, it creates a phylogenetic tree from a distance matrix based on the MACS. Then, it assigns a dissimilarity value to each node of the tree and collapses branches to automatically detect nodes joining distant subtrees. The remaining subtrees represent sequence families in the alignment.

The final output of this pipeline is a high-quality MACS with sequences clustered into potential functional sub-families. The whole procedure is available online at http://bips.u-strasbg.fr/PipeAlign.

6.2.3.2   MACSIMS: comprehensive annotation of multiple alignments

MACSIMS (Multiple Alignment of Complete Sequence Information Management System) is an expert system for the management of all the information related to a protein family (Thompson et al., 2006). It provides an environment that facilitates knowledge extraction from a MACS and the presentation of the most pertinent biological information. This extraction can be split into 4 main parts: data retrieval for sequence features, homology analysis, data validation and data propagation (figure 6-5).

The homology analysis is similar to that integrated in PipeAlign (see 6.2.3.1), using Secator and RASCAL to define high quality aligned regions. Sequence features are compiled from several databases such as GO annotations (Ashburner et al., 2000), Pfam (Punta et al., 2012a) and Prosite domains (Sigrist et al., 2010), PDB secondary structures (Velankar et al., 2012), etc. These are augmented with ab initio calculated characteristics, such as GES hydrophobicity (Engelman et al., 1986), coiled coil predictions, etc. Then, the homology and annotation data are combined to verify and validate all annotated features in the context of the multiple alignment and with the help of a first decision tree (figure 6-6 A). Briefly, each feature associated with each sequence is compared with features for the other sequences in the same sub-group and, based on the first decision tree, the reliability of the feature is investigated. The reliable features are then propagated in the alignment, using a second decision tree (figure 6-6 B). The extended annotation is saved in an XML format, following the MAO (Multiple Alignment Ontology) schema (Thompson et al., 2005b). The

interoperability provided by the MAO format facilitates the analysis and visualisation of propagated features, for example in the Jalview alignment editor.
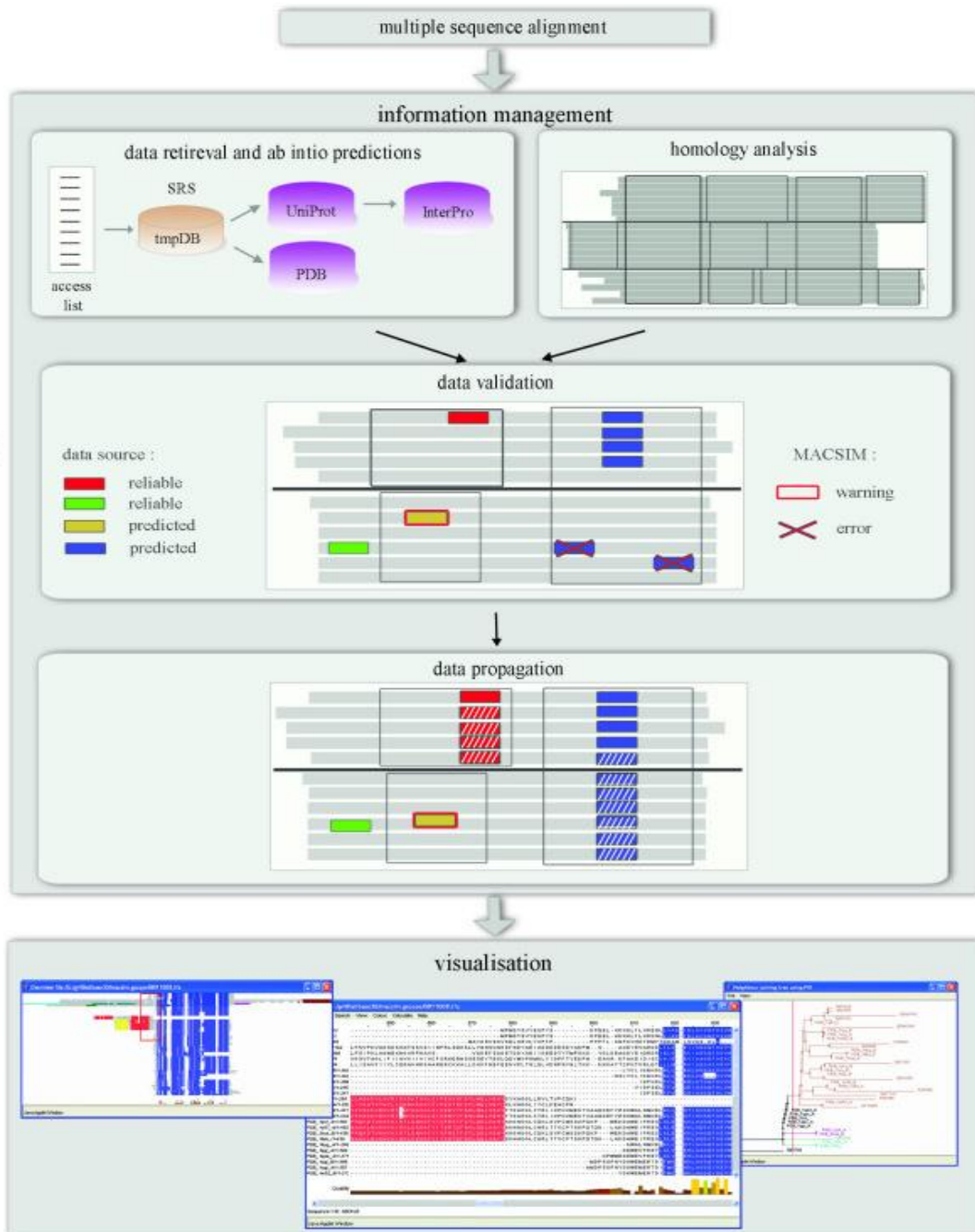
Figure 6-5. Overview of the MACSIMS modules. *Adapted from Thompson et al., 2006.*

A. Feature validation
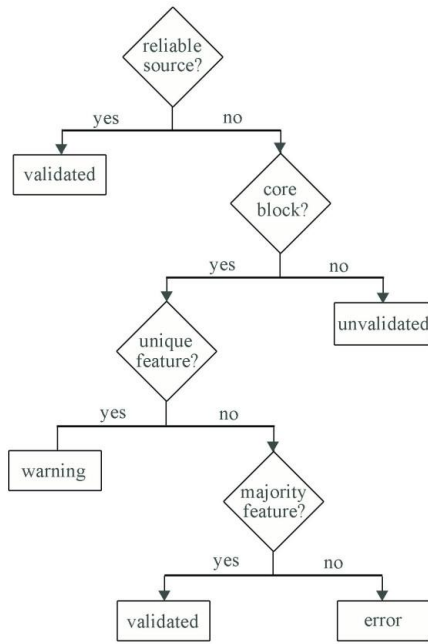
B. Feature propagation
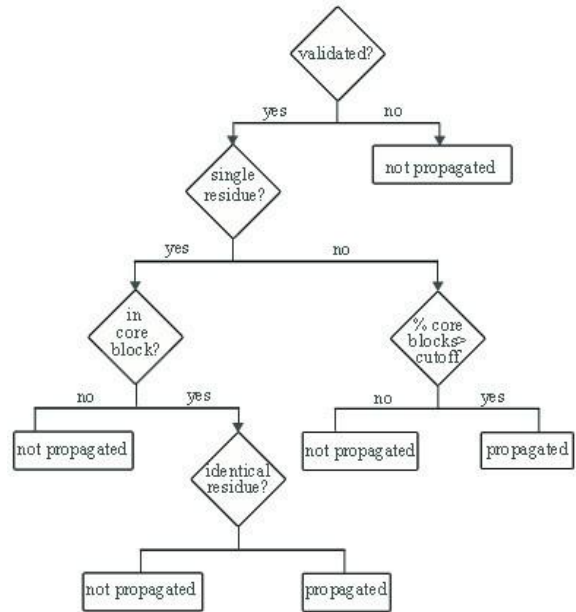
Figure 6-6. MACSIMS decision trees. *(A) A first decision is made to validate or discard a particular feature associated with each sequence included in the MACS. (B) Validated features can be propagated to other sequences based on the homologous block definition using the rules of a second decision tree. Adapted from Thompson et al., 2006.*

6.2.4 Data visualisation

6.2.4.1 Jalview

Jalview is a multiple alignment editor written in Java (Waterhouse et al., 2009). It is used widely in a variety of web pages (EBI ClustalW server, Pfam protein domain database, etc.) but it is also available as a general purpose alignment editor and analysis workbench. It contains numerous tools to view, edit, analyse (e.g. alignments, phylogenetic trees, etc.), annotate (e.g. secondary structures, colour schemes for features) and publish (image, web) MACS related data. We used Jalview to display MACS annotated by MACSIMS on the EvoluCodes website. In the display, each sequence feature provided by MACSIMS is associated with a colour code, allowing human analysis of conserved blocks, domains and key residues.

6.2.4.2 OrdAlie

OrdAlie (Ordered Alignment Information Explorer) is developed in the laboratory by Luc Moulinier. It is a Tcl/Tk program designed to allow a comprehensive analysis and exploration of protein sequences, structures and functions, as well as their evolutionary relationships. Sequences can be clustered automatically into sub-families and a hierarchical analysis of residue conservation can be

performed in each family. Moreover, this information can be viewed in the context of the 3D structure, using the RasMol structure tool (Goodsell, 2005).

6.2.4.3   *Cytoscape*

Cytoscape is an open source software platform for visualizing complex networks and integrating these with any type of attribute data (Smoot et al., 2011). Cytoscape is particularly efficient for large-scale network visualization and can produce publication quality annotations. Another advantage of Cytoscape is its plugin system. Many authors regularly publish new Cytoscape plugins dedicated to specific biological network analysis protocols or add new interfaces to other software. Cytoscape was used to analyse and visualize the evolutionary networks described in chapter 9.

6.2.5   Methods for knowledge extraction

6.2.5.1   *GoMiner for GO enrichment analysis*

We performed a functional enrichment analysis to analyse clusters of Evolucodes describing a similar evolutionary history (see chapter XIII). To do so, we used the GoMiner program, a tool for biological interpretation of 'omics' data, including data from gene expression microarrays (Zeeberg et al., 2003). GoMiner exploits the Gene Ontology (GO) to identify the biological processes, functions and components represented in a gene list. This software contains several clustering and visualisation tools in a graphical interface. For our analysis, we used the command-line version of the program as we were mainly interested in the quantitative and statistical significance of functional enrichment in EvoluCode clusters.

6.2.5.2   *IBM Intelligent Miner for data mining*

Intelligent Miner is an IBM software suite designed for data mining in IBM DB2 databases. It enables users to mine structured data stored in conventional databases or flat files. Its mining algorithms aim to address business problems in such areas as customer relationship marketing or fraud and abuse detection. We used this software suite to investigate the presence of some data structures or models in our EvoluCodes because the 2D matrices can also be represented as database tables. In particular, we analysed our data with self-organizing maps, also known as Kohonen maps, a type of artificial neural network that classifies the training data without any external supervision and produces a low-dimensional representation of the input space of the training samples (Kohonen, 1988). The advantage of this approach is the ability to represent complex multi-dimensional data in a low-dimensional representation (typically 2D maps, see figure 6-7). The topology of the 2D representation is important, since a self-organizing map uses a neighbourhood function to preserve the topological properties of the input space. Similarly, the local density of data is conserved as a data space with more data corresponds to a larger area of the map.
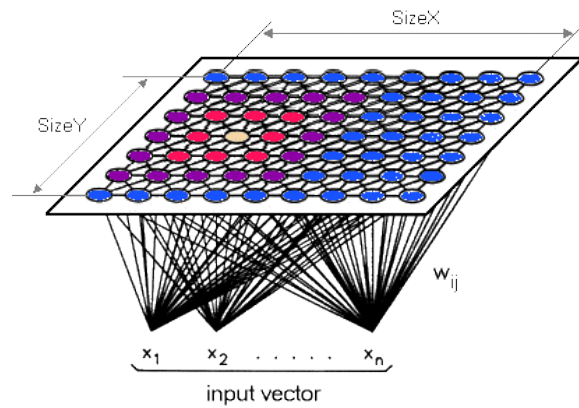
Figure 6-7. Schematic representation of the self-organising map (SOM) projections. *The SOM defines a mapping of the input data space of dimension n to a two-dimensional (X x Y) array of nodes. Adapted from http://www.lohninger.com/helpcsuite/.*

6.2.5.3 LOF: Local Outlier Factor for outlier analysis

In chapter 9, we define 'outlier' EvoluCodes, i.e. outlier evolutionary histories in the context of their biological network using the Local Outlier Factor (LOF) score (Breunig et al., 2000). The LOF assigns to each object a degree of ' outlierness', depending on how isolated the object is with respect to the surrounding neighbourhood (Figure 6-8). The idea of the LOF is to compare the local density of a point's neighbourhood with the local density of its neighbours. The LOF algorithm takes as input the distance from all entities to their k nearest neighbours. Consequently, it can be used with high dimensional spaces, if the user can provide such distances. The 2D projection of figure 6-8 is an example of a LOF calculation for a 2D dataset containing one low density Gaussian cluster of 200 objects and three large clusters of 500 objects each, illustrating the notion of degree of outlierness.
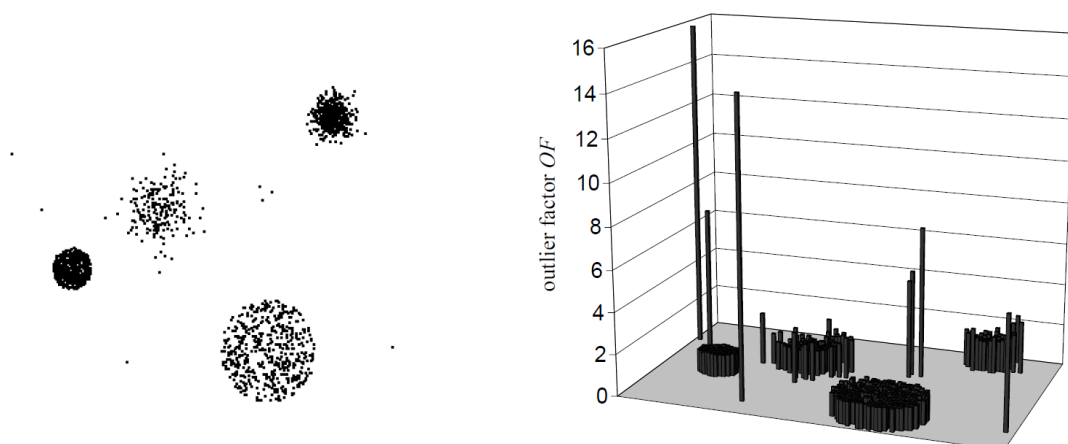
Figure 6-8. Local outlier factor score for points in a sample dataset. *The LOF score attributes a degree of outlierness based on the local point density. Each point of the dataset (left plot) is associated with one LOF score (right graph). Adapted from Breunig et al., 2000.*

6.3   Software development

6.3.1   Java programming

6.3.1.1   Java language

Java is an object-oriented programming language developed by Sun MicroSystems, a filial of Oracle Corporation. It is widely used in mobile applications, games, web-based content and enterprise software. Java applications are semi-compiled to bytecode that is run by an interpreter program generally called 'JAVA Virtual Machine' (JVM) for their execution. This characteristic allows a system-independent run of JAVA programs and a faster portability to different systems, but at the expense of a small loss of performance (ex: C code is 1.5x faster than equivalent java code).

6.3.1.2   Netbeans platform

Netbeans is an open-source Integrated Development Environment (IDE). It has a built-in support for Java platforms, as well as for C/C++, PHP, JavaScript and Groovy. Netbeans is particularly useful for creating visual interfaces because it integrates a standards-based user interface with the NetBeans Swing GUI Builder (figure 6-9).
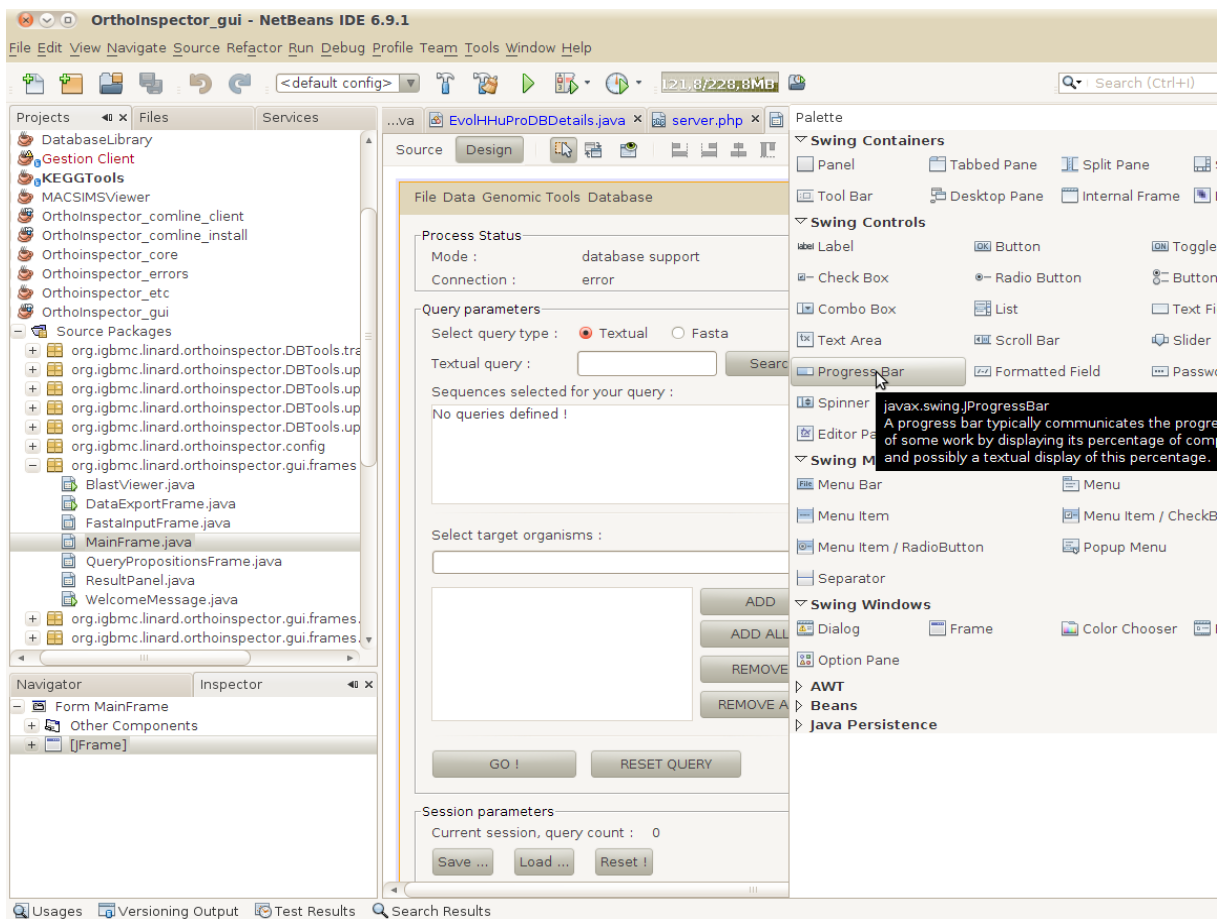
Figure 6-9. Screenshot of the Swing GUI Builder. *The builder facilitates the addition of new components in the interface (middle panel) by a direct drag-and-drop of these elements from a palette (right panel).*

This builder facilitates and accelerates the creation of user-friendly graphical interfaces. Most of the work presented in this thesis was developed in the Netbeans environment.

6.3.1.3   Java Libraries

JAVA programs developed during this thesis include several open-source libraries. The following table lists these libraries.

Library	Ortho-Inspector	Evolucode	Usage	Reference
GradientPainter		yes	Colour gradient scaling for EvoluCode biological parameters	users.erols.com/ziring /java-samp-jgd.html
Jacksum	yes		Checksum transformation of sequences for faster access	sourceforge.net/proje cts/jacksum/
Jdom	yes	yes	Reader/writer for XML format	www.jdom.org
Jung	yes		Graphical and mathematical library for graph analysis and visualisation	jung.sourceforge.net
mysql-connector	yes	yes	MySQL database connection driver	www.mysql.com/prod ucts/connector
OpenCSV	yes		Reader/writer for CSV format	opencsv.sourceforge.n et
OrthoXML	yes		Reader/writer for XML file format based on OrthoXML ontology	(Schmitt et al., 2011) orthoxml.org
postgresql-jdbc	yes	yes	PostgreSQL database connection driver	jdbc.postgresql.org
prefuse	yes		Dynamic graph display and visualisation tools	prefuse.org
rJava		yes	Interface between Java and R processes	www.rforge.net/rJava
venneuler	yes		Venn and Euler diagram calculation and display	(Wilkinson, 2012a)

Table 6-1. Java libraries used in software development.

6.3.2   R programming

6.3.2.1   R language

R is a language and environment for statistical computing and graphics (www.r-project.org/). A compiled installation of R is maintained worldwide by a huge contributor community and provides a wide variety of statistical and graphical techniques. R is particularly useful for rapidly producing publication-quality plots, including mathematical symbols and formulae where needed. Its native C sources allow fast computations and regular updates with new statistical packages produced by users.

6.3.2.2 R librairies

The creation and exploitation of Evolucodes use R scripts that incorporate several libraries. The following table lists these libraries:

Library	Usage	Reference
FactoMineR	Multivariate analysis to decipher most informational parameters in barcodes	(Le et al., 2008) factominer.free.fr
rLOF	Density-based estimation of outlierness degree for Evolucodes (scoring)	(Breunig et al., 2000) cran.r-project.org/web/ packages/Rlof

Table 6-2. R libraries used in software development.

6.3.3 Web development

6.3.3.1 HTML / PHP / CSS

HTML (HyperText Markup Language) is a standard and the main markup language for displaying web pages and other information that can be displayed in a web browser. The main purpose of a web browser is to read HTML documents to transform them into visible web pages. CSS (Cascading Style Sheets) is a standard for defining the appearance and layout of text and other material in a web page. The W3C (World Wide Web Consortium) maintains both the HTML and the CSS standards.

PHP (recursive acronym for PHP: Hypertext Preprocessor) is a widely-used open source general-purpose scripting language that is especially suitable for web development and can be embedded into HTML pages. In contrast to other web-related language such as javascript or java, PHP is executed on the server side and the results of its execution are sent back to the client, saving computational operations for the users. The combination of HTML, CSS and PHP code is the prominent state of the art for web development. Both OrthoInspector and EvoluCodes websites are coded with this combination.

6.3.3.2 Javascript, JQuery and Ajax

JavaScript is a multi-paradigm scripting language, supporting object-oriented, imperative and functional programming styles. JavaScript is mainly used on the client-side for web communication, allowing programmatic access to computational objects within the host environment. Ajax (Asynchronous JavaScript and XML) is a group of web development techniques used on the client-side to create asynchronous web applications. It allows to send and retrieve data in an asynchronous manner without interfering with the display and operations of the current web page.

jQuery is a cross-browser open-source JavaScript library. It is designed to facilitate HTML document navigation, create animations or handle events and develop Ajax applications. It is generally used to create dynamic content in web pages, such as multiple process searches in a database with a roller animation that runs until the data fill the current web page, or the query propositions that appear

when typing text in a textfield. The OrthoInspector and EvoluHHupro projects use these technologies for efficient website navigation and to limit long database searches.

6.4 Analysis protocols

6.4.1 Construction of the BLAST all-vs-all

A BLAST all-vs-all is the complete set of BLAST searches corresponding to a set of proteomes. In the context of the OrthoInspector orthology predictions, we first constructed a database containing all proteomes for the studied organisms. Prior to the BLAST all-vs-all construction, redundant sequences were removed. Each sequence was then compared to all others from the same organism using BLAST. For sequences sharing more than 99% identity, manually-annotated entries from Swissprot were preferred over TrEMBL and RefseqP entries, otherwise the longest sequence was retained. The new NCBI-Blast+ package (Camacho et al., 2009) was used to perform BLAST all-versus-all searches between the proteomes with an E-value cutoff of 1e-9. For the first version of OrthoInspector, the searches were executed on the Décrypthon grid resources (Bard et al., 2010).

6.4.2 Local genome neighbourhood conservation for EvoluCodes

The chromosomal localization of all genes coding for the protein sequences was obtained from Ensembl. Locally developed software was used to identify conserved local synteny between the human genome and each of the 16 other vertebrate genomes. To do this, the chromosomes in each genome are represented as a linear sequence of genes. For each human reference sequence, the local syntenic homolog HREF was defined at position i on the human genome and its upstream and downstream neighbours (HREF-1and HREF+1 respectively) were identified. For each of the 16 vertebrate genomes, the sequences with the highest similarity to HREF-1 and HREF+1 were selected from the MSA, and denoted Vn_Sim-1 and Vn_Sim+1 respectively, where Vn refers to one of the 16 vertebrate genomes. A local synteny homolog, exists for HREF and genome Vn if:

- homologs were found in Vn for HREF-1 and HREF+1,
- the separation between the highest similarity homologs, denoted Vn_Sim-1 and Vn_Sim+1, on the genome was less than 5 genes,
- a homolog of HREF was found on the genome between Vn_Sim-1 and Vn_Sim+1.

The homolog of HREF localized between Vn_Sim-1 and Vn_Sim+1 with the highest similarity to the human reference sequence was then defined as the syntenic homolog. Genes with ambiguous genomic locations, such as scaffolds etc, were discarded since the synteny relationship could not be reliably established. In addition, local or tandem duplications were excluded since the genome contexts of the two gene copies were similar.

7 ORTHOINSPECTOR: COMPREHENSIVE ANALYSIS OF ORTHOLOGY RELATIONS

7.1 Introduction

In the systems biology era, the gene continues to play a central role in large-scale biological studies. This is at least partly due to the fact that, with the ever-increasing performance of sequencing techniques, the gene is now a cheap entity for multiple species comparisons. Genome sequencing has reached new scales and is providing unprecedented opportunities to compare multiple genomes from the same species (intra-population variation) or to perform phylogenetic studies between a rapidly growing number of phyla. In particular, the gene is essential for the exploration of new genomes and their structural and functional annotation.

In this context, homology-based functional inference systems are the most widely used approaches for genome annotation. In particular, the orthology relation represents the most reliable phylogenetic relation for functional annotation. Indeed, orthologous genes are generally assumed to retain similar functions, while paralogous genes are free to evolve differently. Nevertheless, high-quality automated inference of orthologs remains a challenge, because numerous genetic events can produce a complex gene evolutionary history (gene loss, gene fusion, different evolutionary rates…). Moreover, coping with the constant arrival of new genomes on a daily basis requires tremendous computational power, especially in the case of classical phylogenetic inference methods. Consequently, heuristic algorithms such as graph-based methods were developed as an alternative, less computationally intensive approach (for a complete review of orthology inference approaches see chapter III). Unfortunately, heuristic methods are used at the expense of resolution and have a limited sensitivity in the analysis of large gene families and distant species. Another issue is the lack of tools for comprehensive analysis and visualization of orthology predictions. Phylogenetic methods produce individual phylogenetic trees for each protein family, while graph-based methods generally produce flat files containing orthologous relations. Visualisation thus becomes a limiting factor for meaningful biological analyses, as more data require more summarization and better representation in order to highlight interesting evolutionary scenarios.

To address these problems, we developed OrthoInspector, a complete software suite for orthology inference and for comprehensive analysis and visualisation of the resulting data. OrthoInspector is a graph-based method based on a new algorithm which is focused on improving prediction sensitivity. It includes a graphical interface providing an intuitive user interface and several comparative genomics tools. The goal of OrthoInspector is to offer a fast and accurate solution for gene family studies or inter-species gene set analysis.

7.2 Design of OrthoInspector

7.2.1 Inparalogy as a basis to detect orthology

Numerous heuristic methods, generally known as graph-based methods, have been developed (see chapter III). For example, Inparanoid and OrthoMCL are well established orthology inference programs based on pairwise organism comparison and best-hit graph clustering respectively. In both algorithms, the orthology search space is based on the BLAST Reciprocal Best Hits (RBH) existing between genomes. Although RBH guarantees a high specificity, it is also too stringent when co-ortholog genes exist in large protein families. This is particularly true when comparing phylogenetically distant species in which multiple duplication events occurred. For this reason, we decided to consider the in-paralogy relation as the basal homologous relation for detecting orthology (rather than the RBH). The OrthoInspector approach is similar to Inparanoid in that it performs pairwise organism comparison. However, the search for inparalog genes (first step) in each species is done prior to the pairwise organism comparison (second step). Thus, inparalogs are clustered into groups, which are then compared based on the best hits linking them, assigning 1-to-1, 1-to-many or many-to-many relations orthologous relations. The third step uses a decision tree to exclude inconsistent relations. A benchmark study using more than 10 large protein families (e.g. CMGC and TKL kinase families) demonstrated that the OrthoInspector algorithm achieved its goal with a significant gain in sensitivity compared to Inparanoid and OrthoMCL, and a performance close to the phylogeny-based method used in Ensembl Compara (see publication n°1).

The OrthoInspector algorithm is implemented as a Java application and can be accessed via a command-line or a graphical interface. Both of these interfaces allow orthology inference, the installation of an orthology database and querying of the database via textual searches or BLAST sequence searches. The graphical interface is further complemented by novel representations that were developed for a comprehensive analysis of the predicted relations. OrthoInspector includes a number of tools for complex orthology queries, data summarization and data visualization. Two of these tools were described in the publication of the OrthoInspector algorithm (phylogenetic pattern queries and presence/absence diagrams). Other tools have been developed more recently, illustrating the fact that OrthoInspector, initially developed in 2009, is constantly maintained and updated with the integration of new concepts and software libraries.

7.2.2 Facilitating data extraction

Data extraction can be a difficult task when dealing with genome-scale datasets. For the particular case of orthologous genes, the user may want to focus on the phylogenetic patterns of a subset of genes or gene families related to a specific biological function. In OrthoInspector, high-throughput queries can be constructed in a dedicated menu using presence/absence criteria. For example, one can retrieve all genes of an organism A, that are shared by species B and C, but absent in species D and E. Using such combinations allows to retrieve genes corresponding to a specific phylogenetic pattern. Moreover, the user can choose which kind of orthology relation (1-to-1, 1-to-many or many-to-many) corresponding to the pattern should be retrieved. Selected data can be later exported in CSV, FASTA, XML (an OrthoInspector specific XML format). More recently, the OrthoXML

specification has been integrated in OrthoInspector with the help of the java OrthoXML library (Schmitt et al., 2011). Orthology predictions can be saved in XML files constructed with the ontology rules in most OrthoInspector tools. For more details of the OrthoXML format, see paragraph 3.4.3.

7.2.3 Automated processes for data visualization

The phylogenetic pattern queries in OrthoInspector are supported by several visualization tools facilitating human exploration of the data. The phylogenetic distribution of a gene family is a powerful representation for summarizing the duplication and loss events that occurred during the family's evolutionary history. OrthoInspector currently includes three visualization tools allowing a comprehensive analysis of these relations.

The first tool is designed to elucidate the potential sub-families that compose a larger gene family. In order to determine orthologous relations based on BLAST hits, one must choose a BLAST score or E-value threshold to limit the sequence search space for homologs. However, it is impossible to choose a threshold that fits all gene families, as some families are strongly conserved in many phyla and can be delimited by a stringent threshold while other gene families evolved at higher rates and require a more tolerant threshold.
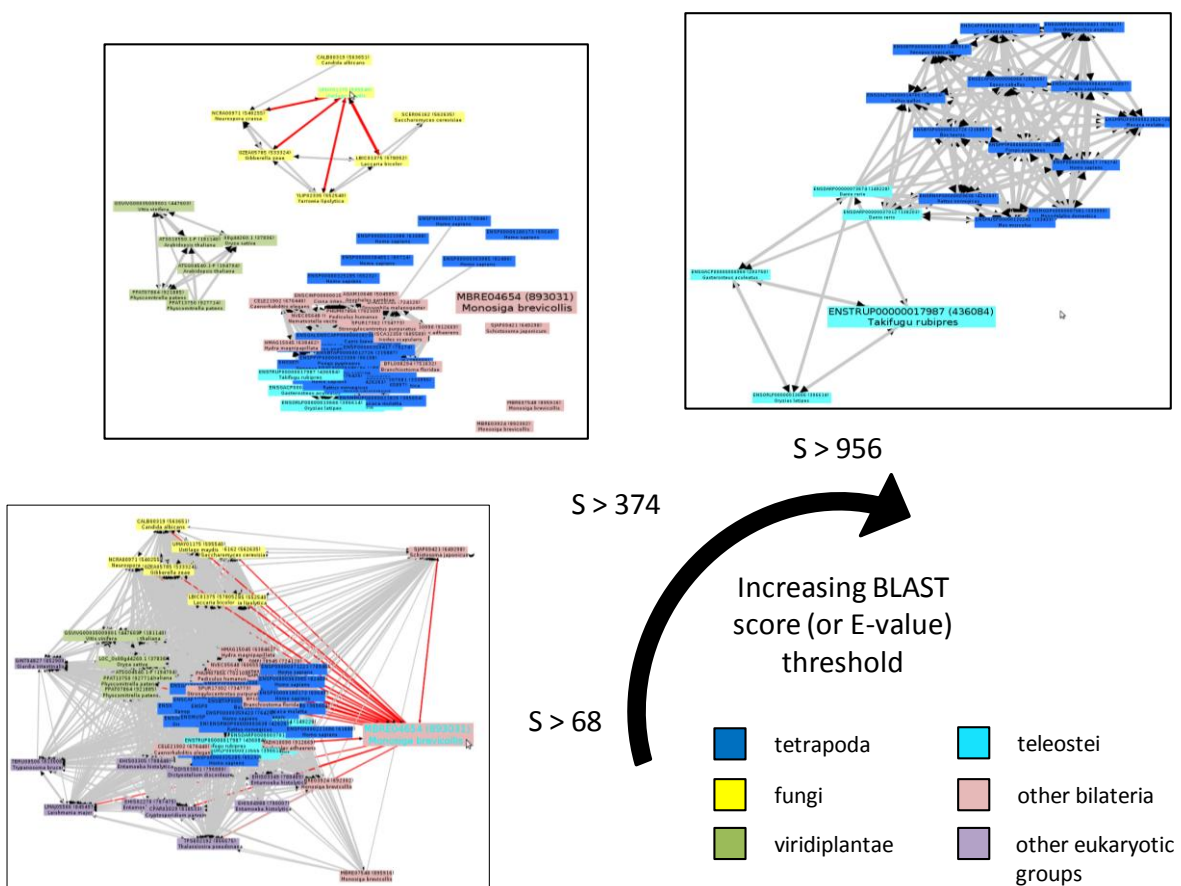
Figure 7-1. A BLAST threshold analysis in the OrthoInspector interface. The *BLAST score or E-value threshold is modified on the fly allowing a human visualization of the sequence similarity levels that separate different phyla in a given gene family. In this example, the score threshold is increased.*

To analyse this phenomenon and to better delimit gene sub-families, we created a graph representation of BLAST best hits linking the genes of a protein family. This representation is dynamic and allows to modify the BLAST score and E-value thresholds on the fly. As shown in figure 7-1, it is particularly useful for the identification of the threshold separating large phylum (plants, fungi, protists…) for a given gene family. This example corresponds to the BLAST hits of the myotubularin-related protein family retrieved using the human myotubularin MTMR1_HUMAN as a query. With a small threshold score (S>68), myotubularins in all phyla are linked by BLAST hits. Increasing this threshold (S>374) separates viridiplantae, fungi and bilateria genes. Increasing this threshold again (S>956), links only fish and other vertebrate myotubularins and highlights two inparalogs in *Danio rerio* that have the highest sequence similarity between animals and fishes. A similar delimitation based Blast E-value can be performed.

Figure 7-2. An extract from the phylogenetic distribution diagram corresponding to MTMR1_HUMAN. *Species are shown in rows and gene family members are in columns. Fusion of a row indicates a single gene co-ortholog for all inparalogs of the query. Several gene IDs in a single box indicate a lineage independent duplication.*

A second visualization tool for gene families is the phylogenetic pattern diagram tool. OrthoInspector can automatically generate such diagrams by using a single gene query and expanding the ortholog search to all related inparalog groups in the compared species. The example in figure 7-2 shows the phylogenetic distribution diagram constructed from the MTMR1_HUMAN query. All 8 human myotubularins are inparalogs compared to the unique myotubularin of fungi and protists. With the help of such relations, the diagram is automatically expanded to the whole myotubularin family (8 in humans, 3 in fishes, 1 in fungi and protists, etc.). The generated diagram identifies, for example, a myotubularin loss in rodents, several duplications prior to the metazoan appearance and an independent duplication in the *Arabidopsis thaliana* lineage. All these events were manually validated in a previous publication deciphering this family (Lecompte et al., 2008).
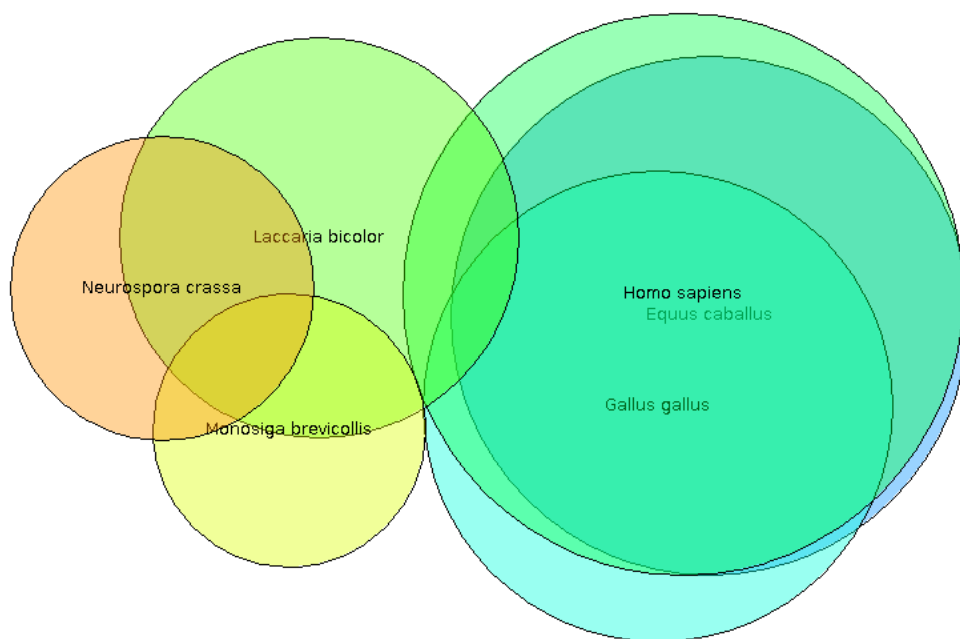
Figure 7-3. An example of a Venn diagram calculated by OrthoInspector. *To produce this diagram, we calculated: all possible intersections between the three fungi (L. Bicolor, N. Crassa and M. brevicollis), all possible intersections between the three vertebrates (H. sapiens, E. caballus and G. gallus) and the intersection between H. sapiens and L. bicolor.*

Finally, a tool is provided for visualization of phylogenetic patterns, namely Euler/Venn diagrams, one of the most common representations used to visualize the genes shared by different species. This tool extends the visualization of phylogenetic patterns to the organism level. Venn diagrams are specialized Euler diagrams in which, for n components, all 2^n hypothetically possible zones corresponding to the various combinations are represented. Euler diagrams do not represent all intersections, but only the possible zones in a given context. OrthoInspector incorporates the VennEuler library (Wilkinson, 2012b) to calculate Euler and Venn diagrams. This library allows to choose which intersections should be represented in the diagram. Consequently, we can represent gene intersections between multiple species corresponding to a particular biological message. Figure

7-3 illustrates this capacity and demonstrates the fact that the fungi *L. bicolor* shares as many genes with humans as it does with other fungi. In contrast, vertebrates share the majority of their genes. This diagram highlights how different the evolutionary distances are between fungi and between mammals. The Venn/Euler diagram calculations are supported by a statistical framework that provides the user with two quality scores estimating the fitness of the data to the proposed representation. Indeed, the fitness may not be complete, since some intersection combinations cannot be completely projected in a 2D plane when comparing a large number of species. A mouse click on any intersection allows to retrieve all orthologs of the corresponding phylogenetic pattern.

7.3 OrthoInspector database

Following the publication of the OrthoInspector software, we created an OrthoInspector database supported by a website. This work was partially performed by a master student, Marc Bigler who contributed especially to the jQuery and Ajax development. The website allows the user to search for orthology relations by textual or BLAST searches. All data and results can be downloaded by users.

7.3.1 Current database content

The first version of the OrthoInspector database (current online version) includes 59 eukaryotic species from the main eukaryotic phyla in Protists, Fungi, Plants and Animals. This represents 940,855 protein sequences. We generated 2,073,328 validated inparalog groups ranging from 1 to 30 proteins. The pairwise comparison of inparalog groups leads to the determination of 8,649,287 1-to-1 relationships, 2,648,403 1-to-many relationships and 469,810 many-to-many relationships. The repartition of these relationships is summarized in figure 7-4.
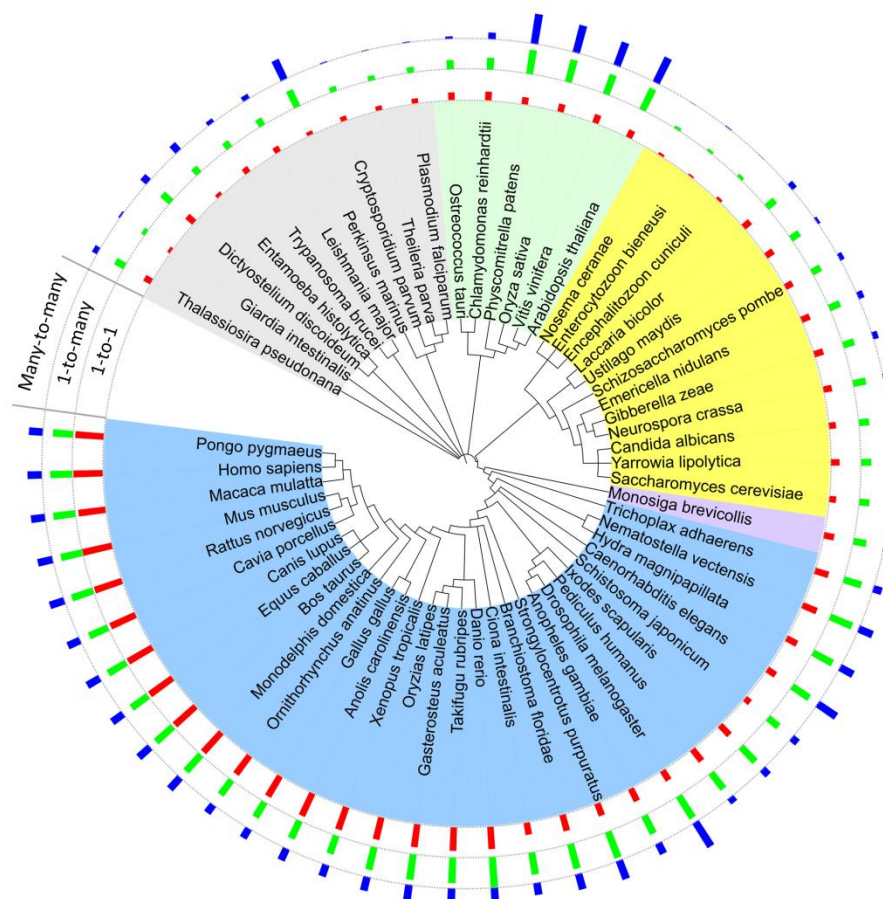
Figure 7-4. The 59 eukaryotic species in the OrthoInspector database. *It contains 6 viridiplantae (green), 9 fungi (yellow), 1 choanoflagellida (purple), 23 metazoa (blue) and 10 species belonging to other eukaryotic groups (grey). The normalized proportion of one-to-one (red), one-to-many (green) and many-to-many (blue) relations is represented by bar-charts. The tree is based on the NCBI taxonomy classification and was generated with the iTol tool (Letunic and Bork, 2007).*

7.3.2 Extending the database to all available eukaryote genomes

A second version of the OrthoInspector database is currently under construction. This update extends each phylum with all the current eukaryotic genomes that have been published, are available in public sequence databases (Uniprot, RefSeq, Ensembl) and have a genome coverage more than 7x (figure 7-5). In particular, we include species that are at the frontier between the main reigns of the tree of life and species that are representative of a recently sequenced phylum. For example, *Capsaspora owczarzaki* is classified in an uncertain phylum between metazoan and fungi. *Batrachochytrium dendrobatidis*, a genome sequenced recently, is the unique representative of the poorly explored *Chytridiomycota* fungi phylum. *Daphnia pulex* and *Ixodes scapularis* are respectively the first crustacean and arachnidan genomes. The three first steps needed for the database update have been completed: species selection, BLAST all-against-all calculation and orthology computation. The remaining steps are the migration of the data to the online database and an update to the website and the software interface. This latter step requires the development of an interactive

species tree interface to facilitate phylum/species selection for a particular analysis. Indeed, dealing with 270 species from very heterogeneous phyla can be difficult for a comprehensive analysis. This work is currently underway, with the help of two master students, Schneider Raphaël and Mörel Can.
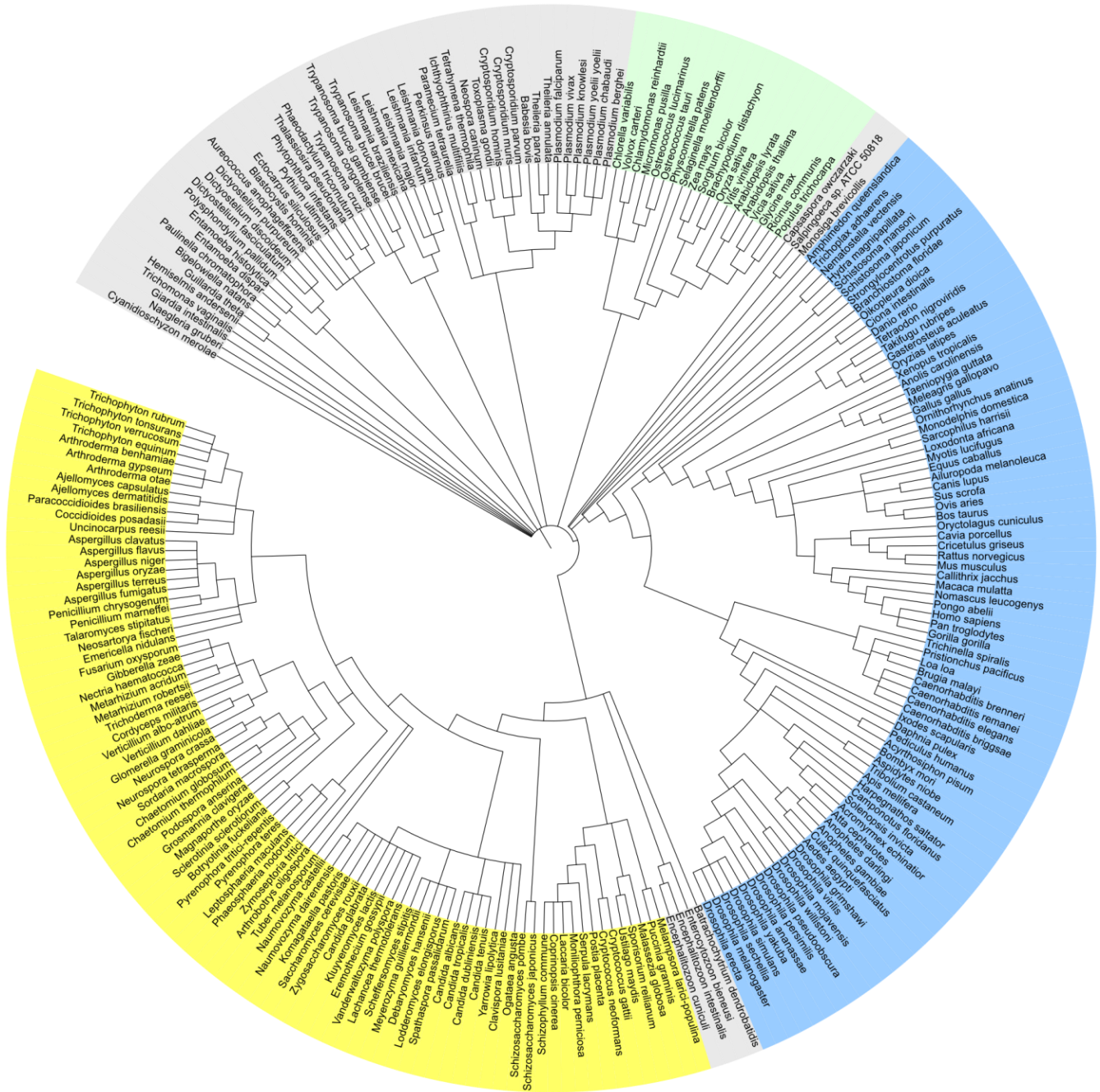
Figure 7-5. The 270 eukaryotic species in the second version of the OrthoInspector database. *It contains 19 viridiplantae (green), 118 fungi (yellow), 79 metazoan (blue) and 51 species belonging to other eukaryotic groups (grey). The tree is based on the NCBI taxonomy classification and was generated with the iTol tool (Letunic and Bork, 2007).*

7.4   OrthoInspector applications

7.4.1   A comparative survey of the TFIIH multiprotein complex

TFIIH is a multiprotein complex composed of two subcomplexes, the CAK and the core-TFIIH. While the CAK, composed of 3 subunits, is mainly involved in the cell cycle and transcriptional regulation, the core, containing 7 subunits, plays a crucial role in both transcription and DNA repair (Zurita and Merino, 2003). An expert phylogenetic distribution of the core-TFIIH in 63 eukaryotic organisms has been performed recently in the laboratory. This analysis revealed that some non-catalytic core-TFIIH subunits were absent in some organisms. To understand the functional significance of these subunits, OrthoInspector was used in a subtractive analysis of the corresponding proteomes. The basic assumption of the subtractive approach is that proteins that function together in a pathway or structural complex tend to co-evolve, i.e. to be present in the same set of species (Pellegrini et al., 1999). Genes sharing a pattern of presence/absence similar to the respective subunits were identified and complemented by a functional analysis. This *in silico* result suggested for the first time a functional link between the p34 subunit of TFIIH and the U1snRNA, but also led to the hypothesis that p34 might be involved either in the earlier first step of mRNA splicing or in the U1 snRNA enhancement of transcription (Alexander et al., 2010). An article describing these results have been submitted (Bedez F, Linard B, Brochet X, Ripp R, Moras D, Lecompte O, Poch O. Functional insights into the core-TFIIH from a comparative survey. Genomics, 2012).

7.4.2   Knowledge extraction for macromolecular complexes

One of the aims of the Puzzle-fit project (funded by the French National Research Agency) is to implement a computational protocol for the integration of diverse data for the structure determination of macromolecular assemblies (figure 7-6). In this context, a pilot study involved the integration of diverse structural information for the elucidation of the molecular architecture of the general transcription factor TFIID. This transcriptional factor has been implicated in several human diseases by virtue of its binding properties with transcriptional activators or repressors or by the histone acetyl transferase (HAT) activity of one of its subunits (Cler et al., 2009). One of the key components of the computational model was the exploitation of the evolutionary context of the TFIID subunits to predict protein-protein interactions at different levels of resolution: protein, domain and residue. Different approaches were used, such as the construction of MACS or the construction of high-quality phylogenetic profiles. OrthoInspector was therefore an important component of the Puzzle-fit pipeline, handling the generation of phylogenetic profiles. The profiles were then used during the knowledge extraction process for the macromolecular complex reconstruction.
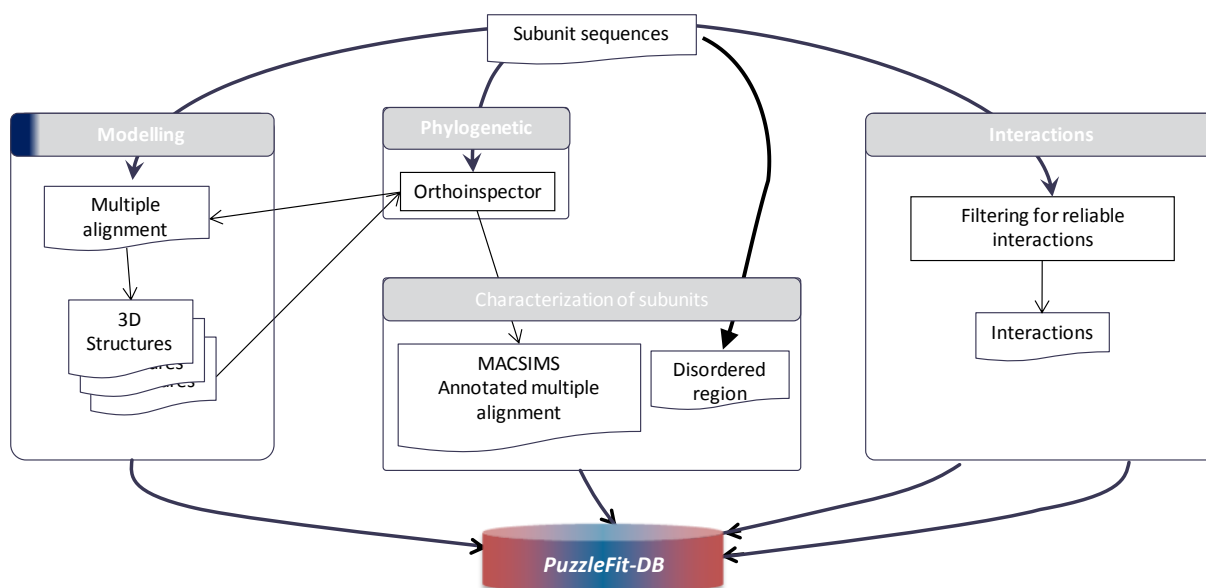
Figure 7-6. Overview of the Puzzle-Fit project pipeline. *An automatic process of knowledge extraction compiles 3D models, phylogenetic and interactome data to predict macromolecular complexes.*

7.4.3   OrthoInspector and Quest for Orthologs Consortium

In 2009, authors of the most perennial orthology databases decided to organize a 'Quest for Orthologs' meeting to discuss and address major limitations and the perspectives for orthology inference (Gabaldon et al., 2009). The LBGI participated in this meeting and joined the consortium (see publication n°5). Subsequently, we implemented several recommendations of this newly formed community. We added the OrthoXML format in OrthoInspector (see sections 3.4.3 & 7.2.3). We also tested OrthoInspector with the 'reference proteome' benchmark provided by Uniprot and made our benchmark predictions publicly available in CSV and OrthoXML format on lbgi.igbmc.fr/orthoinspector/ (section reference proteomes). Calculations for a second and updated version of the benchmark are ongoing.

Following the creation of the first 'reference proteomes' benchmark, Dr. Christophe Dessimoz developed an online benchmarking service allowing the comparison of orthology predictions from several methods with multiple criteria (http://linneus54.inf.ethz.ch:8080/cgi-bin/gateway.pl). We submitted our predictions to the service and obtained the results shown in figures 7-7 and 7-8.
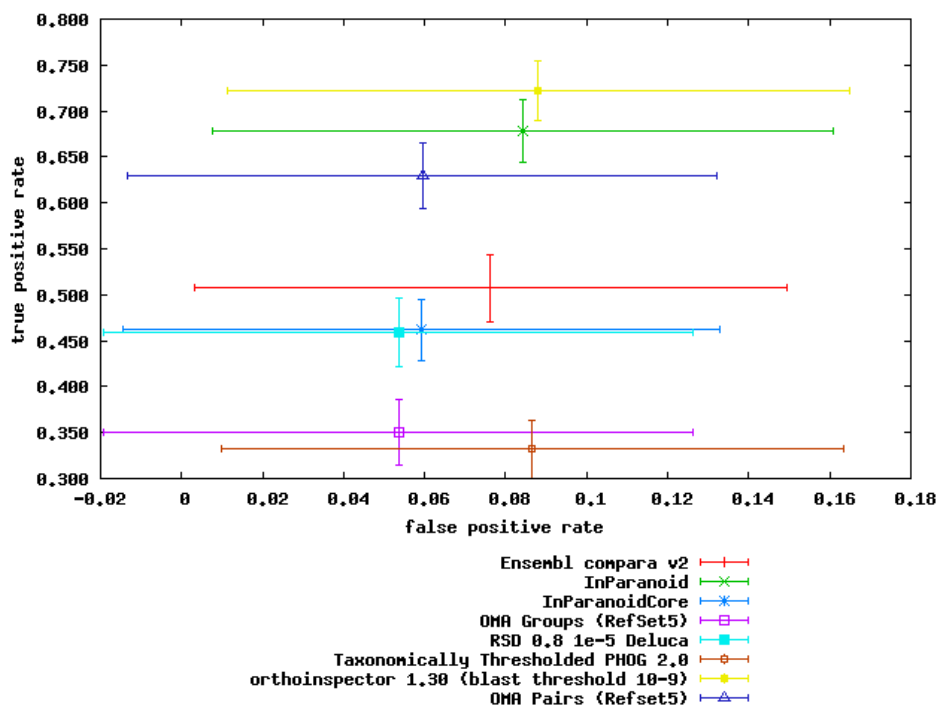
Figure 7-7. Agreement with Reference Phylogeny for 6 protein families. *Orthology predictions from several methods are compared to reference phylogenetic trees. The agreement between phylogeny and orthologous pairs from 8 methods is resumed by true positive (sensitivity) and false positive (specificity) rates.*
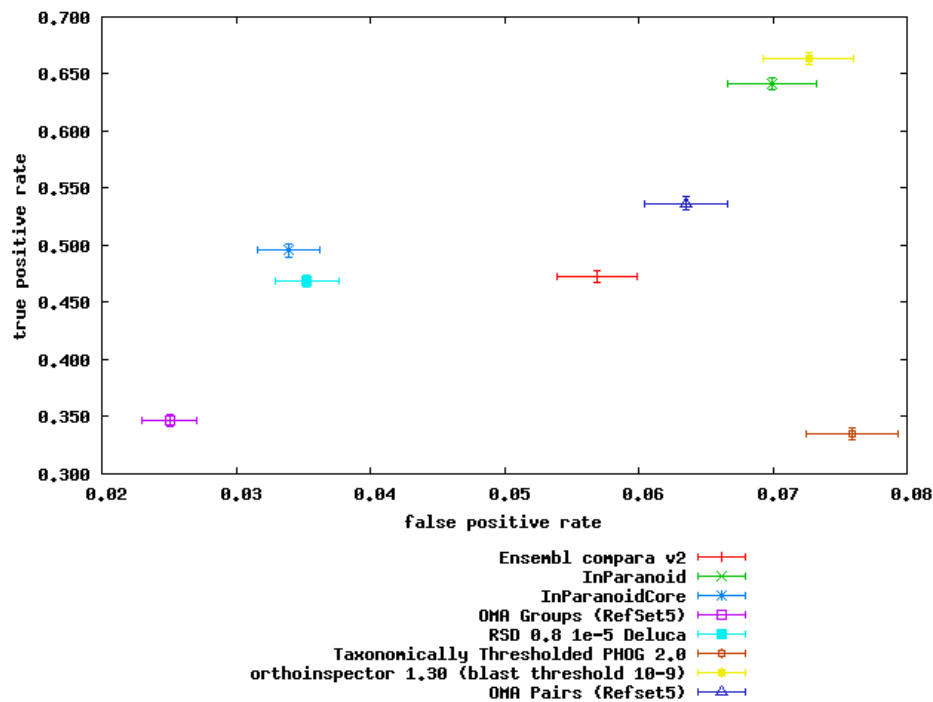
Figure 7-8. Agreement with semi-automated Reference Phylogeny (TreeFam A). *Orthology predictions from several methods are compared to Treefam A trees. The agreement between phylogeny and orthologous pairs coming from 8 methods is resumed by true positive (sensitivity) and false positive (specificity) rates.*

In the dataset corresponding to reference phylogenies (figure 7-7), OrthoInspector has a similar false positive rate interval compared to other methods (0%>rate>17%). However, OrthoInspector performs better than all other methods in terms of the true positive rate (73%), followed by Inparanoid (68%) and OMA pairs (63%). Concerning the agreement with the semi-automated phylogeny dataset (figure 7-8), OrthoInspector predictions correspond to a reasonable false positive rate (<8%). Other methods show similar false positive rates between 2% and 8%. In this second dataset, the gain of sensitivity provided by the OrthoInspector approach is again demonstrated, since our method again achieves the highest true positive rate. Taking into account these results and our published test (publication n°1), we conclude that OrthoInspector has successfully improved orthology inference sensitivity, at the same time retaining a simple and fast graph-based algorithm.

7.5   Conclusions

In the current context of fast and inexpensive genome sequencing, we need powerful algorithms to infer orthology relations on a high-throughput scale and new tools to analyze and visual such relations at the genome-scale. We designed OrthoInspector for this purpose. OrthoInspector predictions perform well compared to other orthology inference methods, finding new co-ortholog relations particularly for large protein families and large evolutionary scales. In contrast to many software systems, OrthoInspector does not only provide flat files with orthology relations. It is a complete software suite and includes several tools to extract, summarize and visualize orthology data from the gene family scale to the multi-genome scale. Such tools are urgently needed in a period when extracting and visualizing data can be extremely time consuming. Reducing this time will facilitate the discovery of more biological knowledge at larger evolutionary scales.

7.6   Publication 1. OrthoInspector: comprehensive orthology analysis and visual exploration.

Publication n° 1

BMC Bioinformatics

OrthoInspector: comprehensive orthology analysis and visual exploration

Benjamin Linard[*], Julie D Thompson, Olivier Poch, Odile Lecompte

Abstract

Background: The accurate determination of orthology and inparalogy relationships is essential for comparative sequence analysis, functional gene annotation and evolutionary studies. Various methods have been developed based on either simple blast all-versus-all pairwise comparisons and/or time-consuming phylogenetic tree analyses.

Results: We have developed OrthoInspector, a new software system incorporating an original algorithm for the rapid detection of orthology and inparalogy relations between different species. In comparisons with existing methods, OrthoInspector improves detection sensitivity, with a minimal loss of specificity. In addition, several visualization tools have been developed to facilitate in-depth studies based on these predictions. The software has been used to study the orthology/in-paralogy relationships for a large set of 940,855 protein sequences from 59 different eukaryotic species.

Conclusion: OrthoInspector is a new software system for orthology/paralogy analysis. It is made available as an independent software suite that can be downloaded and installed for local use. Command line querying facilitates the integration of the software in high throughput processing pipelines and a graphical interface provides easy, intuitive access to results for the non-expert.

Background

New sequencing technologies are dramatically increasing the number of predicted protein sequences available for high throughput comparative analyses, functional annotation or evolutionary studies. All these studies involve a transfer of information between organisms and homology is one of the most popular concepts used to address this problem. In particular, the studies rely on an accurate determination of orthology and paralogy relationships. According to the seminal definition of Fitch [1], orthologs are homologous genes that diverged from a single ancestral gene in their most recent common ancestor via a speciation event, whereas paralogs are homologs resulting from gene duplications. The distinction between orthologs and paralogs refers exclusively to the evolutionary history of genes and does not have functional implications *stricto sensu* [2]. However, from an operational point of view, it is widely accepted that

two orthologs generally share the same function [3]. In contrast, paralogs are generally considered more divergent as new functions can emerge as the result of mutations or domain recombinations. Nevertheless, the multiplication of available genomes has underlined the necessity to distinguish two subtypes of paralogs: inparalogs and outparalogs [4]. Inparalogs are produced by duplication(s) subsequent to a given speciation event, while outparalogs result from an ancestral duplication (relative to the given speciation event). In other words, in-paralogy and out-paralogy are concepts relative to the species under comparison. The distinction is crucial in evolutionary studies since sets of inparalogs derive from orthologs by lineage-specific expansions and thus can be considered to be co-orthologs, while outparalogs do not have orthologous relationships at all.

Today, the most commonly used approach for the prediction of homology relationships between genes and proteins (and thus orthology and paralogy relationships) involves some kind of similarity measure, which can be linked to different types of data, such as sequences, domains or even 3 D structures. In principle, phylogenetic

* Correspondence: linard@igbmc.fr
Laboratoire de bioinformatique et genomique integratives, Département de Biologie et Génomique Structurales CNRS/INSERM/UDS, Institut de Génétique et de Biologie Moléculaire et Cellulaire, 1 rue Laurent Fries, 67404, Illkirch, Cedex, France

tree-based inference represents the most accurate way to determine orthology and paralogy [3-5]. However, its use at the complete proteome scale is computationally expensive and, given the rate at which new genomes are now being sequenced, cannot be considered as a viable option for most laboratories at the present time. As a consequence, alternative algorithms based on graphs or on a combination of tree and graph representations [6], have been developed to infer homology relationships. Most of them involve protein Blast all-versus-all searches and use pairwise distance calculations [7], 3-way best-hits [8-10] or clustering-based approaches [11-13]. In general, comparative studies [14,15] have shown that phylogenetic reconstructions have higher sensitivity and lower specificity than graph-based methods, particularly for distant organisms. Nevertheless, these methods provide good results for both sensitivity and specificity with some datasets [16,17]. However, each of the methods has advantages and disadvantages, and the most appropriate method will depend on the user's purpose [6,18]. Apart from the detection accuracy, other factors need to be taken into account, for example the availability and ease-of-use of the programs. Most of the methods commonly used today are made available as public software binaries and data browsing for the non-specialist is limited to web interfaces that allow remote querying of pre-calculated databases. For the more computer literate, large-scale queries can be performed and results can be retrieved in the form of flat files, although this requires a certain level of programming expertise to parse the data. To address this problem, some efforts have been made to facilitate the querying of data through presence/absence constraints and to provide global views of results via phylum-related tables [10]. Nevertheless, the tools are still available as web-based interfaces and cannot be retrieved locally to support or maintain in-house databases.

Here we describe OrthoInspector, a new software system incorporating an original algorithm for the rapid detection of orthology and in-paralogy relationships between different species. In comparisons with existing methods, it improves detection sensitivity, with a minimal loss of specificity. Moreover, OrthoInspector has a modular design and is provided as an independent software suite that can be downloaded and installed for local use. Command line querying facilities have been developed to allow fast information selection for high throughput studies and to facilitate the integration of the software in other packages or processing pipelines. An enhanced graphical interface is designed to automate the complete software installation and data generation process for non-specialists. Finally, different visualization tools have been designed specifically to allow the in-depth exploration of the complex inter-species orthology/in-paralogy relationships detected.

Implementation

The OrthoInspector suite is coded in Java 1.6.x, which means that it can be run on all Java-supporting platforms (UNIX, Windows, Mac....). Several java packages are incorporated: (i) the Jacksum package is used to encoded sequence data, (ii) the JDOM and opencsv packages are used to format sequence and orthology/paralogy data, (iii) the Jung and Prefuse packages are used to support the visualization tools. OrthoInspector also requires a background database to handle the huge amount of data produced by a Blast all-versus-all analysis. Support for the main "relational database" compatible engines (MySQL, PostgresSQL, Oracle...) is provided via definition of the corresponding java drivers in a configuration file. The only constraint is the predefined database schema that is needed by OrthoInspector. The software suite provides two different user interfaces, a command-line client and a graphical interface that can be used to perform the three steps involved in the complete analysis process (Figure 1):
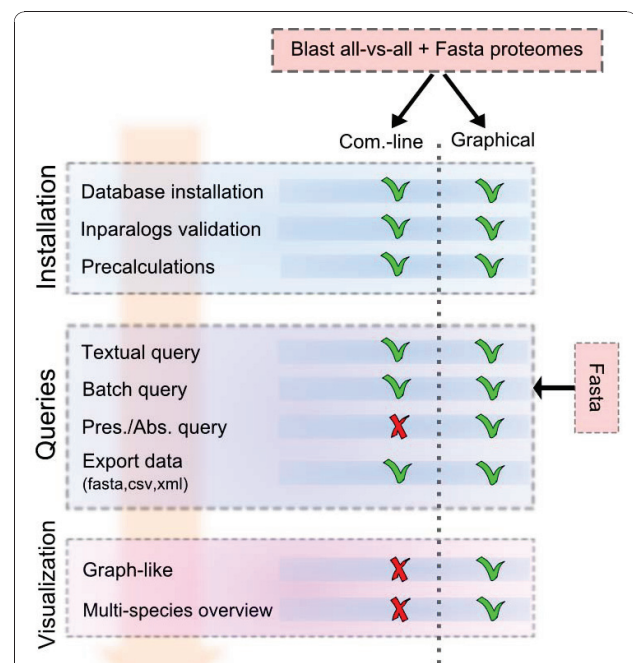
Figure 1 OrthoInspector Suite overview. OrthoInspector provides two user interfaces: a command-line client and a graphical interface. Installation operations include the creation of the database, the calculation of ortholog/inparalog groups and an optional creation of pre-calculated data. Queries include orthologous relationship searches, with or without advanced criteria: textual searches access results trough sequence accession numbers or sequence descriptions, batch queries allow submitting of multiple sequences in FASTA format and constraints of presence/absence of orthologs in specific organisms can be considered. Visualization tools provide different views for comparative studies.

1. Command-line and graphical versions can be used to perform Blast all-versus-all sequence searches and to generate a database containing the search results. Currently, the package is designed to allow the use of both raw and tabbed outputs produced by the classical NCBI Blast package and the recent NCBI blast+ package [19]. Other Blast data formats can be easily added with the help of the Blast parser interface included in the package. OrthoInspector proposes options (i) to directly fill the database with the produced data or (ii) to create intermediate data dumps allowing a considerable speed-up. Sql scripts to use these dumps in mySQL and postgresSQL engines are provided in the OrthoInspector website.

2. After database installation, the command-line version allows fast information retrieval for high throughput studies and the use of the software in other packages. Textual queries (accession numbers, description...), batch queries (Fasta sequences in a file) or queries defining presence/absence of an ortholog in specific organisms can be performed. Both command-line and graphical versions allow the user to export data in FASTA, CSV and XML formats. New output formats can be easily coded with the help of the output interface provided.

3. The graphical version facilitates data querying for non-specialists. In addition, it provides a set of useful tools to retrieve clusters of orthologs covering multiple species, to produce comparative genomics results and to visualize the data.

The whole software suite is available at http://lbgi.igbmc.fr/orthoinspector. Furthermore this website contains tutorials and database dumps for test purposes.

Methods
OrthoInspector algorithm
The OrthoInspector algorithm is divided into three main steps. First, the results of a Blast all-versus-all (proteomes are blasted against each other) is provided by the user and is parsed to find all the Blast best hits for each protein and to create the groups of inparalogs. Second, the inparalog groups for each organism are compared in a pairwise fashion to define potential orthologs and/or in-paralogs. Third, best hits that contradict the potential orthology between entities are detected.

Inparalog group formation and validation
The first step involves the parsing of the Blast all-versus-all results to find all best hits for each protein and to create the groups of inparalogs, i.e. paralogs produced by duplications subsequent to a given speciation event (Figure 2). Inparalog groups are organism-dependant,
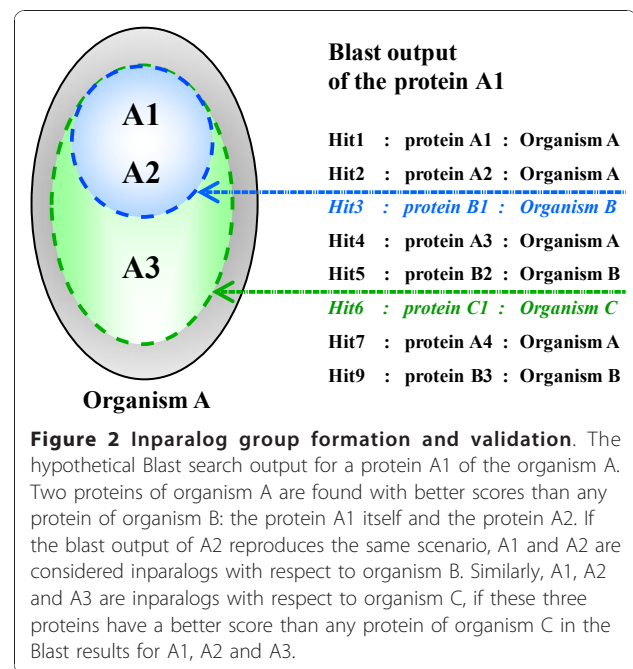
Figure 2 Inparalog group formation and validation. The hypothetical Blast search output for a protein A1 of the organism A. Two proteins of organism A are found with better scores than any protein of organism B: the protein A1 itself and the protein A2. If the blast output of A2 reproduces the same scenario, A1 and A2 are considered inparalogs with respect to organism B. Similarly, A1, A2 and A3 are inparalogs with respect to organism C, if these three proteins have a better score than any protein of organism C in the Blast results for A1, A2 and A3.

which means that a given protein (p_n) can be in different putative groups of inparalogs and we will denote these groups as organism-dependant lists: $\{p1, p2, ..., pn\}^{/organism}$. Given a Blast search result for a protein of organism A, all proteins of A with an E-value inferior to the E-value of the best hit in the organism B will define a potential group of inparalogs in A with respect to the internal node where species A and B coalesce (we will refer to a group of inparalogs in A "with respect to B"). The putative list of inparalogs is then validated if the same minimal hypothesis of inparalogy is verified in the Blast searches for each protein in the list. As an example, we can consider a group of three putative inparalogs in organism A with respect to B (denoted $\{A1, A2, A3\}^{/B}$) that has been defined by the Blast output of the protein *A1*. The entire group will be validated if the Blast outputs of *A2* and *A3* result in the same group. Thus, validation requires that the groups $\{A2, A1, A3\}^{/B}$ or $\{A2, A3, A1\}^{/B}$ are defined by the Blast output of *A2* and that the groups $\{A3, A1, A2\}^{/B}$ or $\{A3, A2, A1\}^{/B}$ are defined by the Blast output of *A3*. If the above condition is not verified, the existence of two-member groups is checked. In the example, if the Blast output of *A1* defines the group $\{A1, A2, A3\}^{/B}$ but the Blast output of *A3* defines a group of two proteins $\{A3, A1\}^{/B}$, only this *A2*-deleted paralog group will be retained in the subsequent steps of the algorithm. Using this method, if n_{orga} organisms are used to create the Blast all-versus-all, each Blast search can define $n_{group} <= n_{orga}$ putative groups of inparalogs, each one being delimited by a best hit in another organism.

Pairwise comparison of inparalog groups

The second step of the OrthoInspector algorithm is the definition of potential (co)-orthology relationships (Figure 3). The definition is based on the detection of best hits existing between the two types of entities determined at the previous step: single proteins (not included in a group of inparalogs), and proteins belonging to one or several inparalog groups. We thus have three types of pairwise entity comparisons (*{protein <-> protein}*, *{protein <-> inparalogs}* and *{inparalogs <-> inparalogs}*), corresponding to the three types of relationships shown in Figure 3. A *1-to1* relationship is described by a best hit between a protein of *O1* and a protein of *O2* complemented by a returning best hit from the protein of *O2* to the protein of *O1*, *known as a reciprocal best hit*. A *1-to-many* relationship is described by a best hit from a given protein of *O1* to any protein member of an inparalog group of *O2* complemented by a returning best hit from any member of the inparalog group of *O2* to the same protein of *O1*. Finally, a *many-to-many* relationship is described by two best hits between proteins of two groups of inparalogs (a group in *O1* and a group in *O2*).

Detection of contradicting information

The third step in the algorithm is the detection of best hits that contradict the potential orthology relationships

defined above. In particular, given two inparalog groups that are potentially orthologous, it is possible to find a best hit from a protein in one of the compared groups to another protein that does not belong to either of the groups. In this case, it is possible that the protein does not belong to the inparalog group. Such contradictions are highlighted by OrthoInspector with a warning signal in the algorithm output: a "red signal" indicates contradictions involving a reciprocal best hit and an "orange signal" indicates contradictions involving a simple best hit. Such signals help the user to discriminate proteins in complex inparalog groups formed by closely related sequences or in cases where the proteome of one of the compared organisms is incomplete and disturbs the precedent formation of validated inparalog groups.

Results

Large-scale proteome analysis

We used the OrthoInspector software to study 59 organisms with approximately complete proteomes covering the main eukaryotic phyla in Protists, Fungi, Plants and Animals. We l incomplete and low coverage genomes to avoid predictions of false gene loss and artefacts in gene duplication inference [20]. The complete list of the 59 studied organisms with their taxonomic identifiers and the number of retained protein transcripts can be found in additional file 1. For 22 higher eukaryotes, protein sequence datasets from Ensembl 56 [21] were used. To avoid multiple transcript issues, the longest protein sequence was selected for each Ensembl-predicted gene annotated as 'protein-coding'. For example, the proteomes of *Homo sapiens* (22384 transcripts), *Mus musculus* (23117 transcripts), *Xenopus tropicalis* (18023 transcripts), *Ciona intestinalis* (14180 transcripts), *Arabidopsis thaliana* (31280 transcripts) or *Oryza sativa japonica* (57995 transcripts) were obtained from Ensembl. For eukaryotes not stored in Ensembl, the NCBI RefseqP [22] and Uniprot (Swissprot +TrEMBL) [23] databases were used. Data from both sources were retrieved using ICARUS scripts on a local SRS server [24] to select sequences according to their taxonomic identifiers. To remove redundant sequences, each sequence was compared to all others from the same organism using Blast. For sequences sharing more than 99% identity, manually-annotated entries from Swissprot were preferred over TrEMBL and RefseqP entries, otherwise the longest sequence was retained. Proteomes built with this protocol include *Plasmodium falciparum* (5234 transcripts), *Trypanosoma brucei* (8928 transcripts), *Ostreococcus tauri* (7974 transcripts), *Encephalitozoon cuniculi* (1903 transcripts), *Emericella nidulans* (9732 transcripts), *Saccharomyces cerevisiae* (6771 transcripts), *Laccaria bicolor* (17698 transcripts), *Caenorhabditis elegans* (22614 transcripts), *Ixodes scapularis*

Figure 3 Comparison of inparalog groups. Blast best hits are used to define the potential relationships existing between inparalog groups. 1-to-1 relationships are equivalent to the classical reciprocal best hits (RBH). 1-to-many relationships are associated with potential duplication(s) after speciation in one of the lineages and cannot always be detected by RBH. Many-to-many relationships result from duplications in both lineages after speciation and again, cannot always be detected by RBH.

(21009 transcripts) and *Drosophila melanogaster* (22430 transcripts). Regardless of the source sequence database, sequences with less than 20 amino acids or more than 10000 amino acids were excluded. Finally, we obtained a pool of 940855 protein sequences.

The new NCBI-Blast+ package was then used to perform Blast all-versus-all searches between the proteomes of the 59 organisms, representing 940855 individual Blast searches in a database of 940855 sequences. Sequences were selected with an E-value cutoff of 1e-9. The searches were executed on the Décrypthon grid resources [25].

The results of the Blast all-versus-all searches, together with the 59 proteomes were then used as input to OrthoInspector. All steps of the algorithm, from Blast parsing to integration of the data in the relational database, took about 20 hours on four 2.67 GHz Intel Xeon CPUs with 6 Go of RAM. This timing is based on an installation of the database using the faster "database dumps" configuration (see Implementation). In more detail, parsing of the Blast results took 5h20, validation of inparalog groups took 2h10 and generation of 1-to-1, 1-to-many and many-many precalculated data for the 59 organisms took 12 h. The inparalog prediction step produced 10342157 putative inparalog groups, themselves generating 2073328 validated groups (Figure 4). Shortest versions of this huge dataset (> 100Go), including 7 proteomes, are available as database dumps (mySQL and postgresSQL) at the OrthoInspector website http://lbgi.igbmc.fr/orthoinspector.

As expected, large-scale proteomes, e.g. in plants (green color in Figure 4), or genome-wide duplications, e.g. in fishes (medium blue), result in an increase in the number of predicted inparalog groups, whereas smaller eukaryote proteomes have relatively few groups. The number of inparalog groups is generally correlated with the proteome size and the phylogenetic distance between organisms, for instance, amniota (dark blue) have a relatively stable number of inparalog groups. Nevertheless, some exceptions can be observed. Despite having the largest proteome in the plant phylum, Oryza sativa has fewer groups than Vitis vinifera or Arabidopsis thaliana and sequences from Oryza sativa are included in relatively few inparalog groups. Further investigation showed that many sequences of this organism had a relatively small number of Blast hits to other organisms compared to other plants (data not shown). This may be partly due to some overprediction of genes in the Oryza sativa proteome, with several protein fragments or pseudogenes predicted as "protein-coding".

Another interesting observation is that all parasitic organisms generate a small number of inparalog groups compared to the other members of their phylum. In arthropods, Ixodes scapularis (deer tick) and Pediculus humanus (body louse) have less inparalog groups than

Anopheles gambiae and Drosophilia melanogaster. In fungi, Encephalitozoon cuniculi, Nosema ceranae and Ustilago maydis have less inparalog groups than other members of this phylum. Lacarria bicolor is another fungus with few inparalogs, although this may be linked to its ectomycorrhizal symbiotic relationship with plant roots.

Unlike parasites or symbionts, some isolated organisms have a relatively large number of inparalog groups. For example, Strogylocentrotus purpuratus has numerous inparalog groups but is currently the only echinodermata genome available, and it is impossible to determine whether this is a characteristic of this phylum. Entamoeba histolytica has a number of inparalog groups similar to that to other organisms with the same proteome size, but individual sequences are included in more inparalog groups compared to other organisms. This might be explained by the lower quality of the proteome and/or the presence of numerous repeats, resulting in multiple Blast hits in all studied species.

In order to identify potential orthology relationships, all the inparalog groups were compared for each pair of organisms. The total number of relationships detected represents 8,649,287 1-to-1 relations, 2,648,403 1-to-many relations and 469,810 many-to-many relations. Figure 5 and additional files 2 and 3 show respectively the number of 1-to-many, 1-to-1 and many-to-many between each proteome pair after normalization. The number of predicted relationships is largely dependent on the composition of the set of selected organisms. As expected, close species present a high proportion of 1-to-1 relationships within their group but few many-to-many relationships (additional files 2 and 3). This is especially obvious for the 18 vertebrates included in our dataset that are phylogenetically very close to each other compared to the other studied phyla. Intergroup relationships highlight lineage-specific duplications. For instance, the 2 whole genome duplications (WGD) encountered by the jawed vertebrates [26] are clearly reflected by the high number of 1-to-many relationships from invertebrates to vertebrates (Figure 5). Similarly, 1-to-many relationships pinpoint the additional round of duplication encountered by the teleostei lineage within vertebrates [27]. The numerous duplication events reported in the land plants [28] explain the extent of 1-to-many relationships between them and most of other species used in our study. Additionally, the abundance of many-to-many relationships between Physcomitrella patens (moss) [29]and flowering plants is in agreement with the independent events that occurred in the moss lineage (simple duplication) and hexaploidy event in flowering plants. Examination of specific sets of relationships (data not shown) is in agreement with dedicated studies. For instance, the functional analysis of the

Figure 4 Distribution of predicted inparalog groups over 59 organisms. Organisms are ordered by their phylum and their decreasing number of inparalogs (green: viridiplantae, dark blue: tetrapoda, blue: teleostei, clear blue: other bilateria, pink: fungi, purple: other eukaryotic groups). The mean number of inparalog groups per sequence (red dots) is linked to the size of the proteomes (negative y axis). The black dotted line represents the corresponding polynomial tendency curve.

human genes exhibiting one-to-many relationships with rodents reveal a significant enrichment in gene related to olfaction as previously reported [30].

Example test case: myotubularin family

To demonstrate the advantages of using inparalog group comparisons to predict orthology, we studied the myotubularin family as a test case. The distribution of myotubularin-related proteins is well established [31] and is represented in Figure 6 for three species with multiple duplication events that occurred during its evolutionary history. OrthoInspector predictions are compared to Inparanoid and OrthoMCL, illustrating the algorithmic differences that lead to some false negatives for the two latter algorithms.

Inparanoid is based on RBH and finds inparalogs having a similarity score equal to or superior to the similarity S defined by the RBH. In the fly/yeast comparison, the three fly myotubularins are more similar to each other than to the yeast myotubularin, thus they are considered as inparalogs. In the human/yeast comparison case, 6 out of 8 human myotubularins have a higher

similarity score than the similarity score defined by the yeast/human RBH, but 2 proteins have lower scores and are thus not considered as inparalogs (false negatives).

The OrthoMCL algorithm begins with the same steps of RBH detection and identification of sequences within the same genome that are more similar to each other than to any sequence from another genome. Then, a graph is constructed, where nodes represent proteins and edges represent the relations, and a Markov clustering is performed. In this example, three clusters are found, with only one fly and three human myotubularins considered to be co-orthologs of the yeast myotubularin.

OrthoInspector does not consider RBHs as a preliminary condition to detect potential inparalogs, instead inparalog groups are inferred directly in each organism. For example, the three fly and eight human myotubularins are identified as inparalogs with respect to yeast. In a second stage, the pairwise comparison of inparalog groups exploits the RBH and BH found between the different organisms to infer many-to-one relations including all the myotubularins.

Figure 5 Distribution of 1-to-many relations over 59 organisms. The normalized number of 1-to-many relations is calculated for each organism pair. Normalisation is done by dividing the observed number of relations by the maximum number of potential relations (the size of the largest proteome of the two compared organisms). The 1-to-many relation is oriented from the x axis to the y axis.

Comparison with existing methods: benchmark data sets

The accuracy of the OrthoInspector predictions was compared to five existing methods, covering the main approaches to infer orthology: namely, Inparanoid (pairwise distance comparisons), eggNOG (3-way best hits), OrthoMCL and OMA (graph clustering) and Ensembl compara (phylogenetic tree inference). Today, these methods are widely used by the community and their databases are cross-referenced in public databases like Uniprot. OrthoInspector is based on a pairwise distance based algorithm which makes it similar to the Inparanoid algorithm in some aspects. However, Inparanoid is directly based on reciprocal best hits (RBH) to find orthologs and inparalogs, as illustrated by the example test case described above. The first step of our algorithm identifies potential inparalog groups independently of RBH, thus exploring a larger search space for the discovery of potential orthology relations. The second step of our algorithm then compares inparalog groups that

are not necessarily linked by a RBH between two organisms.

In order to compare the predictions made by OrthoInspector with the existing methods in a large scale study, we used two benchmarks from the literature [32,33], representing varied protein families (nuclear receptors, hox families, membrane receptors...). The literature benchmarks cover many organisms, including *H. sapiens, M. musculus, G. gallus, D. rerio, D. melanogaster, C. elegans* and *S. cerevisiae*. In addition, we created our own benchmark, performing a detailed study of protein kinase families with complex evolutionary histories that represent a significant challenge for the accurate detection of orthology/paralog relationships. Protein kinases represent an ideal test case for our purposes, since they have been intensively studied and their family relationships are generally known. In fact, protein kinases have been classified into a number of groups sharing broad functional properties, based on sequence
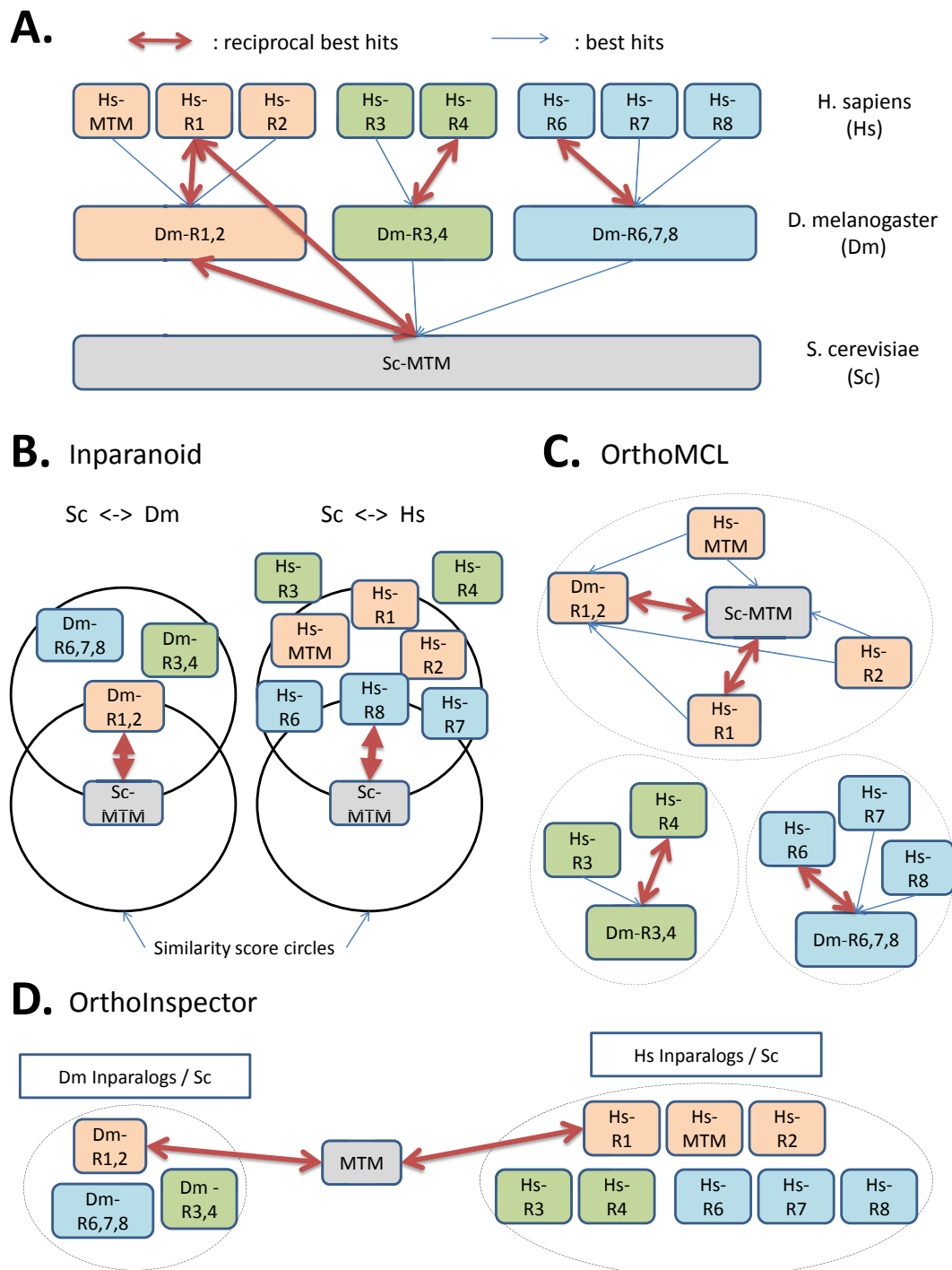
Figure 6 Myotubularin family predictions. A. The myotubularin family distribution is established in three species: *H. sapiens* (Hs), *D. melanogaster* (Dm) and *S. cerevisiae* (Sc), and multiple duplication events have been identified. Reciprocal best hits (RBH) and best hits (BH) linking the sequences are represented as red and blue arrows. Sequences used are Hs-MTM (MTM_HUMAN), Hs-R1 (MTMR1_HUMAN), Hs-R2 (MTMR2_HUMAN), Hs-R3 (MTMR3_HUMAN), Hs-R4 (MTMR4_HUMAN), Hs-R6 (MTMR6_HUMAN), Hs-R7 (MTMR7_HUMAN), Hs-R8 (MTMR8_HUMAN), Dm-R1,2 (Q9VMI9), Dm-R3,4 (Q7YU03), Dm-R6,7,8 (Q8MLR7), Sc-MTM (P47147). B. Inparanoid first identifies RBHs, then searches for putative inparalogs around these RBHs in a search space delimited by similarity score circles. The prediction between Hs and Sc excludes Hs-R3 and Hs-R4 sequences. C. OrthoMCL constructs a graph with proteins as nodes and similarities as edges, then identifies sub-graphs corresponding to co-ortholog groups. In this example, 5 human myotubularins and 2 fly myotubularins are excluded from the group containing the yeast myotubularin. D. OrthoInspector defines inparalog groups, then compares groups between organisms based on RBH and BH relations. Here, inparalog groups containing all the members of the myotubularin family are identified in human and fly, with no false negatives.

similarity in their catalytic domains, the presence of accessory domains and known modes of regulation. Using the standard classification, available at http://kinase.com/kinbase, and by studying the literature, we defined a test set of well annotated protein kinase sequences, from the CMGC group (including cyclin-dependent kinases, mitogen-activated protein kinases, glycogen synthase kinases and CDK-like kinases) and from the TKL (tyrosine kinase-like) group. CMGC kinases represent a homogeneous group, where most proteins possess only the kinase catalytic domain. In contrast, the TKL kinases are more divergent, often having additional domains that regulate kinase activity, link to other signaling modules, or localize the protein in the cell. The CMGC and TKL groups can be further subdivided into several protein families. The distribution of these families was established by a combination of published *in silico* and wet-lab studies in a number of model organisms, including *D. discoideum* [34], *C. elegans* [35], *S. cerevisiae* [36], *D. melanogaster* [37], *M. musculus* [38] and *H. sapiens* [39]. Our test set consisted of 329 manually annotated sequences from these six organisms, covering 31 CMGC sub-families and 16 TKL sub-families (additional file 4).

We then evaluated the predictions made by each of the six methods to the known classifications defined in the four benchmarks. The prediction accuracy was estimated by calculating the Positive Prediction Value (PPV) as a specificity indicator and the sensitivity (Sn) of each method (Figure 7). The benchmark data sets allowed us to highlight a number of advantages and disadvantages of the different methods. For example, OMA achieved the highest specificity, but the lowest sensitivity on average. In contrast, eggNOG obtained the highest sensitivity, although it should be noted that some co-ortholog groups in eggNOG are manually curated, like the COG database on which it is based. On average, the six methods can be classified in two groups. OrthoMCL, Ensembl compara and eggNOG have higher sensitivity than specificity, while Inparanoid, OrthoInspector and OMA have higher specificity than sensitivity. In the second class, OrthoInspector demonstrated higher sensitivity than the other two methods. In fact, OrthoInspector reached a sensitivity level close to that of Ensembl compara (80% and 81% respectively) and superior to OrthoMCL (78%).

Taken individually, the four benchmarks highlighted some contrasting results. For example, OMA obtained a sensitivity <50% for both TKL and CMGC benchmarks,

	Benchmark	PPV	Sn	mean PPV	mean Sn
OMA	A	0,93	0,48		
	B	0,89	0,42		
	C	0,87	0,69	0,91	0,56
	D	0,95	0,65		
Inparanoid	A	0,77	0,83		
	B	0,93	0,61		
	C	0,87	0,81	0,89	0,76
	D	0,97	0,80		
OrthoMCL	A	0,78	0,78		
	B	0,86	0,71		
	C	0,82	0,73	0,71	0,78
	D	0,37	0,91		
OrthoInspector	A	0,71	0,86		
	B	0,95	0,69		
	C	0,77	0,78	0,84	0,80
	D	0,92	0,86		
Ensembl Compara	A	0,90	0,87		
	B	0,74	0,77		
	C	0,67	0,80	0,79	0,81
	D	0,85	0,80		
eggNOG	A	0,83	0,96		
	B	0,69	0,79		
	C	0,78	0,85	0,66	0,88
	D	0,33	0,92		

A : TKL kinases benchmark
B : CMGC kinases benchmark
C : Literature benchmark n°1 (Tim Hulsen & al, 2006)
D : Literature benchmark n°2 (Austin L. Hughes,1998)

Figure 7 Sensitivity and specificity comparison based on 4 benchmarks. Two literature benchmarks and human CMGC and TKL kinases were used to evaluate the prediction accuracy for OrthoInspector and five other methods. Sensitivity (Sn) and Positive Predictive Values (PPV) were calculated for each method on each benchmark. The radar plot resumes the mean PPV (pink) and sensitivity (blue) for each method.

compared to >60% for all other methods. This was due to the fact that OMA failed to predict some orthology relations existing between distant organisms (e.g. human and *C. elegans, S. cerevisiae* or *D. discoideum*). Ensembl compara had higher sensitivity than OrthoInspector for both kinase benchmarks (TKL:+1%, CMGC:+6%) and OrthoMCL had higher sensitivity for the CMGC kinases (+2%), but not TKL kinases (-8%). In the case of the literature benchmark n°1, all methods achieved a good sensitivity and a good specificity, which was not unexpected since the benchmark contains essentially human/ mouse and human/worm relations. For the literature benchmark n°2, the results were more variable. OrthoMCL and eggNOG had high sensitivity (> 90%), but their specificity was surprisingly low (< 40%). In this benchmark, some protein families (heat shock proteins, collagens...) are totally included in a few or a single cluster. This observation is particularly true in the case of distant organism comparisons (human versus *C. elegans, S. cerevisiae*...).

It is clear from these results that the different methods tested here provide complementary approaches for orthology inference. In the future, it should be possible to combine the advantages of the alternative methods to improve both sensitivity and specificity. For example, OrthoInspector could be used as a starting tool to infer orthology relations, since its sensitivity and specificity are well balanced compared to most of the other methods tested here. Furthermore, the orthology inference is less computationally intensive than Ensembl compara, the only other method that achieved similar results. In a subsequent refinement step, the user could then integrate information about true/false positives from lower specificity methods such as eggNOG, OrthoMCL or Ensembl compara and lower sensitivity methods like Inparanoid or OMA methods.

Data management and visualization

The main goal of the OrthoInspector project was to build a complete software suite for orthology and inparalogy prediction and analysis. Nevertheless, in the face of the huge amounts of data being produced by the new sequence technologies, it was clearly crucial to incorporate efficient data management and update procedures in the design of the software. Thus, the complete construction of a database of orthologs can be managed via a four step user-friendly process. OrthoInspector provides administrator tools, accessible via a command-line or a graphical interface, that take as input: (i) the results of a Blast all-versus-all search in a specified directory, (ii) the fasta proteomes of the organisms used in the Blast searches together, with an XML format file describing the organisms (name, source, taxonomic

identifier...). The administrator can then launch the installation procedures that will automatically fill a database with all the required information and calculated data. Subsequent updates of the database are facilitated by the architecture of the database. For example, new proteomes can be added by updating the previously mentioned input data. In contrast to other available systems, after installation the pre-calculated data can be exploited via both command-line and graphical interfaces.

The command-line client interface is designed to allow fast information retrieval for high throughput studies. It also facilitates the incorporation of the software in other packages or processing pipelines. The client provides database querying facilities via a number of different methods: textual searches allow access to results via sequence accession numbers or sequence descriptions, while batch queries permit submission of multiple Fasta-formated sequences. In addition, constraints of presence/absence of orthologs in specified organisms can be defined. Data can be exported in CSV, FASTA or XML formats. New user-defined file formats can easily be added to the software using a java interface included in the source code.

The graphical interface is designed to analyze smaller sets of sequences in more detail. In contrast to the command-line client, the querying functions (textual and FASTA sequence queries) are supported by interactive forms and produce results that can be visualized in more detail. More elaborate queries can also be performed, such as the selection of data according to the presence/absence of orthologous relationships in organisms specified by the user (Figure 8A). For instance, the user can retrieve all *Danio Rerio* proteins having orthologs in *Homo sapiens*, but not in *Mus musculus*. The results can be visualized through a textual description, including cross-references to Ensembl, Uniprot and NCBI-refseqp databases. For ambiguous results, the original Blast search used to generate the prediction can be directly visualized in the interface. Then, the reliable data selected by the user can be summarized using different visualization tools. Currently, two complementary tools are available: (i) a graph representation of the network of predicted relationships (Figure 8B) and (ii) presence/absence diagrams (Figure 8C), but future updates of the software are planned to enhance the visualization capabilities of the software. As in the command-line client, the data can be exported in CSV, FASTA and XML format files. All the visualizations can be exported as image files, the presence/absence diagram can be exported as a CSV matrix and the graph representation can be saved in graphML format. The graphical interface access provides access to other tools, such as batch

Figure 8 OrthoInspector graphical interface. The graphical interface provides visualization tools allowing a global view of the selected data. A. The advanced query interface allows selection of orthology/inparalogy relationships based on presence/absence criteria (pres. = presence, abs. = absence, uncar. = uncaring). B. Graph-based visualization of selected relationships. C. Presence/absence diagrams resume the repartition of orthologs/inparalogs for a family of proteins. Here, the human myotubularin-related protein 6 (mtmr6, Q9Y217) was used as the query. No orthology relationship is found in *P. falciparum*, a 1-to-1 ortholog is found in *M. musculus* (Q8VE11) and a 1-to-many relationship involving four co-orthologs is found in *E. hystolitica* (XP_653520; XP_654089; XP_648834; XP_655157). These sequences are then used as query to find potentially new sequences of the family in these organisms. Here sequences of E. histolytica make a 1-to-many relation with the seven human myotubularins and the six mouse myotubularins (these ones are inparalogs relative to entamoeba histolytica). Here are only represented the human MTMR7 (Q9Y216) and MTMR8 (Q96EF0) and the murine MTMR7 (Q9Z2C9).

generation and exportation of data, generation of database statistics or switching between different OrthoInspector compliant databases.

Conclusions

Various methods have been developed previously to predict the orthology/inparalogy relationships existing between different proteomes. In most cases, the algorithms are made publicly available in the form of binary programs that can generate either simple databases or flat files containing the complete set of predicted relationships. Until now, no comprehensive set of tools has been provided to process, query and update the datasets easily and efficiently. For this reason, we have developed OrthoInspector, incorporating fast and easy-to-use data management tools, as well as a novel algorithm to produce fast and sensitive predictions of orthology/inparalogy. The software suite, portable to any Java-compatible system and easily integrated in any workflow application, is suitable for use in high-throughput studies, which are becoming more and more predominant in the era of systems biology. Its fast and user-friendly procedures

facilitate the production of databases adapted to the user's needs. It also supports more detailed analyses of interesting orthology relationships for non-specialists, who can exploit the generated databases in a graphical interface that provides novel visualization capabilities and comparative genomics tools.

In the future, OrthoInspector will be enhanced to further improve the database update process. Although tools are currently provided to easily incorporate new genomes selected by the user, keeping up with the rate of next generation sequencing will be a major challenge. The most time-consuming step in all orthology prediction algorithms is the generation of the Blast all-versus-all searches for each new update. In spite of the efforts aimed at developing faster parallelized Blast methods [40,41], the Blast all-versus-all computational requirements grow quadratically with the addition of new proteomes. Therefore, one of our future goals will be to develop an incremental update process, minimizing the number of distance calculations required between the thousands of sequences present in the previous version of the database. We also plan to enrich the OrthoInspector system by incorporating functional annotations, such as Gene Ontology terms [42] or links to the Interpro protein domain database [43], facilitating integrated systems biology studies. Finally, to improve the interoperability of OrthoInspector with other software packages, the Ortho-XML format http://orthoxml.org will be included in the next release of OrthoInspector.

Availability and Requirements
Project name: OrthoInspector
　Project home page: http://lbgi.igbmc.fr/orthoinspector/
　Operating system: cross-platform
　Programming language: Java
　Requirements: Java JVM 1.6.x
　License: GNU GPL version 3

Additional material

Additional file 1: The complete list of the 59 studied organisms. Excel file containing the 59 studied organisms in OrthoInspector. They are classified according to their phylum.

Additional file 2: Distribution of 1-to-1 relations over 59 organisms. The normalized number of 1-to-1 relations is calculated for each organism pair. Normalisation is done by dividing the observed number of relations by the maximum number of potential relations (the size of the smallest proteome of the two compared organisms).

Additional file 3: Distribution of many-to-many relations over 59 organisms. The normalized number of many-to-many relations is calculated for each organism pair. Normalisation is done by dividing the observed number of relations by the maximum number of potential relations (the multiplication of the size of the proteomes of the two compared organisms).

Additional file 4: Test set covering 31 CMGC sub-families and 16 TKL sub-families. Excel file describing the 31 CMGC sub-families and 16 TKL sub-families used for benchmarking. Orthology predictions made by all methods for these families are in the file too.

Authors' contributions
LB carried out the algorithm conception and the software programming. JDT participated in the design and testing of the software suite. OP participated in its design and helped to draft the manuscript. OL conceived the study and participated in its design and coordination. All authors participated in writing the manuscript. All authors read and approved the final version of the manuscript

References
1.　Fitch WM: **Distinguishing homologous from analogous proteins.** *Syst Zool* 1970, **19(2)**:99-113.
2.　Peterson ME, Chen F, Saven JG, Roos DS, Babbitt PC, Sali A: **Evolutionary constraints on structural similarity in orthologs and paralogs.** *Protein Sci* 2009, **18(6)**:1306-1315.
3.　Brown D, Sjolander K: **Functional classification using phylogenomic inference.** *PLoS Comput Biol* 2006, **2(6)**:e77.
4.　Koonin EV: **Orthologs, paralogs, and evolutionary genomics.** *Annu Rev Genet* 2005, **39**:309-338.
5.　Gabaldon T: **Large-scale assignment of orthology: back to phylogenetics?** *Genome Biology* 2008, **9(10)**.
6.　Kuzniar A, van Ham RC, Pongor S, Leunissen JA: **The quest for orthologs: finding the corresponding gene across genomes.** *Trends Genet* 2008, **24(11)**:539-551.
7.　Ostlund G, Schmitt T, Forslund K, Kostler T, Messina DN, Roopra S, Frings O, Sonnhammer EL: **InParanoid 7: new algorithms and tools for eukaryotic orthology analysis.** *Nucleic Acids Res* 2010, , **38 Database**: D196-203.
8.　Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA: **The COG database: an updated version includes eukaryotes.** *BMC Bioinformatics* 2003, **4**:41.
9.　Kriventseva EV, Rahman N, Espinosa O, Zdobnov EM: **OrthoDB: the hierarchical catalog of eukaryotic orthologs.** *Nucleic Acids Res* 2008, , **36 Database**: D271-275.
10.　Muller J, Szklarczyk D, Julien P, Letunic I, Roth A, Kuhn M, Powell S, von Mering C, Doerks T, Jensen LJ, Bork P: **eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations.** *Nucleic Acids Res* 2010, , **38 Database**: D190-195.
11.　Roth AC, Gonnet GH, Dessimoz C: **Algorithm of OMA for large-scale orthology inference.** *BMC Bioinformatics* 2008, **9**:518.
12.　Uchiyama I: **MBGD: a platform for microbial comparative genomics based on the automated construction of orthologous groups.** *Nucleic Acids Res* 2007, , **35 Database**: D343-346.
13.　Chen F, Mackey AJ, Stoeckert CJ, Roos DS: **OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups.** *Nucleic Acids Res* 2006, , **34 Database**: D363-368.
14.　Chen F, Mackey AJ, Vermunt JK, Roos DS: **Assessing performance of orthology detection strategies applied to eukaryotic genomes.** *PLoS ONE* 2007, **2(4)**:e383.

15. Altenhoff AM, Dessimoz C: **Phylogenetic and functional assessment of orthologs inference projects and methods.** *PLoS Comput Biol* 2009, **5**(1): e1000262.
16. Huerta-Cepas J, Dopazo H, Dopazo J, Gabaldon T: **The human phylome.** *Genome Biol* 2007, **8**(6):R109.
17. Datta RS, Meacham C, Samad B, Neyer C, Sjolander K: **Berkeley PHOG: PhyloFacts orthology group prediction web server.** *Nucleic Acids Res* 2009, , **37** Web Server: W84-89.
18. Alexeyenko A, Lindberg J, Pérez-Bercoff Å, Sonnhammer ELL: **Overview and comparison of ortholog databases.** *Drug Discovery Today: Technologies* 2006, **3**(2):137-143.
19. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL: **BLAST+: architecture and applications.** *BMC Bioinformatics* 2009, **10**:421.
20. Milinkovitch MC, Helaers R, Depiereux E, Tzika AC, Gabaldon T: **2x genomes - depth does matter.** *Genome Biology* 2010, **11**(2).
21. Hubbard TJP, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, *et al*: **Ensembl 2007.** *Nucleic Acids Research* 2007, **35**:D610-D617.
22. Pruitt KD, Tatusova T, Maglott DR: **NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.** *Nucleic Acids Res* 2005, , **33** Database: D501-504.
23. Boutet E, Lieberherr D, Tognolli M, Schneider M, Bairoch A: **UniProtKB/ Swiss-Prot.** *Methods Mol Biol* 2007, **406**:89-112.
24. Etzold T, Ulyanov A, Argos P: **SRS: information retrieval system for molecular biology data banks.** *Methods Enzymol* 1996, **266**:114-128.
25. Bard N, Bolze R, Caron E, Desprez F, Heymann M, Friedrich A, Moulinier L, Nguyen NH, Poch O, Toursel T: **Decrypthon grid - grid resources dedicated to neuromuscular disorders.** *Stud Health Technol Inform* 2010, **159**:124-133.
26. Kasahara M: **The 2R hypothesis: an update.** *Curr Opin Immunol* 2007, **19**(5):547-552.
27. Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A, *et al*: **Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype.** *Nature* 2004, **431**(7011):946-957.
28. Van de Peer Y, Fawcett JA, Proost S, Sterck L, Vandepoele K: **The flowering world: a tale of duplications.** *Trends Plant Sci* 2009, **14**(12):680-688.
29. Rensing SA, Ick J, Fawcett JA, Lang D, Zimmer A, Van de Peer Y, Reski R: **An ancient genome duplication contributed to the abundance of metabolic genes in the moss Physcomitrella patens.** *BMC Evol Biol* 2007, **7**:130.
30. Church DM, Goodstadt L, Hillier LW, Zody MC, Goldstein S, She X, Bult CJ, Agarwala R, Cherry JL, DiCuccio M, *et al*: **Lineage-specific biology revealed by a finished genome assembly of the mouse.** *PLoS Biol* 2009, **7**(5): e1000112.
31. Lecompte O, Poch O, Laporte J: **PtdIns5P regulation through evolution: roles in membrane trafficking?** *Trends Biochem Sci* 2008, **33**(10):453-460.
32. Hughes AL: **Phylogenetic tests of the hypothesis of block duplication of homologous genes on human chromosomes 6, 9, and 1.** *Mol Biol Evol* 1998, **15**(7):854-870.
33. Hulsen T, Huynen MA, de Vlieg J, Groenen PM: **Benchmarking ortholog identification methods using functional genomics data.** *Genome Biol* 2006, **7**(4):R31.
34. Goldberg JM, Manning G, Liu A, Fey P, Pilcher KE, Xu Y, Smith JL: **The dictyostelium kinome–analysis of the protein kinases from a simple model organism.** *PLoS Genet* 2006, **2**(3):e38.
35. Plowman GD, Sudarsanam S, Bingham J, Whyte D, Hunter T: **The protein kinases of Caenorhabditis elegans: a model for signal transduction in multicellular organisms.** *Proc Natl Acad Sci USA* 1999, **96**(24):13603-13610.
36. Hunter T, Plowman GD: **The protein kinases of budding yeast: six score and more.** *Trends Biochem Sci* 1997, **22**(1):18-22.
37. Manning G, Plowman GD, Hunter T, Sudarsanam S: **Evolution of protein kinase signaling from yeast to man.** *Trends Biochem Sci* 2002, **27**(10):514-520.
38. Caenepeel S, Charydczak G, Sudarsanam S, Hunter T, Manning G: **The mouse kinome: discovery and comparative genomics of all mouse protein kinases.** *Proc Natl Acad Sci USA* 2004, **101**(32):11707-11712.
39. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S: **The protein kinase complement of the human genome.** *Science* 2002, **298**(5600):1912-1934.
40. Balaji P, Feng W, Archuleta J, Lin H, Kettimuthu R, Thakur R, Ma X: **Semantics-based Distributed I/O for mpiBLAST.** *Ppopp'08: Proceedings of the 2008 Acm Sigplan Symposium on Principles and Practice of Parallel Programming* 2008, 293-294, 296.
41. Nguyen VH, Lavenier D: **PLAST: parallel local alignment search tool for database comparison.** *BMC Bioinformatics* 2009, **10**:329.
42. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, *et al*: **The Gene Ontology (GO) database and informatics resource.** *Nucleic Acids Res* 2004, , **32** Database: D258-261.
43. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, *et al*: **InterPro: the integrative protein signature database.** *Nucleic Acids Res* 2009, , **37** Database: D211-215.

8 AN INTEGRATIVE MULTI-SCALE SOLUTION FOR DECIPHERING GENE EVOLUTION

Evolutionary studies aim to decipher the mechanisms that induce the slow transformation of genetic information and its transmission over time. Originally, such studies were based mainly on morphological or behavioural traits, but these have been largely replaced by macromolecular sequences since they provide more detailed measures of the similarities and differences between organisms. The evolutionary relationships between sequences are generally represented by the branching order of a phylogenetic tree, although alternative approaches such as 'bushes' or 'evolutionary networks' have been introduced recently, particularly in the prokaryotic domain (Bapteste et al., 2009). Today, systems biology and the abundance of data in the post-genomic era are providing unique opportunities to integrate data from many biological levels in evolutionary studies (Loewe, 2009). High throughput technologies such as transcriptomics, exomics or interactomics can help us to gain a broader view of how organism systems are evolving (Brawand et al., 2011). In this context, some authors have attempted to compile different types of data in order to confront them with evolutionary parameters and analyse the correlations between evolutionary and functional parameters. As described in chapter 5, this work identified some general evolutionary trends of biological systems, considering the system-level variation of biological sytems as an important component of Evolution (Koonin, 2009a).

The representation of evolutionary histories using phylogenetic trees is not adapted to this new data landscape. Extraction of evolutionary knowledge from trees requires manual analysis and interpretation which is not possible in high-throughput studies. Our goal was therefore to develop a new methodology, the EvoluCode (evolutionary barcode) formalism, that could exploit the new data resources efficiently and provide a multi-scale view of the evolution of genes and complex systems. The new formalism should allow the application of standard knowledge extraction techniques, such as pattern searching, clustering or classification, in evolutionary studies.

8.1 EvoluCode philosophy

The EvoluCode is designed to facilitate the combination of a systems biology approach with an evolutionary framework. We therefore defined 5 necessary criteria:

- **Evolutionary history:** a gene has an evolutionary history that resulted in its present state in a given organism. An EvoluCode associated with a gene must encode this evolutionary history.

- **Multi-scale:** to encode a 'complete' evolutionary message for a gene, we must consider not only the sequence level, but multiple biological levels at the same time. EvoluCodes must be multi-scale, integrating omics data from the gene, clade and network (system) levels.

- **Variation:** The EvoluCode should describe the 'variation' of a gene over multiple species. The variation observed in extant species is the key element that allows to elucidate the ancestral states and innovations that make up the gene evolutionary history. Moreover, the EvoluCode should quantify the variation of a set of biological parameters and differentiate cases of typical or atypical variation.

- **Formalism:** The EvoluCode should summarize the gene evolutionary history described by the variation observed for several biological parameters in several species. Formalisation will allow us to assemble heterogeneous information for each gene, thus providing a well-defined framework to apply knowledge extraction techniques (clustering, data mining…).

- **Knowledge extraction:** the evolutionary history coded in the EvoluCode must be mathematically exploitable in knowledge extraction and high-throughput studies. The multi-level biological data that compose EvoluCodes should summarize biological phenomena with a specific set of parameters, i.e. continuous or qualitative mathematical variables.

- **Visualization:** To allow a human visualization, the EvoluCode formalism must facilitate a comprehensive observation of the variations. In particular, typical or atypical parameter variations discovered in some species should be easily recognizable manually.

In addition to these specifications, we can include two constraints related to the practical construction of EvoluCodes:

- **Reference species:** In order to estimate the variation for a given biological parameter, we need to define a reference species from which to measure it. For example, defining a gene neighbourhood conservation on the genome require a reference genome. Additional genomes are compared to the reference genome and will or will not have a similar gene ordering. Consequently, an EvoluCode quantifies the variations observed in several species by comparing them to a reference species. The corollary is that there is a different set of EvoluCodes for the gene set of each species.

- **Evolutionary scale:** When representing the variation relative to a reference organism in several species, we need to carefully choose the species composing the evolutionary scale of interest. This must correspond to a reliable evolutionary message, avoiding a hazardous species composition such as mixing 10 primates and 1 yeast. Similarly, large phylogenetic scales, for example all bacteria, could be problematic for the barcode approach. In such cases, variation is preponderant to conservation and cannot be quantified in a realistic manner. However, assuming a reasonable and continuous species composition, EvoluCodes can describe many evolutionary scales (e.g. primates, vertebrates, ascomycetes, eudicotyledons, etc.).

To resume, EvoluCode is a synthetic representation for the integration, the visualisation and the analysis of diverse parameters extracted from multiple biological levels in multiple organisms. The barcodes can be easily updated and can be adapted to any kind of biological parameter. A key feature of the EvoluCode formalism is the ability to describe the specific state of the parameters in their "evolutionary context". Thus, for each species, the state of a given parameter is defined as typical or atypical when compared to a reference species with respect to the generally observed state in the same species. The integration of multi-scale parameters defines a combination of typical or atypical states and describes a complex evolutionary scenario that could not be resumed for example with a single parameter such as sequence conservation.

8.2   Collecting evolutionary data

The first step in any knowledge extraction process is the collection of information from a variety of sources for the purposes of analysis. To evaluate the suitability of our EvoluCode approach for evolutionary-based knowledge discovery, we generated a collection of multi-level parameters. Concerning the evolutionary scale, we chose the vertebrate phylum and selected 16 different vertebrate species representing mammal, sauropsid, batrachian and fish phyla. High quality, almost complete genome sequences are available for all of these species. As a reference species, we chose the human for the annotation quality of its genome and proteome. For each of the protein coding genes (19778 genes), we generated a *de novo* dataset for all biological parameters, using three protocols briefly described in the following paragraphs.

8.2.1   Synteny data

The genomic context is an important parameter in gene evolutionary histories since chromosomal rearrangements are a key mechanism in genome evolution (Tang, 2007). In particular, the conservation of genomic context is widely used in the reconstruction of ancestral genome states (Muffato and Crollius, 2008; Rascol et al., 2007) and has sometimes been related to gene functions (Michalak, 2008; Zaslaver et al., 2011). The conservation of local gene neighbourhood (also known as microsynteny) can also be related to the functional aspect of genomes. It has been previously shown that neighbouring genes can be related to a coordinated transcription, describing genomic regulatory blocks. Cis-acting regulatory elements that control multiple genes are frequent in prokaryotes and are generally organised into functional operons (Osbourn and Field, 2009). More recently, a similar organisation has been observed in plants, fungi, insects and mammals with a much lower frequency (Engstrom et al., 2007; Field and Osbourn, 2008). Recently, a study showed that ~12% of the ancestral bilaterian genome contained genomic regulatory blocks characterized by transcriptional enhancers controlling developmental genes and that cis-regulatory constraints are crucial in determining metazoan genome architecture (Irimia et al., 2012). We developed local software using data from the Ensembl genome database (Hubbard et al., 2007) for the identification of local synteny between the human genome and each of the 16 other vertebrate genomes. Thus, we determined whether synteny exists to the right and left of each gene, between each species and the human reference.

8.2.2   Orthology data

Gene gains and gene losses are an important driving force in genome evolution (Kaessmann, 2010) and are thought to be a major contributor to evolutionary innovation. A single gene family can describe complex evolutionary patterns with independent expansion in some phyla and gene loss in another (Ruano-Rubio et al., 2009) and these patterns have been used for example, to study co-adaptation and co-evolution of proteins (Pazos and Valencia, 2008). To take into account such phenomena, we extracted orthology and inparalogy data from OrthoInspector for the human reference and the 16 vertebrate proteomes and used this knowledge to describe the stability of the protein family.

8.2.3   Multiple Alignment data

The comparison of protein and nucleic acid sequences plays a major role in the understanding of sequence/structure/function/evolution relationships (Lecompte et al., 2001). Originally used in evolutionary analysis, multiple alignments are now essential to highlight conserved functional features (Levasseur et al., 2008) or to improve the prediction of 3D structures (Moult et al., 2005). When exploited in expert annotation processes, multiple alignment of complete sequences (MACS) are a compact source of information from which numerous evolutionary parameters can be extracted (figure 8-1). Thus, for each human reference gene, we constructed a protein MACS by including all vertebrate homologous genes in the alignment. High-quality MACS were built using a computational pipeline, called PipeAlign, and the alignments was annotated by the MACSIMS information management system (see material and methods, paragraph 6.2.3).
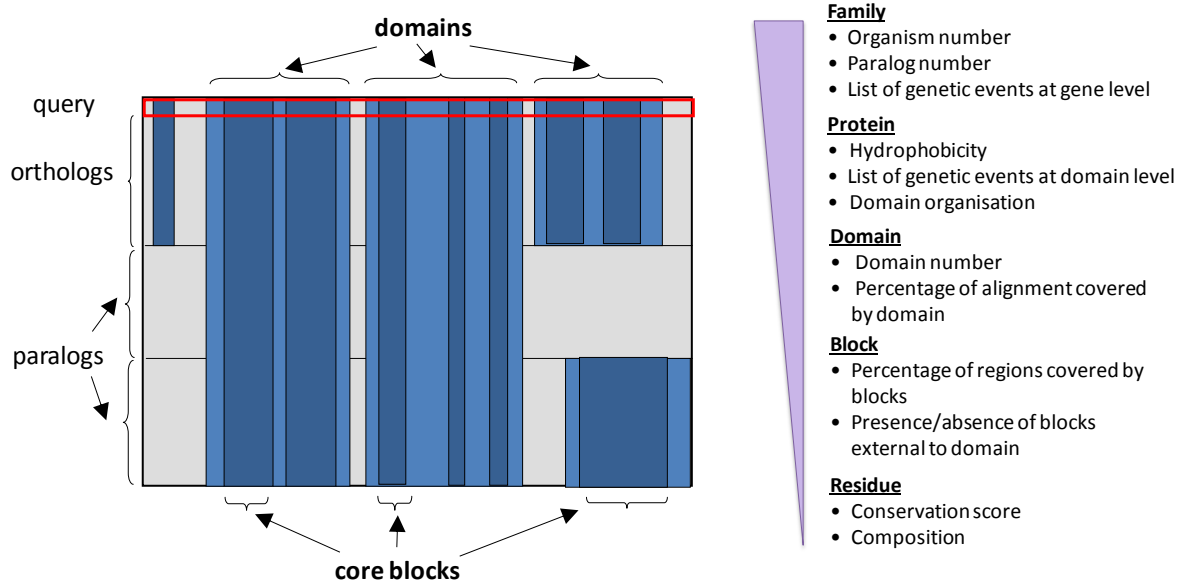
Figure 8-1. Some example of evolutionary parameters that can be automatically retrieved by MACSIMS when analyzing a MACS.

In the context of this work, we performed an in-depth study of state-of-the-art multiple alignment methods (publication n°4), which resulted in a refinement of the original PipeAlign. The modified pipeline allowed us to perform an automated high-throughput extraction of numerous biological parameters for the whole human proteome. The complete data for this work represents more than 280,000 genome context mappings, more than 11,000,000 orthologous relations and more than 500 000 vertebrate sequences aligned in 19778 high-quality and completely annotated MACS.

8.2.4 Data quality

The quality of the data collected for the construction of the EvoluCodes is a major issue, particularly since most the data was produced by high-throughput technologies, which are notoriously error-prone, inconsistent and incomplete (Pop and Salzberg, 2008). In collaboration with the team of P. Pontarotti (Marseille), we studied the impact of genome sequencing and protein sequence prediction errors on evolutionary studies, specifically in the analysis of asymmetric evolution after duplication (AED) events. An AED event can be observed after a duplication event, when the homologous sequence with higher similarity is relocated in the genome and the positional homolog (with conserved gene neighbourhood) shows a lower similarity (figure 8-2). In fact, AED describe events where the local homolog has evolved significantly faster than the relocated homolog, which is unexpected since it is generally hypothesized that the gene copy that retains the genome context will be more conserved.
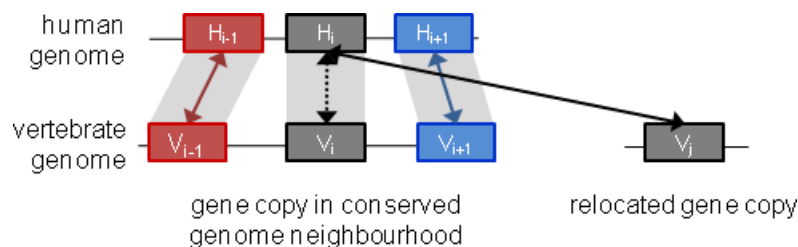
Figure 8-2. Asymmetric evolution after duplication (AED). *H_i is the human reference gene and its homolog V_i is detected in a vertebrate genome that maintain similar genome neighborhood. Full arrows indicate homologs based on sequence similarity and gray shadows link positional homologs.*

To test this hypothesis, we searched for AED events in the 16 vertebrate genomes. However, most of the potential AED events we detected were in fact false positives, due to the low quality of the underlying protein sequences. In fact, up to half of all protein sequences in these vertebrate genomes contained putative erroneous insertions, deletions or suspicious segments. In publication n°3, we discussed this quality issue, its implication for evolutionary analysis and how the true functional significance of AED events is masked by the sequence errors.

8.3 EvoluCodes and high-throughput analysis

8.3.1 Evolutionary histories of the human proteome

The EvoluCode formalism was initiated in the context of a French ANR project, EvolHHuPro (Evolutionary History of Human Proteome) in collaboration with the groups of Pierre Pontarotti (Marseille) and Anthony Levasseur (Aix Marseille). The aim of this project was to provide a complete set of the evolutionary histories (cascade of phylogenetic events) for the human proteome and their genome-scale analysis. In this context, high-quality annotated multiple alignments were built for each protein in the human proteome and our collaborators constructed expert annotated phylogenetic trees based on the alignments. Then, with the help of the DAGOBAH platform, all vertebrate genetic events leading to the current state of the human proteome were inferred (domain gain/loss, rearrangements…). The phylogenetic trees and genetic events are publicly available in a database dedicated to chordate evolutionary histories (Levasseur et al., 2012b).

8.3.2 Human proteome EvoluCodes

EvoluCode is a data formalism that can be used to summarize the evolutionary history of a gene and to facilitate automatic, high throughput analysis. In order to construct EvoluCodes for the complete human proteome, we created the pipeline shown in figure 8-3. We integrated and normalized all the data described in section 8.2 and selected 10 representative parameters from different levels (genome, protein sequence, family, etc.) (see publication n°2 for parameter descriptions). For a given gene, we created an EvoluCode as a 2D matrix representing the variation existing between a particular vertebrate gene and its corresponding human reference, independently for each biological parameter. This matrix structure provides an efficient formalism for mathematical explorations of the combined data and the set of human EvoluCodes can be used as input for knowledge extraction strategies.

A parallel step of the EvoluCode pipeline is the statistical description of all the biological parameters in all vertebrate species. This analysis highlights what are typical or atypical values when comparing a vertebrate parameter value to the human reference value (figure VIII-3). For example, a sequence identity conservation of 60% between a zebrafish gene and its human reference is a relatively common case when looking at all sequence conservation of fish and human genes. This value will be labelled as 'typical' in zebrafish (green colour). In contrast, this level of sequence identity between primate genes is rare and will be labelled as an atypically high value (red colour). Applying this statistical description for all parameters transforms the 2D EvoluCode matrix into a mathematical description summarizing all the variation that a gene underwent during its vertebrate evolution. In the course of this work, we tested several statistical models for the description of the 'typical' nature of a value (figure 8-3). However, the heterogeneous parameters composing the EvoluCode present different statistical distributions (normal, skew normal or even multimodal distributions). This heterogeneity made it difficult to apply a global statistical model and the current version of the EvoluCode uses the descriptive non-parametric properties of boxplots to estimate atypical values.

Consequently, EvoluCodes describe a profile of typical or atypical variations that correspond to a particular evolutionary scenario. Figure 8-4 shows several examples of EvoluCode profiles that can be directly associated with an evolutionary message.
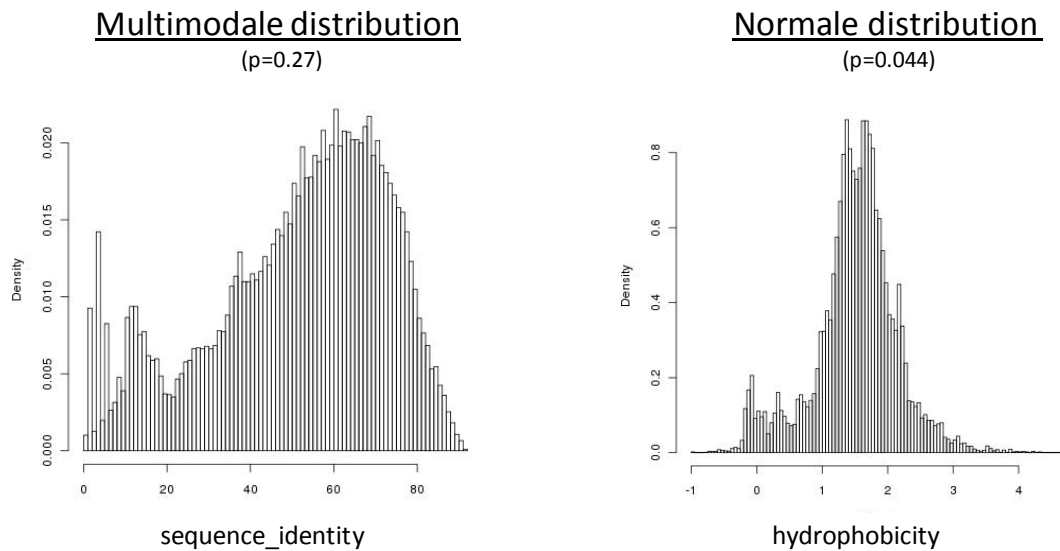
Figure 8-3. Some examples of statistical distributions observed in EvoluCode parameters. *The p values correspond to Kruskal-Wallis test of normality.*

The human EvoluCodes that we created in this work respect the five criteria that emerged from our reflexion into studying evolution in a high-throughput way and with multi-level data corresponding to the systems biology philosophy. Our first version of the EvoluCode is a formalised representation of the multi-scale variation of a human gene and effectively describes its vertebrate evolutionary history.

Human genes are reference genes
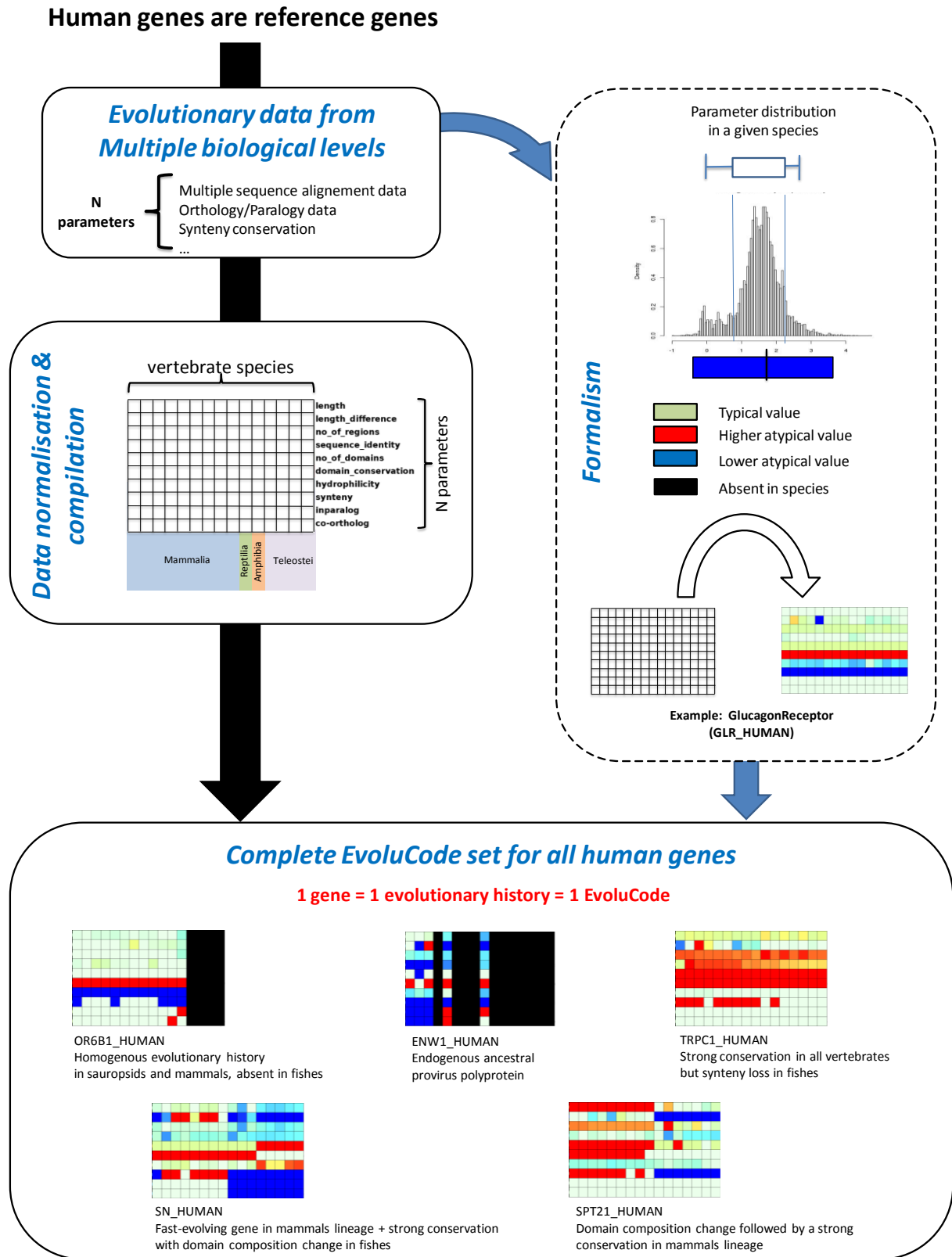
Figure 8-4: Overview of the EvoluCode construction process. *Different evolutionary parameters are compiled from several sources and organised in the EvoluCode framework. A parallel process statistically describes the variation of these parameters when comparing the vertebrate's parameter states compared to the human reference. The statistical model is used to create the EvoluCode formalism that allows a direct human visualization of variation profiles corresponding to different gene evolutionary histories.*

8.4   EvoluCodes and extraction of evolutionary knowledge

Classical phylogenetic approaches study gene evolutionary histories at the family level for different purposes, such as predicting gene functions (Jiang, 2008), gene interactions (Pellegrini, 2012) or for the analysis of evolutionary mechanisms (Keeling et al., 2005). The EvoluCode approach provides the possibility to extend such analyses to the genome level or even to the species level. EvoluCodes summarize the variation observed during a gene evolutionary history, by integrating normalized parameters in a 2D matrix. As a consequence, we can rapidly perform large scale evolutionary analysis using various standard approaches for formal knowledge extraction. The goal of knowledge extraction techniques is to discover interesting patterns or relationships in large datasets (even in complex and high-dimensional ones), to use these relationships to make predictions and to present the discovered knowledge in a comprehensible form. Among the numerous possibilities, we explored three different approaches (more extensively described in publication n°2):

8.4.1   Identification of interesting relationships in human evolutionary histories

By applying a knowledge extraction technique such as clustering, we can regroup genes with similar evolutionary histories and investigate their functional significance using standard functional enrichment software. We applied several non-supervised clustering techniques on our EvoluCodes, notably a neural network approach based on Kohonen clustering and a paramagnetic approach using Potts clustering and performed a preliminary study to establish the coverage of the clusters predicted by the two methods. The Potts algorithm constructed an optimal clustering with 303 clusters (the improved Potts clustering that we used automatically predicts the optimal number of clusters). Consequently, in order to obtain comparable results, we imposed a neural network grid of 17x18 cells for the Kohonen clustering (306 clusters). The Jaccard (Jaccard, 1901) and Baroni-urbani (Baroni-Urbani, 1980) similarity coefficients were calculated for all cluster pairs. Figure 8-5 shows the Jaccard similarity between all clusters containing an intersection of one EvoluCode (a similar profile is described with the Barroni coefficient). About one third of the clusters present a good similarity between the 2 clustering methods. Two thirds of the clusters calculated by Potts clustering are dispersed in many clusters produced by Kohonen clustering. However, 59% of the EvoluCodes are shared by the 33% most similar clusters. Highly similar clusters are larger than low similarity clusters. These results showed that both methods cluster 59% of the EvoluCodes in similar large clusters. This was confirmed later by the functional enrichment analysis performed with the clusters of both methods, where comparable functional enrichments were observed in most similar clusters.

We decided to perform the subsequent studies with the results of the Potts clustering, as the method implements statistical approaches to estimate the optimal number of clusters. The simple structure of the EvoluCode allowed a fast Potts clustering of evolutionary histories (<15minutes) at the human genome scale. The 303 clusters inferred by the Potts clustering indicate that there are a limited number of evolutionary histories for human protein-coding genes (around 300). Moreover, most clusters were characterized by a significant functional enrichment, denoting an unexpectedly strong correlation between the evolutionary profile of a gene and its function.
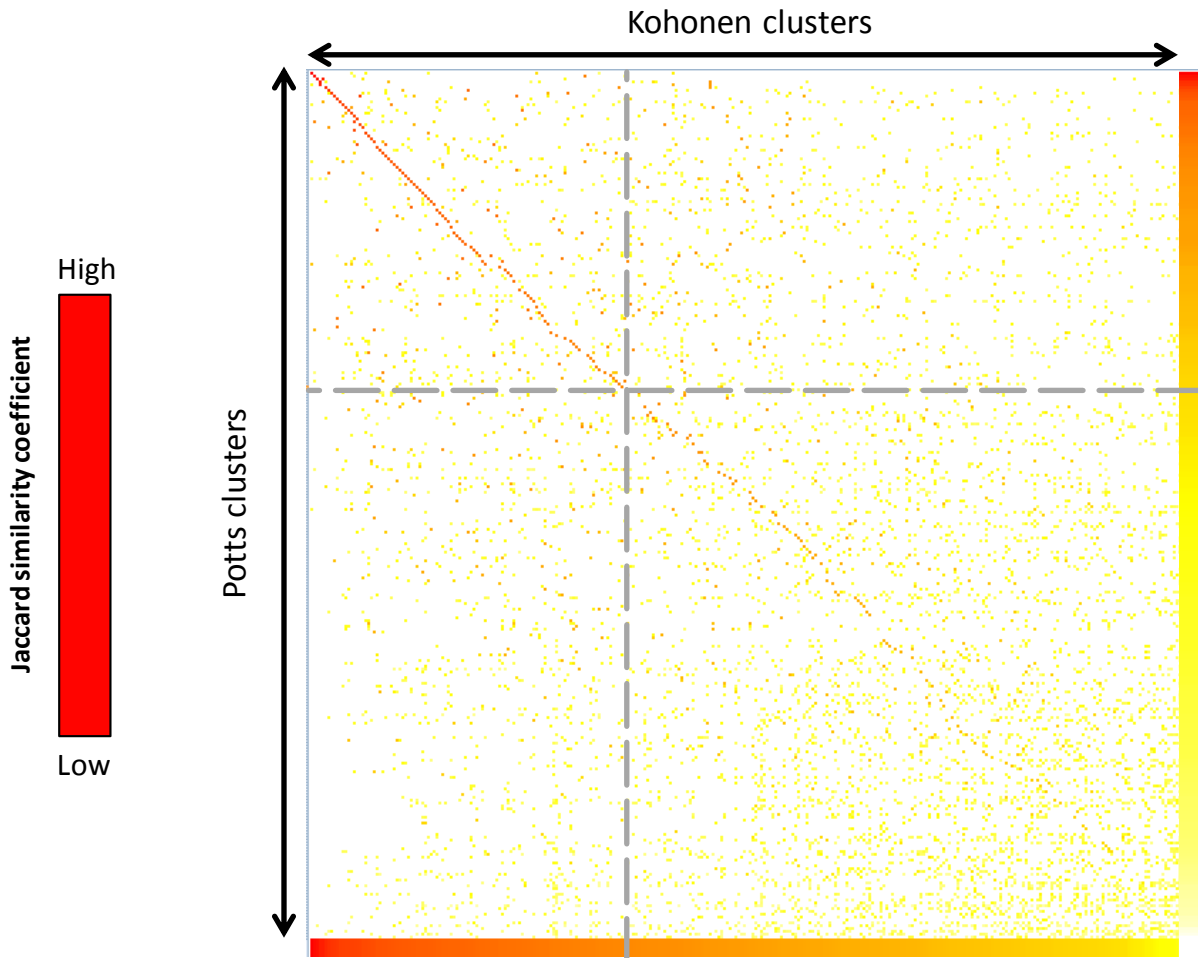
Figure 8-5. Jaccard similarity coefficient between all EvoluCode clusters predicted by Potts and Kohonen clusterings. Each cell of the matrix corresponds to an intersection between two clusters. A similarity value is assigned to the corresponding pair. Cluster pairs with a large intersection are red and cluster pairs with a low intersection are yellow.

8.4.2   Classification and prediction of protein function

The previous sections illustrated the application of knowledge extraction techniques for description purposes. However, knowledge extraction can also be used for prediction of new information, by using a reduced set of variables or parameters to predict unknown values of other variables of interest. This first step in this process is the use of data reduction or projection methods to find pertinent features that represent the data depending on the goal or specific question of the user. To explore this possibility, we analysed the EvoluCodes corresponding to all the human genes coding multi-pass transmembrane proteins. These proteins are characterized by alternate intra-membrane/extra-membrane domains with conserved hydrophobic intra-membrane domains and less conserved extra-membrane domains (figure 8-6). We extracted all the human genes annotated as 'multi-pass membrane protein' in Uniprot and performed a Multiple Correspondence Analysis (MCA) of the corresponding EvoluCodes in order to perform a data reduction of this high-dimensional dataset. We constructed a 2D projection of Evolucodes with the help of the MCA and

performed a clustering on this projection. The two axes of the projection represent radically different biological information. One axe is mainly related to the structural and chemical characteristics of multi-pass proteins, whereas the second axe represents the evolution of the family (synteny, duplications...).
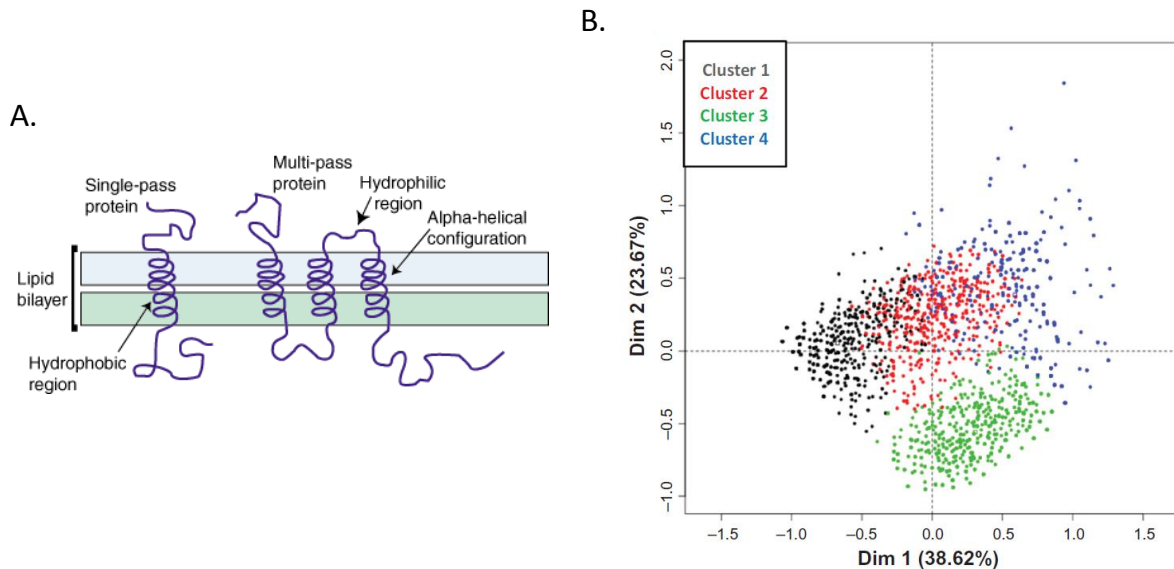
Figure 8-6. Using EvoluCode and its multi-level perspective to create evolutionary predictive models. (A) A representation of structural and physic-chemical properties of multi-pass transmembrane proteins. Adapted from http://www.sparknotes.com/biology/. (B) The model calculated by multiple correspondance analysis that separates multi-pass proteins in 4 classes based on different evolutionary and structural characteristics.

This result illustrates the potential for knowledge discovery of the EvoluCodes. In a single analysis, we link structural and evolutionary characteristics of a large group of proteins that are not clearly related at the sequence level. The 2D projection can then be used as a predictive model for the functional annotation of multi-pass genes that are currently annotated as 'putative' in Uniprot. Indeed, 3 out of the 4 multi-pass protein clusters demonstrate a significant functional enrichment. For example, a large proportion (55%) of cluster 1 genes is known to be associated with olfactory and taste perception.

8.4.3   Presentation of knowledge in a comprehensible form

The summarization capabilities of the EvoluCodes can be illustrated by observing the repartition of gene evolutionary histories over the human chromosomes. To simplify the EvoluCodes and facilitate visualization, the 2D EvoluCodes were compressed into a single 1D vector and mapped to the chromosomes (figure 8-7). Several chromosomal clusters with similar evolutionary histories were highlighted, including a number of published clusters such as the olfactory receptor clusters (Hasin et al., 2008) and the keratin and keratin-associated protein clusters (Wu et al., 2009; Zimek and Weber, 2005).
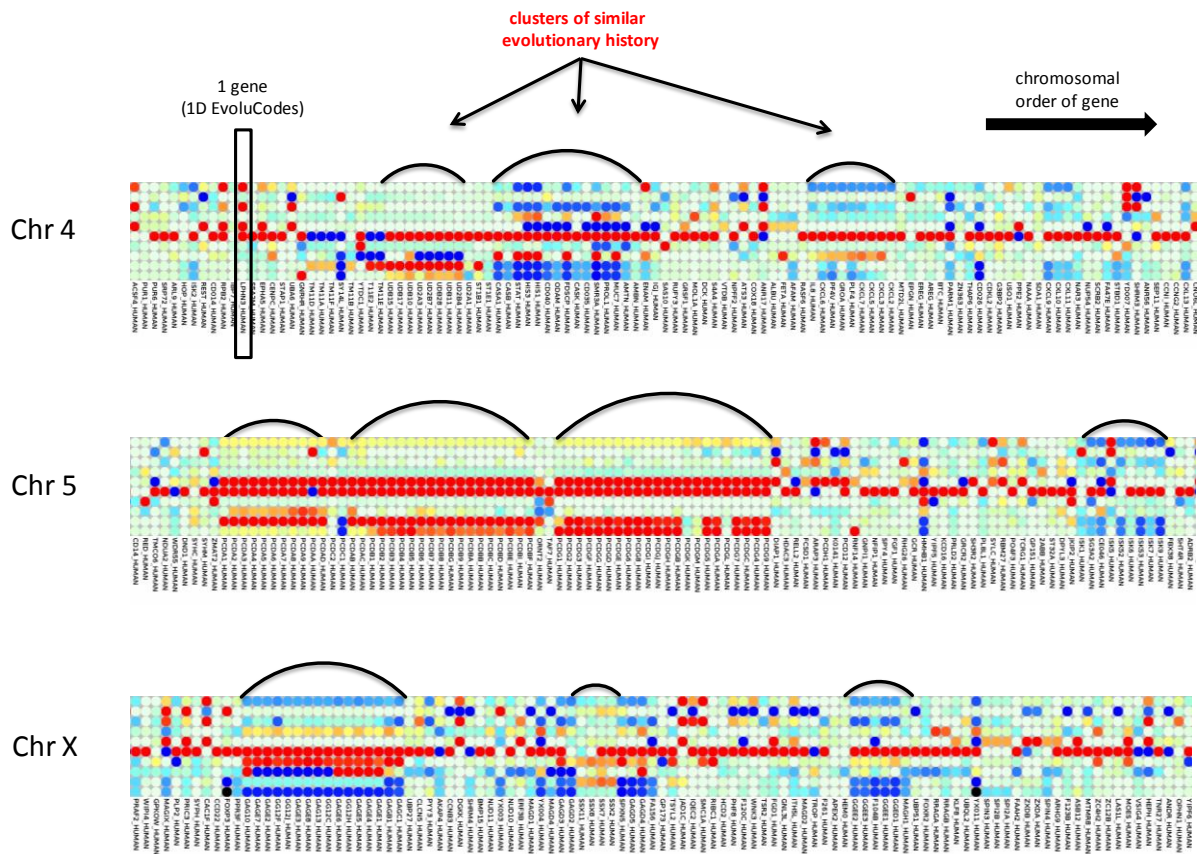
Figure 8-7. Several examples of chromosomal clusters with similar evolutionary histories. 2D EvoluCodes are reduced to 1D vertical vectors for visualization purposes. The gene order corresponds to the chromosomal order, but genetic distances are not taken into account. Black arcs highlight evolutionarily-related clusters.

8.5 Conclusion

The example applications described here demonstrate that the EvoluCode barcode formalism represents a powerful tool for the visualization and quantitative analysis of complex evolutionary histories in high throughput studies. We have constructed EvoluCodes representing the evolutionary histories of the complete human proteome at the vertebrate evolutionary scale. However, the EvoluCode approach can be applied to any reference species and to different evolutionary scales. Similarly, we integrated 10 evolutionary parameters in the first version of the EvoluCode but it can easily be extended with new biological parameters. In contrast to standard molecular phylogenies and their inherent reliance on a certain amount of residue conservation, EvoluCodes represent a methodological advantage as they allow to compare multi-level variables. Future versions will compile new data, such as genomic context (CNVs, number of exons, inter-genic distances...) or interactomic data that are available in the laboratory databases (gps.igbmc.fr).

Finally, we performed several case studies to test the reliability of the evolutionary knowledge that is summarized by our EvoluCodes. Using formal knowledge extraction techniques, we confirmed their

suitability for answering diverse evolutionary questions in large-scale studies. The exploitation of EvoluCodes in the context of cellular biological systems will be discussed in the next chapter.

8.6 Publication 2. EvoluCode: Evolutionary Barcodes as a Unifying Framework for Multilevel Evolutionary Data

Publication n° 2

ORIGINAL RESEARCH

EvoluCode: Evolutionary Barcodes as a Unifying Framework for Multilevel Evolutionary Data

Benjamin Linard[1], Ngoc Hoan Nguyen[1], Francisco Prosdocimi[2], Olivier Poch[1] and Julie D. Thompson[1]

[1]Laboratoire De Bioinformatique Et Génomique Intégratives, Institut de Génétique et de Biologie Moléculaire et Cellulaire CNRS/INSERM/UDS, Illkirch, France. [2]Federal University of Rio de Janeiro, Rio de Janeiro, Brazil. Corresponding author email: julie.thompson@igbmc.fr

Abstract: Evolutionary systems biology aims to uncover the general trends and principles governing the evolution of biological networks. An essential part of this process is the reconstruction and analysis of the evolutionary histories of these complex, dynamic networks. Unfortunately, the methodologies for representing and exploiting such complex evolutionary histories in large scale studies are currently limited. Here, we propose a new formalism, called EvoluCode (Evolutionary barCode), which allows the integration of different evolutionary parameters (eg, sequence conservation, orthology, synteny …) in a unifying format and facilitates the multilevel analysis and visualization of complex evolutionary histories at the genome scale. The advantages of the approach are demonstrated by constructing barcodes representing the evolution of the complete human proteome. Two large-scale studies are then described: (i) the mapping and visualization of the barcodes on the human chromosomes and (ii) automatic clustering of the barcodes to highlight protein subsets sharing similar evolutionary histories and their functional analysis. The methodologies developed here open the way to the efficient application of other data mining and knowledge extraction techniques in evolutionary systems biology studies. A database containing all EvoluCode data is available at: http://lbgi.igbmc.fr/barcodes.

Keywords: systems biology, evolutionary history, multilevel data analysis, data representation, data visualization, data mining

Introduction

Systems biology aims to understand the structure and dynamic behavior of complex biological systems by modeling the components and their interactions at different functional levels.[1,2] Such a comprehensive understanding requires the integration of large-scale experimental data with computational analyses and mathematical modeling approaches.[3] In particular, successful systems biology will rely on our ability to integrate different types of multi-scale data across various levels of complexity,[4] from individual molecules such as proteins, metabolites, etc. to cells, tissues, organisms or even ecosystems. These different levels are now being described by the large volumes of experimental data resulting from genomics technologies such as next-generation sequencing, transcriptomics, interactomics, etc. This high throughput data is characterized by a low signal-to-noise ratio and data mining and extraction of significant, pertinent knowledge are major challenges. In this context, the field of evolutionary systems biology aims to combine the modeling aspects of current systems biology with the long-standing quantitative experience in evolutionary genetics in order to uncover the general trends and principles underlying the evolution and function of complex biological networks.[5,6]

Evolutionary based inference provides an incredibly powerful tool for comparing multiple sources of data, since features that are maintained in several organisms tend to be functionally important while variations or differences may indicate key innovations. Comparative studies of individual components, such as proteins, have been widely used and are generally based on multiple sequence alignments and the subsequent reconstruction of a phylogenetic tree. Evolutionary histories are then typically represented by mapping major events (duplications, speciations, gene loss, domain reorganization, etc.) onto the tree. Some recent work has applied these methodologies at the genome scale, for example to build the complete collections of gene phylogenies (phylomes) in the PhylomeDB database,[7] or in the construction of the Chordate Proteome History Database (ioda.univ-provence.fr). At the level of protein networks or pathways, the reconstruction of the evolutionary histories is more complex, since the interactions between the different

molecular components have to be taken into account and changes at one biological level often have consequences on the evolution of other levels.[8–11] Therefore, additional information concerning genome context, gene expression, molecular interactions, etc. is needed to successfully model the dynamic behavior of the system.

A number of groups have performed genome-scale studies aimed at investigating the potential correlations between variables characterizing different aspects of protein network functions and evolution.[12–14] For example, positive correlations were observed between gene essentiality, duplicability and protein connectivity, estimated by the number of interaction partners in the networks.[15,16] Other recent studies have shown negative correlation between expression breadth, ie, the number of tissue types in which genes are expressed, and protein evolutionary rates.[17] While these studies were limited to the correlations observed between two variables, others have attempted to compile more diverse sets of evolutionary variables. Thus, principal component analysis was used to investigate the relationships between seven genome-related variables, identifying three main axes reflecting a gene's "importance", "plasticity" and "adaptability".[18] Waterhouse et al also examined the links between evolutionary and functional traits, by classifying metazoan orthologs as "essential" or "non-essential" and confronting these classes with various evolutionary variables.[19] Although these studies have revealed several interesting trends, new standardized methodologies and tools are now needed that allow the integration of larger, more diverse sets of multi-level data and efficient, quantitative analyses at the genome scale. Similarly, despite some attempts to develop tools providing global overviews of complex evolutionary scenarios,[20] original visualization tools will be required to facilitate rapid identification of specific behaviors.

Here we describe a novel formalism, called EvoluCode, or the Evolutionary barCode, which allows the integration of different data types in a unifying framework. Thus, a barcode is assigned to each component in a biological system and diverse evolutionary parameters from different biological levels can be incorporated, facilitating multi-scale evolutionary analyses. Visualization tools have also

been developed to allow the human expert to view the barcodes and to identify interesting patterns in both low and high throughput studies. In order to evaluate the pertinence of the evolutionary barcodes and to test their ability to represent complex evolutionary histories, we constructed evolutionary barcodes for the complete proteomes of 17 vertebrate species. In this context, we incorporated a number of different evolutionary variables, including primary sequence data, genome neighborhood and evolutionary conservation, but the barcode formalism can be easily extended to incorporate other variables representing different biological features. At this stage, the values of the barcode parameters are normalized to allow quantitative analyses and automatic comparisons, using standard data mining techniques such as clustering or classification. We show that, in addition to highlighting general evolutionary trends, the barcodes facilitate the identification of specific evolutionary histories, such as strict conservations or significant gene family expansions. Two genome-scale analyses were then performed. First, by mapping the protein barcodes onto the human genome and visualizing the results in our barcode visualization tool, we were able to identify a number of previously described chromosome gene clusters. Second, automatic barcode clustering and functional enrichment analysis allowed us to identify specific sets of proteins that have experienced similar evolutionary histories. In a more detailed study, automatic clustering of multi-pass membrane proteins highlighted a number of particular evolutionary trends that are inherent to these protein families. Finally, as a proof of concept we demonstrate the potential of our evolutionary barcodes for biological pathway analysis. All data described in this publication are available online at: http://lbgi.igbmc.fr/barcodes.

Material and Methods
Protein test set
A reference set of human proteins was retrieved from the Human Protein Initiative (HPI) project.[21] This project defined a master human proteome set, according to the quality standards set by the UniprotKB/Swiss-Prot[22] databases, resulting in a total of 19778 human reference protein sequences (with 1 protein reference per coding gene). We created our own database of vertebrate proteomes, by selecting an additional 16 vertebrate species that best represent major vertebrate phyla, ie, fish, batracia, sauropsida and mammals (species list in supplementary Table 1). The complete proteomes for these organisms were downloaded from Ensembl (version 51),[23] to create a local database with more than 500,000 sequences. Each human protein was then used as a query for a BlastP[24] search in this local protein sequence database.

Multiple sequence alignment construction
For each human reference sequence, a modified version of the PipeAlign[25] protein analysis pipeline was used to construct a MACS (Multiple Alignment of Complete Sequences) for all sequences detected by the BlastP search with E < 10–3 (maximum sequences = 500). PipeAlign integrates several steps, including post-processing of the BlastP results, construction of a MACS with DbClustal,[26] verification of the MACS with RASCAL[27] and removal of unrelated sequences with LEON.[28] In this modified version, DbClustal was replaced by the MAFFT program,[29] since the computational speed of MAFFT is better suited to high throughput projects. The MACS obtained from this pipeline were then annotated with structural and functional information thanks to MACSIMS,[30] an information management system that combines knowledge-based methods with complementary ab initio sequence-based predictions. MACSIMS integrates several types of data in the alignment, in particular Gene Ontology annotations,[31] functional annotations and keywords from Swissprot, and functional/structural domains from the Pfam database.[32]

Local genome neighborhood conservation
The chromosomal localization of all genes coding for the protein sequences was obtained from Ensembl. Locally developed software was used to identify conserved local synteny between the human genome and each of the 16 other vertebrate genomes. To achieve this, the chromosomes in each genome are represented as a linear sequence of genes. For each human reference sequence, the local syntenic homolog HREF was defined at position i on the human genome and its upstream and downstream neighbors (HREF−1

and HREF+1 respectively) were identified. For each of the 16 vertebrate genomes, the sequences with the highest similarity to HREF−1 and HREF+1 were selected from the MSA, and denoted Vn_Sim−1 and Vn_Sim+1 respectively, where Vn refers to one of the 16 vertebrate genomes. A local synteny homolog, exists for HREF and genome Vn if:

i. homologs were found in Vn for HREF−1 and HREF+1,
ii. the separation between the highest similarity homologs, denoted Vn_Sim−1 and Vn_Sim+1, on the genome was less than 5 genes,
iii. a homolog of HREF was found on the genome between Vn_Sim−1 and Vn_Sim+1.

The homolog of HREF localized between Vn_Sim−1 and Vn_Sim+1 with the highest similarity to the human reference sequence was then defined as the syntenic homolog. Genes with ambiguous genomic locations, such as scaffolds etc, were discarded since the synteny relationship could not be reliably established. In addition, local or tandem duplications were excluded since the genome contexts of the two gene copies were similar.

Orthology data

Orthologs are homologous genes that diverged from a single ancestral gene in their most recent common ancestor via a speciation event, whereas paralogs are homologs resulting from gene duplications.[33] Paralogs are considered as "inparalogs" when they are produced by duplication(s) subsequent to a given speciation event. In this context, several inparalogs of a given species (recently duplicated genes) are "co-orthologs" relative to the non-duplicated ortholog of a second species.

Orthologous relationships were generated with the OrthoInspector software.[34] Orthology inference is based on a blast all- vs. -all generated with a 10-9 Expect value threshold. Each human reference sequence was used as a query to retrieve human inparalogs and co-orthologs in each of the 16 vertebrate organisms.

Barcode construction for the human proteome

Evolutionary barcodes were constructed for all human reference proteins. Each barcode includes

a number of different evolutionary parameters that were extracted from the annotated multiple alignments, synteny analysis and orthology data described above (Fig. 1A). For each of the vertebrate organisms included in this work, the most closely related homolog (based on percent residue identity) was identified in the MACS and seven parameters were extracted:

- *length*: the length of the vertebrate sequence.
- *length_difference*: the difference in length between the human reference protein and the vertebrate
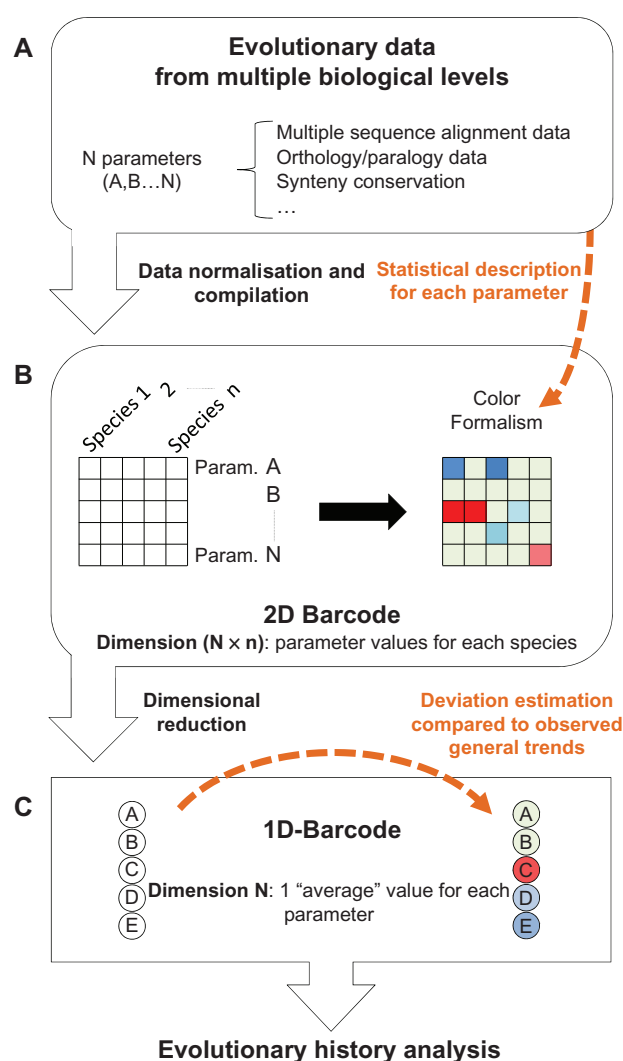
Figure 1. Schematic view of the methodology used to produce the barcodes representing the evolutionary histories of the human proteome. Three main steps are shown. (**A**) Multiple evolutionary parameters are selected and described statistically. (**B**) The values of these parameters for different species are compiled in a 2D barcode. The statistical description of these parameters is used to define a colour code for the barcode. (**C**) For each barcode, a lower dimensional barcode (1D-barcode) is generated.

sequence. This parameter may indicate potential genetic events, such as exon/domain gains or losses, but may also highlight protein fragments or sequence prediction errors.

- *no_of_regions*: the number of conserved regions defined by MACSIMS and shared between the human reference protein and the vertebrate sequence.
- *sequence_identity*: the percent residue identity shared between the human reference protein and the vertebrate sequence.
- *no_of_domains*: the number of known protein domains in the vertebrate sequence. These domains are based on annotations from the Pfam database.
- *domain_conservation*: a qualitative parameter indicating changes in the domain structure of the vertebrate sequence compared to the human reference protein. This parameter identifies an unchanged domain organization, domain gains, domain losses or domain shuffling.
- *hydrophilicity*: the average hydrophilicity of the vertebrate sequence.

Two parameters, representing orthology/paralogy data were also extracted from the OrthoInspector database:

- *inparalog*: the number of human inparalogs with respect to the specific vertebrate organism. This parameter represents the recent duplicability of a human gene compared to the other species.
- *co-ortholog*: the number of co-orthologs in the specific vertebrate species with respect to human. This parameter indicates the number of gene duplications in the non human lineage.

Finally, a parameter representing the genome neighborhood between the human and each vertebrate species was calculated:

- *synteny*: categorical parameter with 3 values: (i) synteny on both sides of the gene, (ii) synteny either downstream or upstream of the gene (iii) no synteny.

All these evolutionary parameters were then organized in a 2D matrix, which we will refer to as the "2D-barcode" (Fig. 1B). Each row of the 2D-barcode represents one parameter (denoted A, B … to N). Each column of the 2D-barcode represents one species (denoted 1, 2 … n) and the intersection between rows and columns corresponds to the value or the state of one specific parameter, in one particular species.

To facilitate visualization of the 2D-barcode, a color is assigned to each matrix cell representing typical or atypical parameter values (Fig. 1B). To do this, the distribution of each parameter in each organism is first described by the sample percentiles, using the Emerson-Strenio formulas[35] implemented in the R software. These nonparametric statistics are used to avoid bias due to non-Gaussian distributions of some of the parameters. The Emerson median, whiskers and hinges are then used to define three intervals that are assigned color gradients. The first interval (IT1) is assigned a blue-to-green gradient and represents values that are lower than what is generally observed for a specific parameter in a specific organism:

$$IT1 = \left\{ x \in \mathbb{R} \mid lower\_whisker \leq x \leq lower\_hinge + \left(\frac{median - lower\_hinge}{2} \right) \right\}$$

The second interval IT2 (green color) represents values that correspond to what is generally observed for a specific parameter in a specific organism.

$$IT2 = \left\{ x \in \mathbb{R} \mid IT1 < x < IT3 \right\}$$

The third interval (IT3) is assigned a green-to-red gradient and represents values that are higher than what is generally observed for a specific parameter in a specific organism.

$$IT3 = \left\{ x \in \mathbb{R} \mid median + \left(\frac{upper\_hinge - median}{2} \right) \leq x \leq upper\_whisker \right\}$$

Finally, the 2D-barcodes are reduced to a single dimension (Fig. 1C), called the 1D-barcode. The 1D-barcode is a simple vector representing the "average" state of each evolutionary parameter for the complete set of vertebrate species considered and is designed to facilitate inter barcode comparisons and clustering. The 1D-barcode values are produced by calculating phylum-weighted means: (i) for each parameter, a mean is calculated for 4 phyla: mammals, sauropsida, amphibians and teleostei, (ii) these phylum means are used to calculate a new mean that is the final value for a specific parameter of the 1D-barcode. As in the 2D-barcode, a color is assigned to each 1D-barcode parameter value based on the sample percentiles, for visualization purposes. However, in contrast to the 2D-barcodes, these percentiles are not organism related. They are based on the phylum weighted mean parameter values from the complete set of 1D-barcodes.

Barcode clustering and GO enrichment analysis

The complete set of 1D-barcodes representing the human proteome were used for the clustering analysis, although barcodes with missing values were removed from the test set, leaving a total of 19465 barcodes. Each 1D-barcode was represented by a vector of real values, $X = (x_1, x_2, \ldots x_n)$ and the distance, $d(X,Y)$ between two barcodes was defined as:

$$d(X,Y) = \sqrt[2]{\sum_{i=0}^{n} (x_i - y_i)^2}$$

The distance between each pair of barcodes was calculated and the complete pairwise distance matrix as used as input to a clustering program that implements an improved Potts clustering model.[36] The Potts clustering approach, also known as super-paramagnetic clustering, is based on the physical behavior of an inhomogeneous ferromagnet.[37] No assumptions are made about the underlying distribution of the data. Briefly, a Potts spin variable is assigned to each data point and short range interactions between neighboring points are introduced. Spin-spin correlations are measured by a Monte Carlo procedure and are used to partition the data points into clusters.

The GoMiner software[38] was then used to analyze the GO enrichment of the resulting barcode clusters. The complete set of human reference sequences was used as a background gene list. As stated by the GoMiner authors, the calculated P-values should be considered as heuristic measures, useful as indicators of possible statistical significance, rather than as the results of formal inference. The P-values can be used, for example, to sort categories to identify those of the most potential interest. In this work, a cluster was considered to be enriched in a GO term if the associated P-value was <0.05, the recommended value for high-throughput GoMiner. We then sorted the clusters according to their mean P-values and selected several top ranking clusters for further manual analysis.

Barcode website

All the data presented in this publication are available online at the following address: http://lbgi.igbmc.fr/barcodes. The website interface allows the user to browse all the human barcodes, as well as the annotated multiple alignments corresponding to each barcode. Barcodes can be selected by textual searches with Uniprot and Ensembl identifiers or by uploading a Fasta sequence followed by a BlastP search. The results of two high throughput analyses are also available: the mapping of all the 1D-barcodes on the human chromosomes and the clustering of the 1D-barcodes generated by the Potts model.

Results and Discussion
Design of the barcode

The objective of the EvoluCode evolutionary barcode is to integrate heterogeneous biological data from different biological levels in order to highlight new evolutionary patterns or scenarios that could not be detected using only one kind of data (genomic context data, sequence data, expression data …).

In this study, we applied the barcode formalism to the human proteome to study vertebrate evolution. This barcode (described in detail below) includes data from 17 vertebrate species and 10 evolutionary parameters, representing different biological levels, from the genomic level (synteny) to the clade level (number of co-orthologs). Nevertheless, the barcode can theoretically be of any dimension N × n, with a parameter and species composition depending on the objectives or evolutionary scale (eg, primates, vertebrates, eukaryotes…) of the study.

The barcode combines both continuous parameters, such as sequence conservation or hydrophobicity, and discontinuous parameters, such as local synteny conservation or domain organization. Since the different parameters have very heterogeneous distributions (multi-modal, exponential, normal distribution…) they cannot be described using a single statistical model. We therefore developed a methodology to normalize the values of any given parameter using simple percentile statistics, which are suitable for any kind of parameter distribution. For visualization purposes, the normalized parameters are color-coded to highlight values that are inferior or superior to what is generally observed in a given species.

In order to summarize the diverse data inherent to the 2D-barcode approach, each barcode can also be represented in 1D. The 1D-barcode is thus a vector of continuous values representing the phylum-weighted average state of each evolutionary parameter. In the case of the human proteome barcodes, the 1D-barcode represents the average values observed during the vertebrate evolutionary history. As in the 2D-barcode, the parameters are color-coded to highlight the "expectedness" of a particular value.

Representation of complex evolutionary histories: the human proteome

To demonstrate the applicability of the EvoluCode formalism, we constructed barcodes to represent the evolutionary histories of the complete human proteome since the appearance of the vertebrates. Thus, for 19778 human genes, a representative reference protein was selected and homologs were identified in 16 complete genomes of vertebrate organisms (see Material and Methods). We then constructed 19778 multiple sequence alignments that were annotated with known structural and functional

information. In addition, we estimated the synteny between the 19778 human genes and the 16 vertebrate genomes. Finally, orthologous relationships between human and the 16 vertebrates were inferred. Based on these data, we extracted various evolutionary parameters, representing primary sequence characteristics, domain organization, phylogenetic distribution and genome neighborhood conservation. These parameters were then integrated to form an evolutionary barcode representing each human reference protein. Some typical examples of barcodes, representing genes with heterogeneous and complex evolutionary histories, are shown in Figure 2 and described in detail below.

The first example (Fig. 2A) corresponds to the glucagon receptor (reference protein GLR_HUMAN). This receptor is essential for blood glucose level regulation, an essential function for all vascular animals.[39] For all parameters; the 2D-barcode displays homogenous states over all vertebrates, implying that relatively few genetic events have affected this gene during vertebrate evolution.

The second example (Fig. 2B) corresponds to the barcode of a gene integrated from an endogenous retrovirus (reference protein POK12_HUMAN). In our barcode construction procedure, the human gene was associated with genes from the other vertebrate species that have also integrated endogenous retrovirus genes, characterized by specific sequence motifs. Consequently, the phylogenetic distribution of this barcode is dispersed. Moreover, these genes generally produce polyprotein products, explaining the heterogeneity observed for the number of domains and the fact that these sequences are not detected as orthologs.

The third example (Fig. 2C) represents a gene specific to the rodent and primate lineages (reference protein DPPA3_HUMAN). This gene appeared recently in the mammalian lineage and was previously characterized as playing a role in developmental cell pluripotency and in adult sexual organs.[40] The protein product of this gene has several unusual characteristics. Despite its recent evolutionary history, it has very low sequence conservation, with 78% percent identity between human and macaque and only 37% between human and mouse. This is supported by heterogeneous hydrophobicity scores in the different species. Such rapid divergence for reproductive proteins is a well-known phenomena.[41]
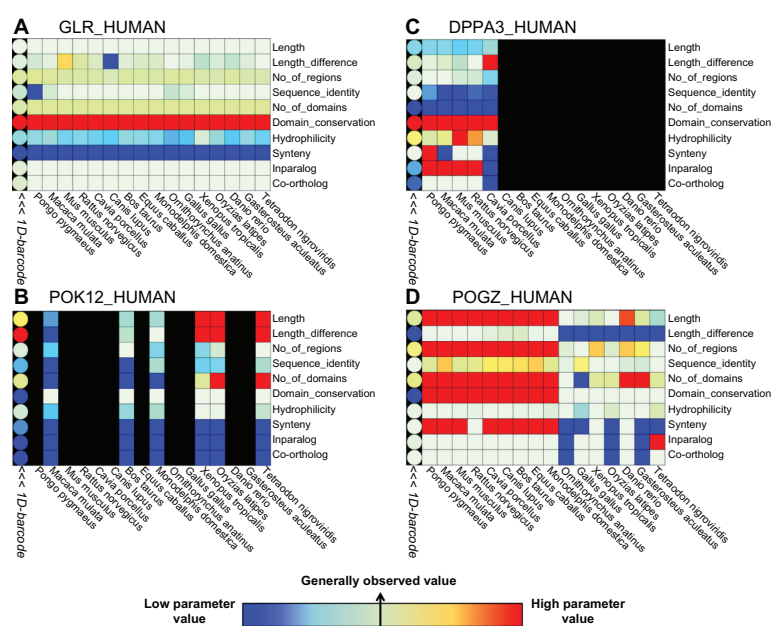
Figure 2. Four examples of 2D-barcodes (square cells) are shown. Rows represent evolutionary parameters and columns represent vertebrate species. A color gradient highlights parameters having respectively, values lower than what is generally observed (blue), generally observed values (green), values higher than what is generally observed (red).The 1D-barcode is shown on the left hand side (round cells). Each barcode is associated with one human protein: (**A**) the glucagon receptor (GLR_HUMAN), (**B**) the HERV-K_1q22 provirus ancestral Pol protein (POK12_HUMAN), (**C**) the developmental pluripotency-associated protein 3 (DPPA3_HUMAN), and (**D**) the Pogo transposable element with ZNF domain (POGZ_HUMAN).

The last example (Fig. 2D) illustrates the ability of our multi-level barcode approach to highlight a potential genetic event. The 'Pogo transposable element with ZNF domain' gene (reference protein POK12_HUMAN) is involved in kinetochore assembly.[42] The genetic event highlighted by the 2D-barcode occurred just after the separation of the theria and prototheria lineages. Two different blocks can be distinguished in the 2D-barcode of POGZ_HUMAN. The first block includes all theria and for these species, the gene is characterized by long sequences with conserved synteny and one ortholog in each species. The second block is less homogeneous, characterized by shorter sequences with fewer domains and low percent identities compared to human. The barcode thus suggests a potential domain gain for this gene in the marsupial and placental mammal lineages. This genetic event is particularly interesting because it occurred in a gene implicated in a fundamental process (mitosis) but indicates recent mammalian innovation in this process.[42]

These examples illustrate the wide range of information that can be extracted using the barcode formalism. By visualizing the evolutionary histories of the different proteins in the form of 2D-barcodes, general evolutionary trends can be observed and

specific evolutionary events such as genetic events can be easily identified. The following sections will describe some large-scale analyses of the complete set of barcodes representing the evolutionary histories of the human proteome.

Large scale visualization of evolutionary barcodes

Although the 2D-barcode is a useful tool for visualizing the evolutionary histories of a small number of genes, it is too complex for large-scale visualization. To address this issue, we designed a 1 dimensional version of the evolutionary barcode, called the 1D-barcode. To estimate whether these 1D-barcodes can usefully represent global evolutionary histories, we mapped the human proteome 1D-barcodes to the 24 human chromosomes, resulting in a barcode map of the complete genome.

The visual inspection of this map allowed us to distinguish several previously published gene clusters. One example is the case of the keratin I and keratin II gene clusters. Early chordates had one keratin I gene and one keratin II gene.[43] During vertebrate evolution, these genes evolved to form gene clusters with evidence of cluster expansion from amphibia and birds to mammals.[44] A second gene

family appeared during mammalian evolution and separates the type I KR chromosomal cluster in two parts. This family contains keratin associated proteins (KRAP) and represents one of the major components of hair, playing essential roles in the formation of rigid and resistant hair shafts.[45] Figure 3 shows the consecutive 1D-barcodes corresponding to the human type I keratin (KR) cluster and highlights different evolutionary histories. The older KRs are the cytokeratins, which are present in the amphibian and bird KR clusters. The number of human inparalogs and the number of co-orthologs in other species have higher values (shown in red) for these cytokeratins compared to the values observed in other human genes. In particular, the number of human inparalogs is relatively high compared to the other vertebrate species, indicating that numerous duplications occurred after the cytokeratin duplications in early vertebrates. Interestingly, the values of these parameters are much lower for hair KR and inner root sheath KR, implying that these genes duplicated more recently. The KRAP cluster splitting the keratin cluster in two parts has very different barcode profiles. The unusual values of the corresponding 1D-barcodes suggest original evolutionary histories. Indeed, the values of the synteny, inparalog, co-ortholog and sequence conservation parameters are low, indicating a gene family that appeared recently with high variability between the species. In fact,

these genes are specific to mammals and have evolved and diverged rapidly.[45] Thus, this example illustrates the ability of the 1D-barcodes to identify local chromosomal regions that have experienced similar evolutionary histories. Such an approach could be used in the future to identify other chromosomal features, for example evolutionary breakpoints.[46]

Genome-level clustering of evolutionary histories

The goal of this analysis was to identify subsets of genes in the full set of 19778 human genes that share similar barcodes, ie, similar evolutionary histories. To achieve this, we defined a Euclidean distance metric between any two barcodes based on the phylum-weighted mean values of each evolutionary parameter in the 1D-barcode. Since no a priori assumptions can be made about the statistical models underlying the parameter value distributions, we used a clustering algorithm based on nonparametric techniques: the Potts clustering model, also known as super-paramagnetic clustering. The Potts model was first developed for physical systems,[47] then recently adapted for clustering purposes in neuroscience and bioinformatics.[48–52] The advantage of this technique is that the user does not need to specify the number of clusters required, because this number is estimated in a probabilistic framework. In particular, we used an
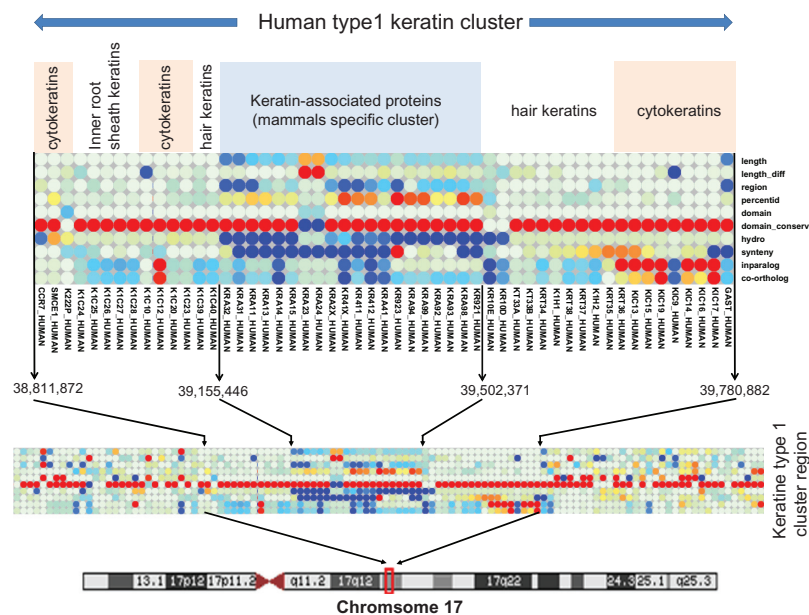
Figure 3. The 1D-barcodes corresponding to the human type I keratin cluster.
Notes: Each column represents one 1D-barcode of one protein. Several keratin subfamilies are delimited by white vertical lines. The boundaries of the keratin cluster are delimited by black arrows.

improved version of this clustering technique called Conditional-Potts Clustering Model.[53] This model is based on an improved Potts clustering model[37] with an additional prior estimation of the most suitable parameters for an efficient clustering. Using the Potts clustering model, 303 clusters were generated with a maximum cluster size of 380 proteins.

To investigate the potential functional significance of these barcode clusters, we performed a GO enrichment analysis of the 303 generated clusters using the GoMiner software.[38] Figure 4 shows the distribution of the mean enrichment P-values obtained by considering all GO terms with a P-value <0.05 (the lower the P-value, the better the enrichment). Most clusters are enriched in at least one GO term, with 75% of the clusters having mean P-values <0.025 and 98% of the clusters having mean P-values <0.03. Several examples of the most enriched clusters are described in Table 1 and some of these clusters are clearly related to specific gene families. One striking example is the cluster 15, which groups numerous olfactory receptors. The family of olfactory receptors experienced a vast expansion during the chordate evolution, with the number of olfactory receptors ranging from a dozen in fishes to over a thousand in rodents.[54] Moreover, pseudogenization and decline of olfactory functions has occurred in some lineages and it is thought that half of all primate receptor genes may be pseudogenes.[55] The evolutionary history of this family is characterized by barcodes
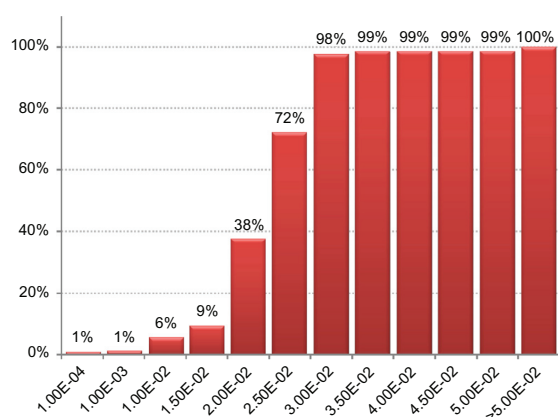
with high hydrophobicity scores, high domain conservation and a variable number of co-orthologs in mammalian species. Interestingly, some keratin-associated proteins, implicated in hair development were clustered together with the olfactory receptors, possibly reflecting their similar, recent expansion during mammalian evolution. Other enriched clusters correspond to highly conserved systems in vertebrates. For example, cluster 46 is enriched in genes linked to the mitochondrial respiratory chain. Similarly, clusters 67 and 153 are enriched in genes linked to translation and mRNA splicing respectively. Interestingly, the barcodes associated with these two clusters are mainly differentiated by the synteny conservation. The synteny tends to be conserved for genes linked to mRNA splicing complexes, but not for the genes involved in translation.

In this example analysis, we have studied the functional significance of the barcode clusters, based on GO term enrichment. In the future, we also plan to investigate the correlations between the barcode clusters and other functional data, including gene expression profiles, interactomic data and biological networks.

Multi-dimensional analysis highlights new evolutionary trends

To further illustrate the power of the multi-level barcode analyses, we analyzed the barcodes corresponding to multi-pass membrane proteins. These proteins have strong physico-chemical constraints with a predominant conservation of hydrophobic residues in their alpha helix compared to soluble proteins.[56] We extracted from our sequence dataset, the 2674 human proteins that are annotated as "Multi-pass membrane protein" in Uniprot (Uniprot search engine keywords: "location: SL-9909"). In this protein subset study, we wanted to investigate in more detail the contributions of each of the individual parameters to the clustering process. We therefore performed a Multiple Correspondence Analysis (MCA) clustering of the 1D-barcodes, using the FactoMineR R package.[57] This package provides visualization tools to display the clustering results. In particular, we can clearly illustrate the correlations between the barcode parameters and the inferred barcode clusters.

Using the 2674 "multi-pass membrane protein" barcodes, the MCA clustering produced 4 barcode

Figure 4. Percentage of clusters with a mean GO term enrichment *P*-value below a given threshold.
Note: We calculated the mean *P*-value of all GO terms having a *P*-value <0.05. 98% of the clusters have at least one enriched GO term and a mean *P*-value $<3.00\text{E-}2$, indicating potential biological meaning for most clusters.

Table 1. Some examples of barcode clusters with high GO enrichment. The most enriched terms for each cluster are shown with their corresponding *P*-value (10log(p)) and false discovery rate (FDR). The lower the *P*-value and FDR, the better is the enrichment.

Cluster id	Representative sequence	Go accession	Go terms	10log(p)	FDR
46	NDUA7_HUMAN	GO:0022904	repiratory electron transport chain	−10.894378	0
		GO:0006796	phosphate metabolic process	−5.162176	0.003
15	OR2L5_HUMAN	GO:0007608	sensory perception of smell	−69.573133	0
		GO:0007606	sensory perception of chemical stimulus	−66.771345	0
		GO:0007186	G-protein coupled receptor protein signaling pathway	−55.368505	0
95	D104A_HUMAN	GO:0042742	defense response to bacterium	−10.156822	0
		GO:0009607	response to biotic stimulus	−5.232461	0
		GO:0006950	response to stress	−4.145167	0.018
207	MYH3_HUMAN	GO:0030029	actin filament-based process	−8.190798	0
		GO:0007265	Ras protein signal transduction	−3.375746	0.015
		GO:0014065	phosphoinsitide 3-kinase cascade	−2.923239	0.031
67	TF2H2_HUMAN	GO:0006414	translational elongation	−14.67022	0
		GO:0042273	ribosomal large subunit biogenesis	−5.260087	0
		GO:0016072	rRNA metabolic process	−4.21555	0
153	RL15_HUMAN	GO:0044260	cellular macromolecule metabolic process	−8.336618	0
		GO:0000398	nuclear mRNA splicing via spliceosome	−6.719139	0
		GO:0006807	nitrogen compound metabolic process	−5.889665	0

clusters, as shown in Figure 5. The first axis represents parameters linked to the evolutionary history, while the second axis is linked to sequence characteristics. Details of the cluster compositions are provided in supplementary Table 2. All 4 clusters contain similar numbers of barcodes, respectively: 30.3%, 23.7%, 26.6% and 19.4%. Clusters 1, 3 and 4 correspond to three different barcode profiles and are described in detail below. Cluster 2 contains barcodes that are intermediates between clusters 1, 3 and 6.
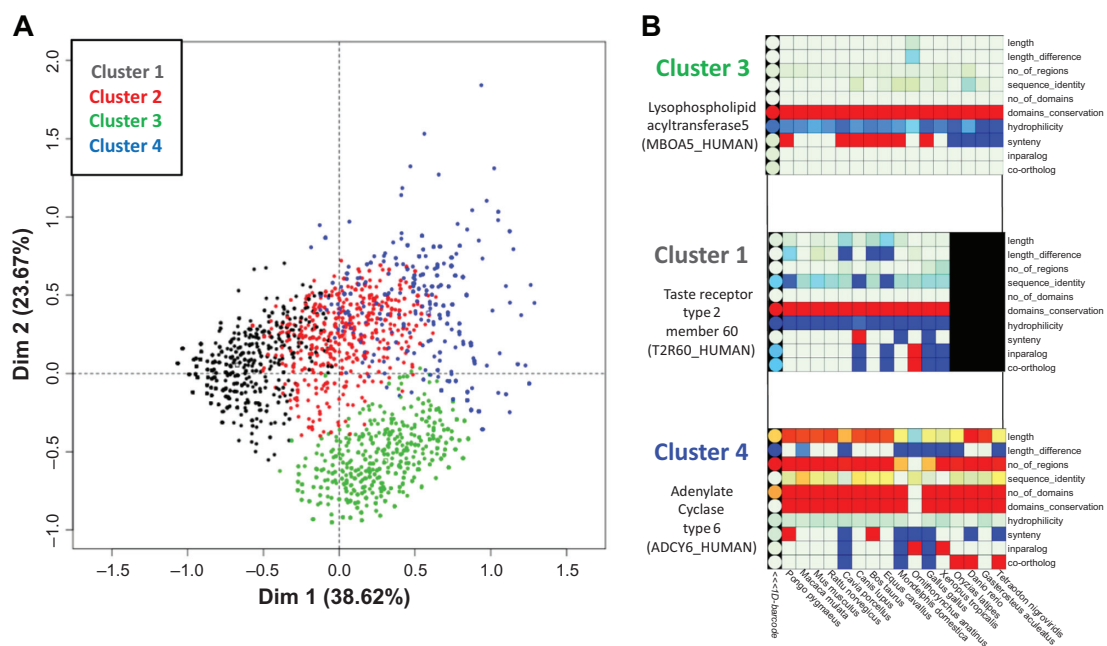
Figure 5. (**A**) MCA clustering of 2674 human membrane multi-pass proteins. (**B**) Representative barcodes for 3 clusters are shown to illustrate the major differences between the barcodes in each cluster.
Note: Each dot represents one 1D-barcode.

- Cluster 1 (black) contains 30% of the 2674 integral membrane proteins and corresponds to proteins with short sequences and low hydrophilicity. From an evolutionary point of view, they are less well conserved, with early mammals, sauropsida and fish often sharing as little as 50% sequence identity. Their phylogenetic distribution is very heterogeneous, with gene gains and losses in many phyla, represented by a wide range of values for the inparalog and co-ortholog parameters. A large proportion (55%) of this cluster is composed of G-protein coupled receptors (GPCRs), mainly olfactory and taste receptors.

- Cluster 3 (green) contains 27% of the proteins and is the most homogeneous cluster. It groups barcodes with the number of domains of conserved regions, conserved synteny in most mammals and a single ortholog in most vertebrate species. Thus, the cluster corresponds mainly to genes that are highly conserved in vertebrates with fewer genetic events compared to other multi-pass membrane proteins. To investigate the potential functional significance of this cluster, we mapped the corresponding genes to the KEGG pathway database.[58] This analysis linked 41% of the 293 mapped proteins to basal metabolic processes and neural processes (eg, hsa01100-Metabolic systems, hsa04080-Neuroactive ligand-receptor interaction).

- Cluster 4 (blue) contains 19% of the proteins and represents a wider distribution of barcodes. It contains average to long sequences, with numerous conserved regions. The associated proteins are not necessarily conserved in vertebrates (heterogeneous sequence identity between barcodes in the cluster), but generally have lower hydrophobicity than the other multi-pass membrane proteins. In fact, the cluster contains many proteins with multiple intra/extracellular regions, which are more conserved and hydrophilic than the hydrophobic α-helix transmembrane regions. Interestingly, 29% of cluster 4 proteins map to KEGG pathways involved in secretion processes (eg, hsa04724-Glutamatergic synapse; hsa04972-Pancreatic secretion; hsa04976-Bile secretion; hsa04970-Salivary secretion; hsa02010-ABC transporters).

This in-depth analysis of the barcodes corresponding to multi-pass membrane proteins identified important evolutionary trends and their correlations with protein function. For example, the proteins in cluster 3 have evolved little during vertebrate evolution and are mostly involved in essential processes, such as metabolic or neural processes. In contrast, cluster 1 highlights a subset of integral membrane protein families, such as GPCRs, that have experienced more genetic events. Interestingly, such behavior seems to be correlated with shorter, more hydrophobic sequences containing few intra/extracellular regions. Thus, membrane proteins that have fewer extramembrane regions are observed to be more divergent. This seems to contradict previous studies indicating that the transmembrane regions of membrane proteins are highly constrained and diverge at slower rates than the extramembrane regions.[56]

EvoluCode in systems biology: a proof of concept

Systems biology aims to analyze genes and proteins in the context of their biological networks. As a proof of concept, we mapped our evolutionary barcodes to the KEGG pathway corresponding to the cysteine and methionine metabolism (hsa00270), in order to identify branches or 'hot spots' having particular evolutionary behaviors. Figure 6 shows the human methionine salvage sub-pathway, involving 13 human proteins. This sub-pathway is found in many phyla, such as plants, fungi, mammals, and bacteria (for a review, see Albers, 2009). We then calculated a normalized Euclidean distance between each pair of barcodes and constructed a neighbor-joining tree from the resulting distance matrix (Fig. 6A). This distance between barcodes represents the differences between the corresponding protein evolutionary histories and takes into account, not only sequence similarity, but also other factors, such as domain conservation, gene duplicability and genome context. In the context of the methionine salvage pathway, two barcodes corresponding to the *adi*1 and *il4i*1 genes are relatively distant compared to the other barcodes of this metabolic pathway.

First, the ADI1 protein (MTND_HUMAN) is an acireductone dioxygenase. Depending on the ion used as a cofactor, Fe2+ or Ni2+, this enzyme performs different reactions, introducing an "off-pathway" branching.[59] Its barcode demonstrates very high hydrophilicity and short sequences for all species,
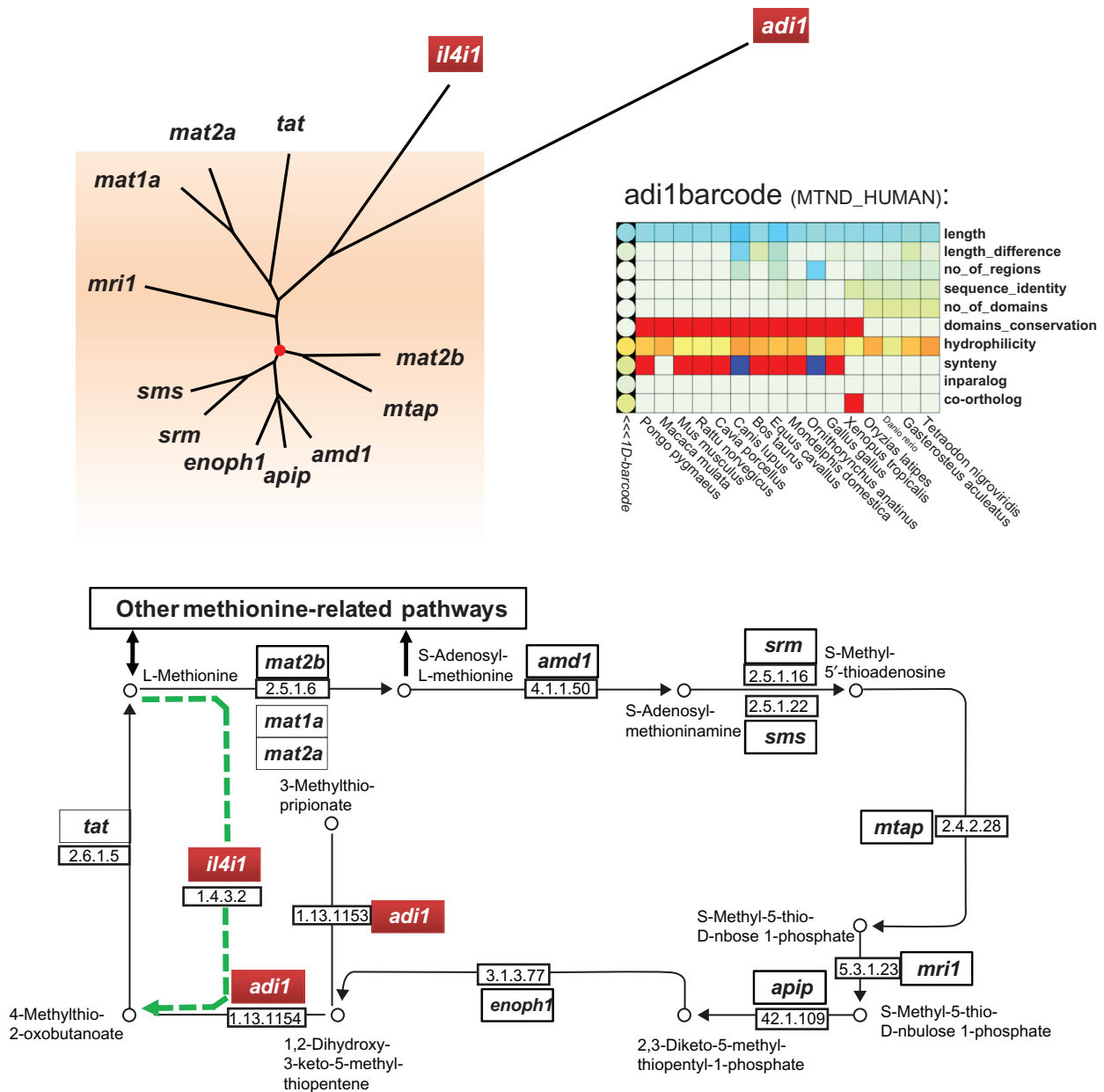
Figure 6. (**A**) Neighbor-joining tree of barcodes corresponding to genes in the KEGG human methionine salvage sub-pathway (hsa00720). The root of the tree is indicated by a red circle. The most distant barcodes from the root are shown in red boxes. (**B**) KEGG sub-pathway map, highlighting the positions of the genes corresponding to the most distant barcodes.

but a variable number of conserved regions and an additional domain in the fish lineage. Interestingly, this enzyme is also implicated in several other processes: the compound produced by this enzyme can cause apoptosis[60] and the *adi*1 gene has been implicated in prostate cancers.[61] Thus, it not only generates a new branch in the methionine salvage pathway, but it is also involved in other pathways. These interactions can lead to different evolutionary constraints compared to the other genes implicated in the "canonical"

methionine salvage pathway, which might explain its position as an outlier in this analysis.

Second, the IL4I1 protein (OXLA_HUMAN) is an L-amino acid oxidase (LAO). Despite its presence in the KEGG methionine salvage pathway, this protein is mainly expressed in immune defenses of vertebrates and mollusks, in particular in immune system cells and B-cell lymphomas.[62] As IL4I1 is not directly implicated in the basal metabolic processes, it is not surprising that the corresponding barcode is seen

as an outlier. Moreover, a recent study have shown that the LAO families have undergone repeated duplications and deletions.[63] This study supported the hypothesis that IL4I1 and the ancestor of LAO1 and LAO2 arose from an ancient duplication prior to the origin of tetrapods and that IL4I1 was lost in many non-mammalian tetrapods, whereas LAO1 and LAO2 were lost in mouse and human. This evolutionary pattern is in fact characteristic of many families involved in vertebrate immune processes.[64]

The mapping of the barcodes on the methionine salvage sub-pathway demonstrates their ability to highlight unusual evolutionary patterns, not only related to genomic data, but also to concepts such as centrality in networks or patterns of expression. Interestingly, both outlier barcodes are located in non linear parts of the pathway. Such correlation might indicate different evolutionary constraints for multi-connected pathway nodes. However, this hypothesis will require further investigation. In particular, the identification of such patterns currently requires human expert analysis. Further developments will be needed to automate the process, involving high throughput comparison of the evolutionary barcodes with network and expression data, as well as rigorous mathematical analyses to identify breakpoints and barcode outliers.

Conclusions and Perspectives

The EvoluCode barcode formalism is a powerful tool for the visualization and quantitative analysis of complex evolutionary histories in high throughput studies. Three major advantages are: (i) diverse parameters from different biological levels can be combined in a unifying framework, (ii) the parameter set can be easily modified, facilitating the construction of different barcodes for different purposes, (iii) the parameter values are normalized based on their specific distributions to allow direct comparisons within and between barcodes and to facilitate the rapid identification of typical/atypical values by the user.

We have constructed barcodes representing the evolutionary histories of the complete human proteome. The analysis was restricted to the vertebrate evolutionary scale to ensure the production of high quality multiple alignments, from which several barcode parameters are extracted. Although in principle, the barcode could be applied to higher evolutionary

scales (eg, metazoa, eukaryotes …), such an extension would require more robust protocols to evaluate and validate the quality of the alignments.

One critical question that had to be addressed during the design was the selection of pertinent evolutionary parameters. The human proteome barcodes incorporate various multilevel parameters from 17 vertebrate organisms, covering genomic context, primary sequence characteristics, sequence/domain conservation and phylogenetic distributions. However, both the species set and parameter set can be easily adapted to the goals of a specific study. The data mining technique used for the subsequent analysis of the barcodes may also influence the choice of parameters to include. For example, some methods may be sensitive to highly correlated parameters, and a correspondence analysis (CA) may be necessary to select a subset of parameters with low dependency.

The combination of heterogeneous parameters is able to highlight more original and complex evolutionary trends, which could not be detected based on a single parameter such as sequence conservation or orthology. We have demonstrated this in two large scale analyses: chromosome mapping and clustering. However, the EvoluCode formalism opens the way to the application of a wide range of standard data mining or machine learning techniques that have not been possible in evolutionary studies. To illustrate the potential of EvoluCode barcodes in systems biology studies, we described the analysis of a small metabolic pathway. This proof of concept provides the basis for future studies. The automation of such analyses at the scale of all pathways in an organism should provide valuable information for pathway evolution analysis. In particular, the ability to calculate distances between barcodes will allow us to estimate parameters such as pathway "evolutionary rates" and to highlight rapidly evolving sub-pathways.

Future developments will include on the study of other distance metrics, in addition to the Euclidean distance used here. In particular, we will use the Pearson correlation coefficient to estimate the linear dependency between the barcode parameters. This would lead to a barcode clustering based on relative changes in the parameter values, rather than their scale. We will also apply more rigorous mathematical theories to identify outlying parameter values, as well as shifts or breakpoints in the barcode behavior.

For example, a formal description of the different blocks in the barcode corresponding to POGZ_HUMAN (Fig. 2) could be a first step towards automatically detecting genetic events. Similarly, the stochastic or heterogeneous nature of a given barcode could be estimated based on the frequency of parameter state changes in the different phyla. This could lead to the development of quantitative indicators of the rate of evolution for a particular gene, facilitating the automatic identification of "original" evolutionary scenarios and signatures of adaptation or innovation. The analysis of the proteome is thus expected to shed more light on the fundamental aspects of the evolutionary processes and the factors that shape contemporary vertebrate genomes.

In the longer term, the methodologies developed here should facilitate, not only the analysis of proteomes from other species, but also the efficient exploitation of evolutionary information in functional genomics (notably, in interactomics and transcriptomics comparisons or in high throughput promoter studies) and large scale systems biology projects.

Acknowledgements

Disclosures

Author(s) have provided signed confirmations to the publisher of their compliance with all applicable legal and ethical obligations in respect to declaration of conflicts of interest, funding, authorship and contributorship, and compliance with ethical requirements in respect to treatment of human and animal test subjects. If this article contains identifiable human subject(s) author(s) were required to supply signed patient consent prior to publication. Author(s) have confirmed that the published article is unique and not under consideration nor published by any other publication and that they have consent to reproduce any copyrighted material. The peer reviewers declared no conflicts of interest.

References

1. Kitano H. Systems biology: a brief overview. *Science*. Mar 1, 2002;295 (5560):1662–4.
2. Snoep JL, Bruggeman F, Olivier BG, Westerhoff HV. Towards building the silicon cell: a modular approach. *Biosystems*. Feb–Mar 2006;83(2–3): 207–16.
3. Kohl P, Noble D. Systems biology and the virtual physiological human. *Molecular Systems Biology*. Jul 2009;5.
4. Hoehndorf R, Dumontier M, Gennari JH, et al. Integrating systems biology models and biomedical ontologies. *Bmc Systems Biology*. Aug 11, 2011;5.
5. Loewe L. A framework for evolutionary systems biology. *Bmc Systems Biology*. 2009;3:27.
6. Medina M. Genomes, phylogeny, and evolutionary systems biology. *Proc Natl Acad Sci U S A*. May 3, 2005;102(Suppl 1):6630–5.
7. Huerta-Cepas J, Capella-Gutierrez S, Pryszcz LP, et al. PhylomeDB v3.0: an expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. *Nucleic Acids Res*. Jan 2011;39(Database issue):D556–60.
8. Cork JM, Purugganan MD. The evolution of molecular genetic pathways and networks. *Bioessays*. May 2004;26(5):479–84.
9. Medina M. Genomes, phylogeny, and evolutionary systems biology. *Proceedings of the National Academy of Sciences of the United States of America*. May 3, 2005;102:6630–5.
10. Yamada T, Bork P. Evolution of biomolecular networks—lessons from metabolic and protein interactions. *Nature Reviews Molecular Cell Biology*. Nov 2009;10(11):791–803.
11. Lercher MJ, Pal C, Papp B. An integrated view of protein evolution. *Nature Reviews Genetics*. May 2006;7(5):337–48.
12. Knight CG, Pinney JW. Making the right connections: biological networks in the light of evolution. *Bioessays*. Oct 2009;31(10):1080–90.
13. Koonin EV, Wolf YI. Evolutionary systems biology: links between gene evolution and function. *Current Opinion in Biotechnology*. Oct 2006;17(5):481–7.
14. Herbeck JT, Wall DP. Converging on a general model of protein evolution. *Trends Biotechnol*. Oct 2005;23(10):485–7.
15. Jeong H, Mason SP, Barabasi AL, Oltvai ZN. Lethality and centrality in protein networks. *Nature*. May 3, 2001;411(6833):41–2.
16. Liang H, Li WH. Gene essentiality, gene duplicability and protein connectivity in human and mouse. *Trends in Genetics*. Aug 2007;23(8):375–8.
17. Park SG, Choi SS. Expression breadth and expression abundance behave differently in correlations with evolutionary rates. *BMC Evol Biol*. 2010;10:241.
18. Wolf YI, Carmel L, Koonin EV. Unifying measures of gene function and evolution. *Proc Biol Sci*. Jun 22, 2006;273(1593):1507–15.
19. Waterhouse RM, Zdobnov EM, Kriventseva EV. Correlating traits of gene retention, sequence divergence, duplicability and essentiality in vertebrates, arthropods, and fungi. *Genome Biol Evol*. Jan 2011;3:75–86.
20. Procter JB, Thompson J, Letunic I, Creevey C, Jossinet F, Barton GJ. Visualization of multiple alignments, phylogenies and gene family evolution. *Nat Methods*. Mar 2010;7(3 Suppl):S16–25.
21. O'Donovan C, Apweiler R, Bairoch A. The human proteomics initiative (HPI). *Trends Biotechnol*. May 2001;19(5):178–81.
22. Boutet E, Lieberherr D, Tognolli M, Schneider M, Bairoch A. UniProtKB/Swiss-Prot. *Methods Mol Biol*. 2007;406:89–112.
23. Hubbard TJ, Aken BL, Ayling S, et al. Ensembl 2009. *Nucleic Acids Research*. Jan 2009;37(Database issue):D690–7.
24. McGinnis S, Madden TL. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res*. Jul 1, 2004;32(Web Server issue):W20–5.
25. Plewniak F, Bianchetti L, Brelivet Y, et al. PipeAlign: A new toolkit for protein family analysis. *Nucleic Acids Research*. Jul 1, 2003;31(13):3829–32.

26. Thompson JD, Plewniak F, Thierry J, Poch O. DbClustal: rapid and reliable global multiple alignments of protein sequences detected by database searches. *Nucleic Acids Research*. Aug 1, 2000;28(15):2919–26.

27. Thompson JD, Thierry JC, Poch O. RASCAL: rapid scanning and correction of multiple sequence alignments. *Bioinformatics*. June 12, 2003;19(9):1155–61.

28. Thompson JD, Prigent V, Poch O. LEON: multiple aLignment Evaluation of Neighbours. *Nucleic Acids Research*. 2004;32(4):1298–307.

29. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*. Jul 15, 2002;30(14):3059–66.

30. Thompson JD, Muller A, Waterhouse A, et al. MACSIMS: multiple alignment of complete sequences information management system. *BMC Bioinformatics*. 2006;7:318.

31. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*. May 2000;25(1):25–9.

32. Finn RD, Mistry J, Tate J, et al. The Pfam protein families database. *Nucleic Acids Research*. Jan 2010;38(Database issue):D211–22.

33. Koonin EV. Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet*. 2005;39:309–38.

34. Linard B, Thompson JD, Poch O, Lecompte O. OrthoInspector: comprehensive orthology analysis and visual exploration. *BMC Bioinformatics*. 2011;12:11.

35. Emerson JD, Strenio J. Boxplots and batch comparison. (Editors. *D. C. Hoaglin, F. Mosteller and J. W. Tukey). Wiley, New York*. 1983:pp. 58–96 in Understanding Robust and Exploratory Data Analysis.

36. Alejandro M, Nicolas W. The Conditional-Potts Clustering Model.

37. Murua A, Stanberry L, Stuetzle W. On Potts model clustering, kernel K-means, and density estimation. *Journal of Computational and Graphical Statistics*. Sep 2008;17(3):629–58.

38. Zeeberg BR, Feng W, Wang G, et al. GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol*. 2003;4(4):R28.

39. Lok S, Kuijper JL, Jelinek LJ, et al. The human glucagon receptor encoding gene: structure, cDNA sequence and chromosomal localization. *Gene*. Mar 25, 1994;140(2):203–9.

40. Clark AT, Rodriguez RT, Bodnar MS, et al. Human STELLAR, NANOG, and GDF3 genes are expressed in pluripotent cells and map to chromosome 12p13, a hotspot for teratocarcinoma. *Stem Cells*. 2004;22(2):169–79.

41. Swanson WJ, Vacquier VD. The rapid evolution of reproductive proteins. *Nat Rev Genet*. Feb 2002;3(2):137–44.

42. Nozawa RS, Nagao K, Masuda HT, et al. Human POGZ modulates dissociation of HP1alpha from mitotic chromosome arms through Aurora B activation. *Nat Cell Biol*. Jul 2010;12(7):719–27.

43. Karabinos A, Zimek A, Weber K. The genome of the early chordate Ciona intestinalis encodes only five cytoplasmic intermediate filament proteins including a single type I and type II keratin and a unique IF-annexin fusion protein. *Gene*. Feb 4, 2004;326:123–9.

44. Zimek A, Weber K. Terrestrial vertebrates have two keratin gene clusters; striking differences in teleost fish. *European Journal of Cell Biology*. Jun 2005;84(6):623–35.

45. Wu DD, Irwin DM, Zhang YP. Molecular evolution of the keratin associated protein gene family in mammals, role in the evolution of mammalian hair. *BMC Evol Biol*. 2008;8:241.

46. Veron AS, Lemaitre C, Gautier C, Lacroix V, Sagot MF. Close 3D proximity of evolutionary breakpoints argues for the notion of spatial synteny. *BMC Genomics*. 2011;12:303.

47. Blatt M, Wiseman S, Domany E. Superparamagnetic clustering of data. *Phys Rev Lett*. Apr 29, 1996;76(18):3251–4.

48. Stanberry L, Murua A, Cordes D. Functional connectivity mapping using the ferromagnetic Potts spin model. *Human Brain Mapping*. Apr 2008;29(4):422–40.

49. Getz G, Levine E, Domany E, Zhang MQ. Super-paramagnetic clustering of yeast gene expression profiles. *Physica A*. May 1, 2000;279(1–4):457–64.

50. Einav U, Tabach Y, Getz G, et al. Gene expression analysis reveals a strong signature of an interferon-induced pathway in childhood lymphoblastic leukemia as well as in breast and ovarian cancer. *Oncogene*. Sep 22, 2005;24(42):6367–75.

51. Radjiman S, Han LY, Wang JS, Chen YZ. Super paramagnetic clustering of DNA sequences. *Journal of Biological Physics*. Jan 2006;32(1):11–25.

52. Tetko IV, Facius A, Ruepp A, Mewes HW. Super paramagnetic clustering of protein sequences. *BMC Bioinformatics*. Apr 1, 2005;6.

53. Murua A, Wicker N. The Conditional-Potts Clustering Model. *(submitted)*. 2011.

54. Zhang X, Firestein S. Genomics of olfactory receptors. *Results Probl Cell Differ*. 2009;47:25–36.

55. Kambere MB, Lane RP. Co-regulation of a large and rapidly evolving repertoire of odorant receptor genes. *BMC Neurosci*. 2007;8 (Suppl 3):S2.

56. Oberai A, Joh NH, Pettit FK, Bowie JU. Structural imperatives impose diverse evolutionary constraints on helical membrane proteins. *Proc Natl Acad Sci U S A*. Oct 20, 2009;106(42):17747–50.

57. Le S, Josse J, Husson F. FactoMineR: An R package for multivariate analysis. *Journal of Statistical Software*. Mar 2008;25(1):1–18.

58. Kanehisa M. The KEGG database. *Novartis Found Symp*. 2002;247:91–101; discussion 101–103, 119–128, 244–152.

59. Albers E. Metabolic characteristics and importance of the universal methionine salvage pathway recycling methionine from 5'-methylthioadenosine. *IUBMB Life*. Dec 2009;61(12):1132–42.

60. Tang B, Kadariya Y, Murphy ME, Kruger WD. The methionine salvage pathway compound 4-methylthio-2-oxobutanate causes apoptosis independent of down-regulation of ornithine decarboxylase. *Biochem Pharmacol*. Sep 28, 2006;72(7):806–15.

61. Oram SW, Ai J, Pagani GM, et al. Expression and function of the human androgen-responsive gene ADI1 in prostate cancer. *Neoplasia*. Aug 2007;9(8):643–51.

62. Carbonnelle-Puscian A, Copie-Bergman C, Baia M, et al. The novel immunosuppressive enzyme IL4I1 is expressed by neoplastic cells of several B-cell lymphomas and by tumor-associated macrophages. *Leukemia*. May 2009;23(5):952–60.

63. Hughes AL. Origin and diversification of the L-amino oxidase family in innate immune defenses of animals. *Immunogenetics*. Dec 2010;62(11–12):753–9.

64. Nei M, Rooney AP. Concerted and birth-and-death evolution of multigene families. *Annu Rev Genet*. 2005;39:121–52.

Supplementary Tables

Supplementary Table 1. Data collected from Ensembl release 51 (Nov 2008).

Ensembl identifier	Common name	Scientific name	Number of genes	Number of transcripts
ENSP	Human	Homo sapiens	21971	60953
ENSPPY	Orangutan	Pongo pygmaeus	20068	29256
ENSMMU	Macaque	Macaca mulatta	21905	42370
ENSMUS	Mouse	Mus musculus	23873	43630
ENSRNO	Rat	Rattus norvegicus	22503	37672
ENSCPO	Guinea pig	Cavia porcellus	18673	24334
ENSCAF	Dog	Canis familiaris	19305	29804
ENSBTA	Cow	Bos taurus	21036	29517
ENSECA	Horse	Equus caballus	20322	28128
ENSMOD	Opossum	Monodelphis domestica	19471	34132
ENSOAN	Platypus	Ornithorynchus anatinus	17951	29227
ENSGAL	Chicken	Gallus gallus	16736	22945
ENSXET	Xenopus	Xenopus tropicalis	18023	28619
ENSORL	Medaka	Oryzias latipes	19686	25174
ENSDAR	Zebrafish	Danio rerio	21322	35967
ENSGAC	Stickleback	Gasterosteus aculeatus	20787	29096
ENSTNI	Tetraodon	Tetraodon nigroviridis	19602	23909

Supplementary Table 2.
Supplementary Table 2 is available from 8814SupplementaryFile.zip

9 TOWARDS AN EVOLUTIONARY VIEW OF HUMAN SYSTEMS

Achieving a system-level description of biological phenomena is a significant challenge that requires the integration of omics data from different biological levels. Evolutionary systems biology goes even further to try to understand the evolutionary mechanisms that shaped the complex systems we observe today. In this context, EvoluCodes provide a multi-scale overview of gene-level evolutionary histories. By integrating the information contained in EvoluCodes with system-level information, we can transfer the evolutionary information to a higher biological level and open the way to an innovative system level extraction of evolutionary knowledge.

9.1 Defining biological systems and their evolutionary context

Today, many biological networks are available to describe biological phenomena at the systems level and their number is constantly expanding thanks to the possibilities offered by the new high throughput technologies. Unfortunately, we are still far from having an exhaustive description of complete cellular networks. Despite the limitations induced by their simplification, schematic models of biological networks at the cellular level are currently being built, with more and more genes implicated in multiple cellular processes (see chapter 4). For example, recent biomedical research has highlighted the complexity of cellular systems, as disease dysfunctions can be linked to several functional network modules (Menon and Farina, 2011). In this context, studying network evolutionary history can offer many opportunities and understanding the phenomena that have shaped and extended biological networks over billions of years is one of the keys for the development of network-based therapies (Barabasi et al., 2011). A number of pathway evolutionary histories have been reconstructed by expert integration of phylogenetic analysis based on sequences, shared structural domains and interactomic data, for example (Comparot-Moss and Denyer, 2009; Gazave et al., 2009; Oberst et al., 2008; van Dam et al., 2011). These studies demonstrate that the topology of a network is tightly linked to its evolutionary history, with new interactions appearing after gene duplication, enzyme recruitment or the derivation of a pathway module (Yamada and Bork, 2009). However, this evolutionary knowledge has emerged from single pathway studies and no methodologies exist to perform high-throughput pathway-level evolutionary studies and to extract evolutionary information from many pathways simultaneously. Developing new methodologies to exploit biological networks and their evolutionary histories will pave the way to understanding the more complete and complex networks that will be described in the future.

9.1.1 Towards a conceptual system-level evolutionary map

A biological system can be defined as a set of interactions between different compounds implicated in a biological phenomenon. The interactors include RNA, metabolites, proteins and other chemical compounds. Several biological databases have been developed for the description of such systems (for a complete review see chapter 4). The networks and pathways described in these databases often correspond to manually defined sub-parts of a complete biological system and consist of

molecular components that interact to initiate a particular biological response. Thus, biological phenomena such as 'carbohydrate synthesis', 'ATP production' or 'response to viral infection' can be associated with a list of genes implicated in the process. These definitions are clearly linked to a subjective view of biological processes and most current network definitions are an artificially bound sample of the true system structure and behaviour. Indeed, most genes implicated in manually defined networks also interact with other partners in biological processes corresponding to other phenomena. A good illustration of this is the class of genes called 'hub' genes. Hubs often have multiple biological functions and may play a role in various unrelated processes. Their expression depends on many factors, such as tissue specificity or the response to different external stimuli (Lehner et al., 2006). Nevertheless, such a network description is an effective simplification for human readability and their delimitation facilitates systems-level evolutionary studies.
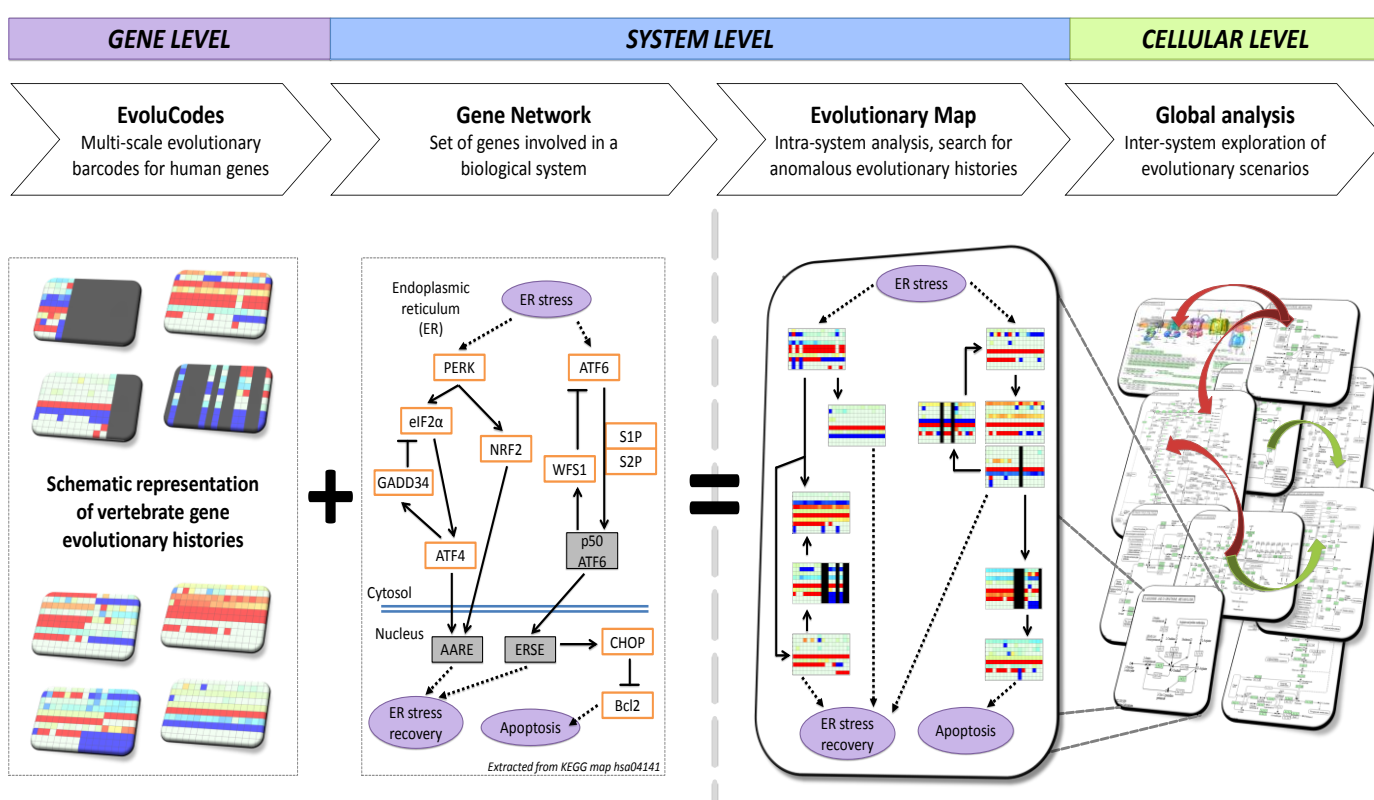
Figure 9-1. Framework to construct and explore multi-level evolutionary network maps. *EvoluCodes are assigned to individual genes and mapped onto a known gene network, such as a KEGG pathway map.*

Starting from this observation, we wanted to explore whether the evolutionary information contained in the EvoluCodes could be used in the context of a biological pathway to transfer gene-level knowledge to the system level, allowing new insights into the evolutionary history of the system. We thus conceived the concept of the 'evolutionary map', a combination of the EvoluCode and pathway data (figure 9-1). An 'evolutionary map' describes the evolutionary context corresponding to a particular biological phenomenon. As genes can participate in different networks, the same EvoluCode, corresponding to a single gene, can be considered in different biological

contexts. This fact opens up many possibilities. For example, we can analyze systems evolutionary behaviour in an intra-network framework or even perform an inter-network analysis, thus approaching a cellular level evolutionary analysis.

9.1.2 Knowledge extraction at the system-level

 The evolutionary maps provide a framework for studying evolutionary histories from a system-level point of view. A single map can contain heterogeneous gene evolutionary histories, i.e. EvoluCodes with different profiles. One of the advantages of EvoluCodes in the context of biological pathways is that they describe evolutionary histories based on multi-level parameters. In a pilot study (described in detail in publication n°2), we mapped our evolutionary EvoluCodes to the KEGG human methionine salvage pathway, a sub-pathway of the methionine pathway (KEGG id: hsa00270). We calculated pairwise distances between EvoluCodes in order to identify branches or 'hot spots' of the network, corresponding to unusual evolutionary histories compared to the rest of the network. We highlighted two genes, *adi1* and *il4i1,* for which EvoluCode discrepancies are observed. An in-depth analysis of the functions and localization of the corresponding genes suggested that the observed evolutionary discrepancies might be linked to unusual evolutionary patterns not only related to genomic data, but also to concepts such as centrality in networks, patterns of expression or different evolutionary constraints related to the topology of the pathway. This study highlighted the potential of EvoluCodes for the study of evolutionary patterns, not only in a set of genes but also in a complete biological system. This exciting result motivated us to extend our approach to a larger set of biological systems.

We thus set out to develop a methodology to automatically apply our approach to all human pathways. To investigate the events that may have occurred in the evolution of the genes participating in a specific biological phenomenon, we used a knowledge extraction approach on the evolutionary maps (figure 9-2). We applied an anomaly detection algorithm called the Local Outlier Factor (LOF) to identify 'outlier' EvoluCodes, i.e. genes with an unusual EvoluCode compared to the other genes in the pathway. The basic concept of LOF is to map the objects under study, here the EvoluCodes, into a Euclidean space and then calculate the local density of the objects, where the locality is defined by k nearest neighbours. By comparing the local density of an EvoluCode to the local densities of its neighbours, we can identify regions of similar density, as well as EvoluCodes that have a substantially lower density than their neighbours. The latter are assigned a degree of outlierness (see chapter VI for LOF algorithm details). The LOF score thus represents the cohesiveness of the EvoluCode in the context of its pathway. The authors of the LOF algorithm consider that a score less than 1 indicates a clear inlier object, i.e. a cohesive EvoluCode. EvoluCodes with a LOF score significantly greater than 1 are considered as outliers. However, the threshold determining a clear outlier depends on the dataset. Here, we defined the outlier threshold value as the upper quartile of the LOF scores for all EvoluCodes in the context of the 248 human pathways studied, which was 1.037.
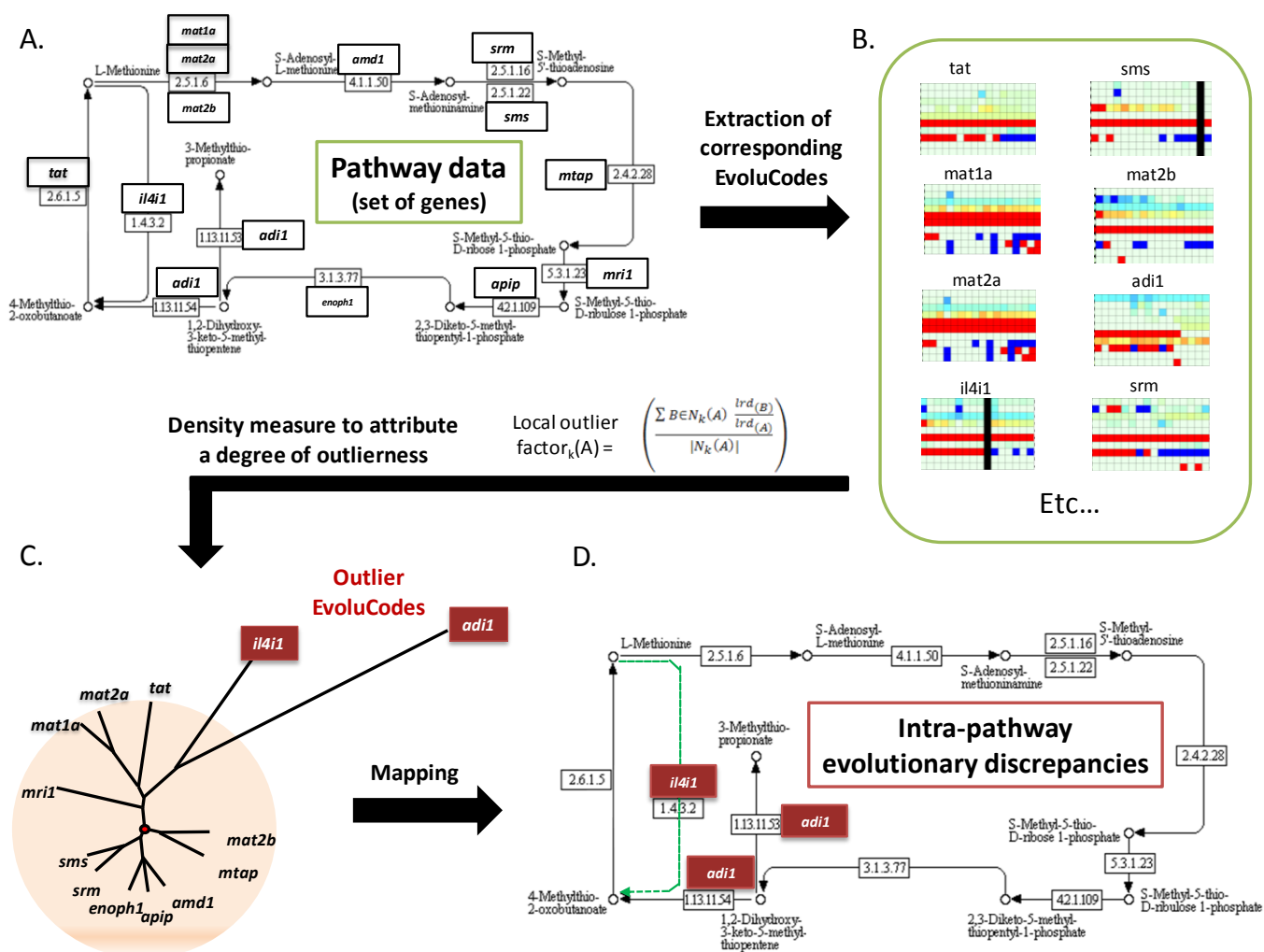
Figure 9-2. Methodology to attribute an 'outlier' status to a gene evolutionary history in the context of its pathway. *A. The data extracted from the KEGG database defines a set of pathway genes. B. The corresponding EvoluCodes are extracted and define the evolutionary context of the map. C. The Local Outlier Factor (LOF) is used to highlight EvoluCodes with the highest discrepancy relative to the pathway evolutionary context. D. The analysis of the evolutionary map of the pathway provides new insights into systems evolution.*

The 'outlier' status of an EvoluCode strongly depends on the evolutionary context in which it is considered, e.g. the scope of the considered pathway. Consequently, defining a biologically relevant pathway is a crucial step in extracting reliable evolutionary information. This idea is illustrated in the following example.
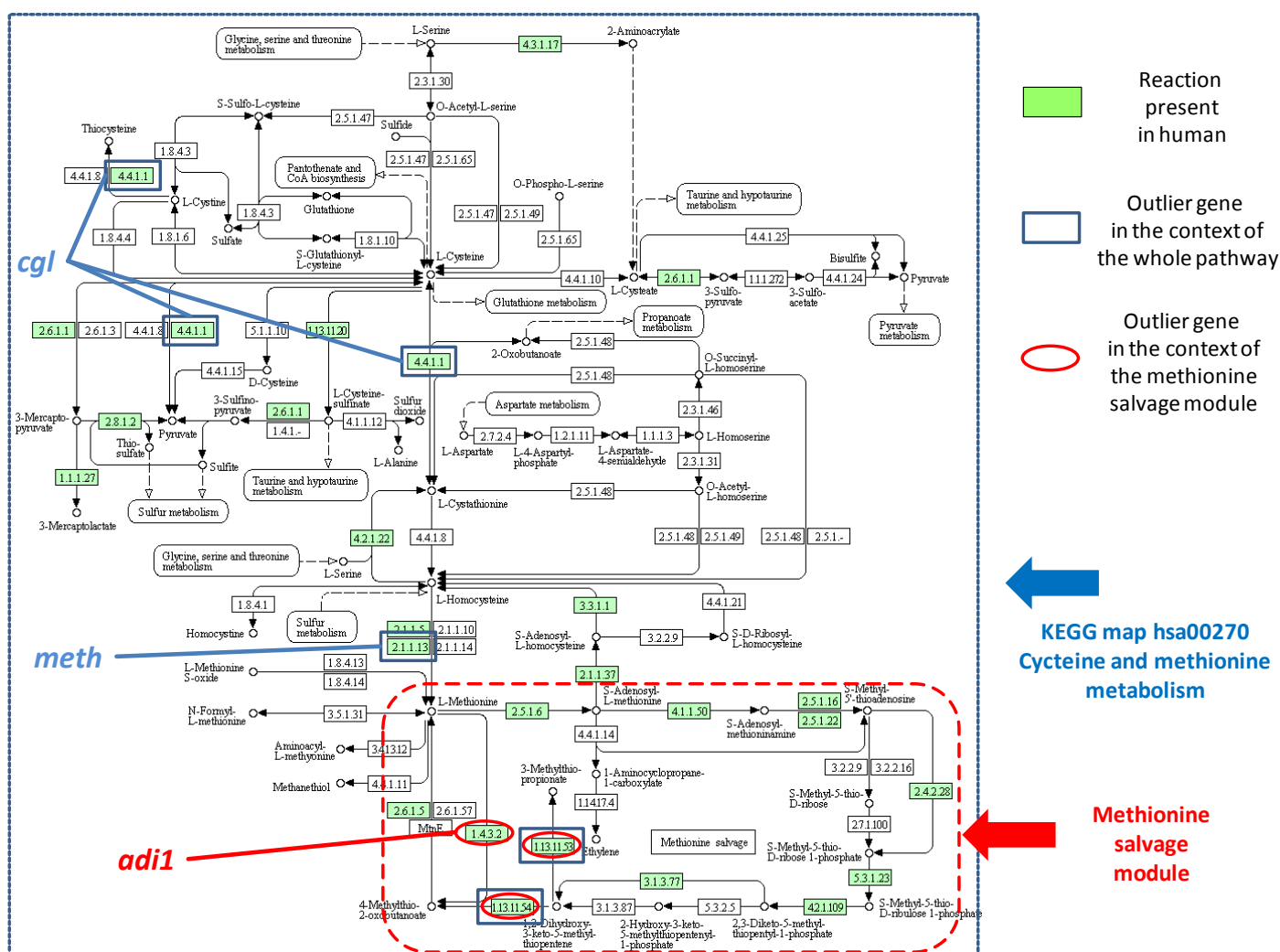
Figure 9-3. Relation between the pathway context and the definition of outlier genes. *The use of the LOF algorithm produces different outliers depending on the pathway context that is considered. Outliers in the methionine pathway are indicated by blue rectangles and outliers in the methionine salvage pathway by red circles.*

Figure 9-3 shows the methionine pathway and one of its modules: the methionine salvage pathway. We calculated outlier EvoluCodes in the context of both of these pathways. The methionine salvage pathway is common to all eukaryotes and is a module of the basal cellular metabolism (Thomas et al., 2000). We observe that the *adi1* gene is considered as an outlier within the context of this module and in fact, *adi1* is known to be expressed only in lymphocytes. This discrepancy, linked to patterns of expression, is not apparent in the context of the whole methionine pathway. However, new information can be extracted as two new outlier genes are detected. First, the *mtr* gene codes for a long modular enzyme (METH_HUMAN), with five domains catalyzing different functions and is highly conserved in all vertebrates. In contrast, most metabolic enzymes in this pathway have unique single catalytic domain and are often less conserved. We hypothesize that the *mtr* gene may be subject to specific evolutionary constraints due to its complex domain organisation. This is supported by the fact that the gene has been implicated in many human diseases (Doolin et al., 2002). Second, the *cth* gene is characterized by a protein product with broad substrate specificity (CGL_HUMAN) as suggested by the three reactions in the KEGG map and in the literature (Krishnan et al., 2011). The

corresponding EvoluCode indicates higher sequence divergence compared to the other genes in the whole pathway, possibly due to the low affinity nature of the protein. These examples demonstrate that the methodology can be applied to different levels of networks to extract different evolutionary information. A large-scale approach can be used to highlight global evolutionary trends, while a more focused analysis may reveal detailed features of local networks.

9.2   Exploiting the EvoluCodes to elucidate system-level evolution

In this section, I will describe in more detail a large-scale analysis of pathway evolutionary maps, involving all human pathways defined in the KEGG PATHWAY database. Most of the processes in the database correspond to metabolic pathways, signalling pathways or cellular responses to stimuli. These are well-defined pathways, corresponding to well-defined biological phenomena. We thus constructed 248 pathway-level evolutionary maps corresponding to all human KEGG pathway definitions, with a total of 5849 EvoluCodes mapped to the genes in these pathways. This set of human 'evolutionary maps' was our basis for the results described in the following paragraphs.

To extract evolutionary knowledge at the system-level, we calculated LOF scores for all genes in the 248 evolutionary maps and identified a total of 1147 outlier genes. Most evolutionary maps contain between 10 and 20% of outlier genes (figure 9-4).
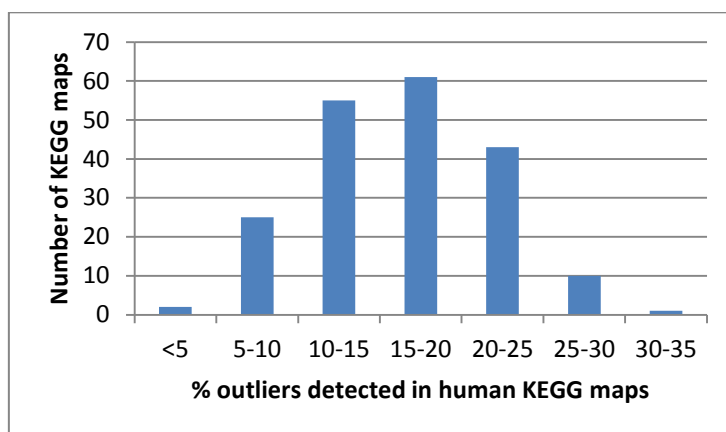
Figure 9-4. Percentage of genes with anomalous, outlier EvoluCodes in 248 human metabolic pathways from the KEGG database.

We then calculated the evolutionary cohesiveness of the pathways, i.e. the proportion of outliers (figure 9-5). Maps with the highest cohesiveness are typically involved in universal biological processes such as translation or cell growth/death, in line with previous observations (Fokkens and Snel, 2009).
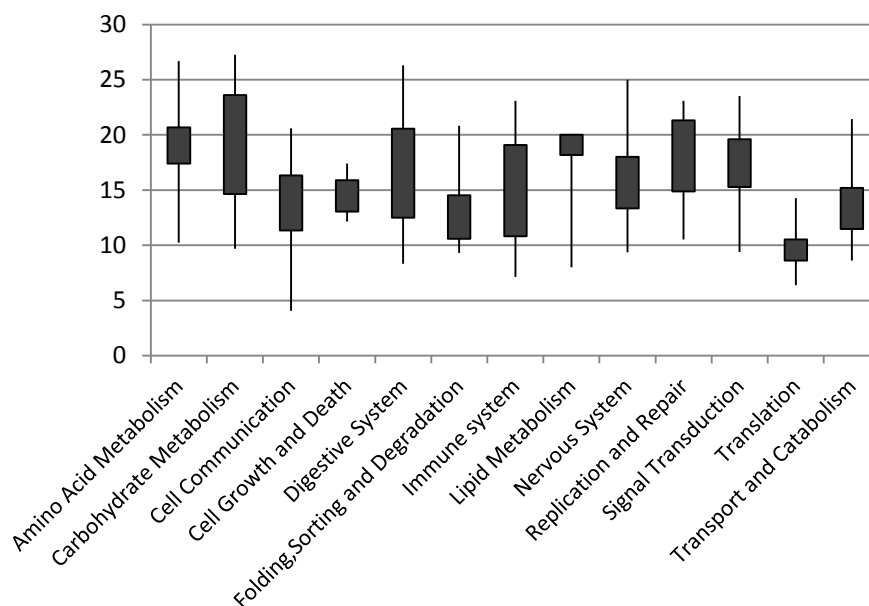
Figure 9-5. Percentage of genes with outlier EvoluCodes in 248 human metabolic pathways from KEGG. *Pathways are classified into functional groups. (Pathways with a larger number of outlier genes have less cohesive evolutionary histories).*

To investigate the biological significance of genes with anomalous evolutionary histories, we performed several studies to measure the correlations between outlier evolutionary histories and intra-pathway and inter-pathway criteria.

9.2.1 Links between gene evolutionary history and network topology

Network topology is an important feature of biological systems. Some genes can be highly connected to many other interactors in the same or different biological processes (Bock et al., 2012). Other genes are found in long linear chains and have few interactors, for example genes coding for specialized enzymes in metabolic network modules. The framework of the evolutionary maps provides an opportunity to confront gene evolutionary history and network topology and allows to perform this analysis on a large scale.

To study the relationship between gene evolutionary histories and network topology, we focused on the metabolic networks defined in KEGG. These pathways are represented by graphs of metabolite nodes connected by enzymatic reactions (one or several genes being associated with the reaction). All KEGG metabolic pathways describe the same entity (metabolites) with the same type of link (enzymatic reactions). This is not the case for other KEGG pathways, where links can represent many concepts (enzymatic reaction, transport, activation, inhibition…), and therefore the evolution of the topology of these graphs is difficult to interpret.

We restricted our topological study to the 41 human metabolic pathways containing more than 20 genes. This condition was introduced to ensure statistical power during the LOF estimation of 'outlier' EvoluCodes. To study network topology, we used the network reaction and its underlying gene as a basal unit. We defined 6 topological classes of metabolic reactions, depending on the local redundancy and connectivity properties of the network (figure 9-6). A reaction can be assigned to more than one topological class. For example, multiple genes can catalyze the same reaction (class A) to produce a metabolite that is also used in another pathway (class C). We therefore annotated reactions associated with a combination of local topologies, provided that the combination is observed more than 20 times over all pathways. Only the combinations A&C and C&E reach this count.
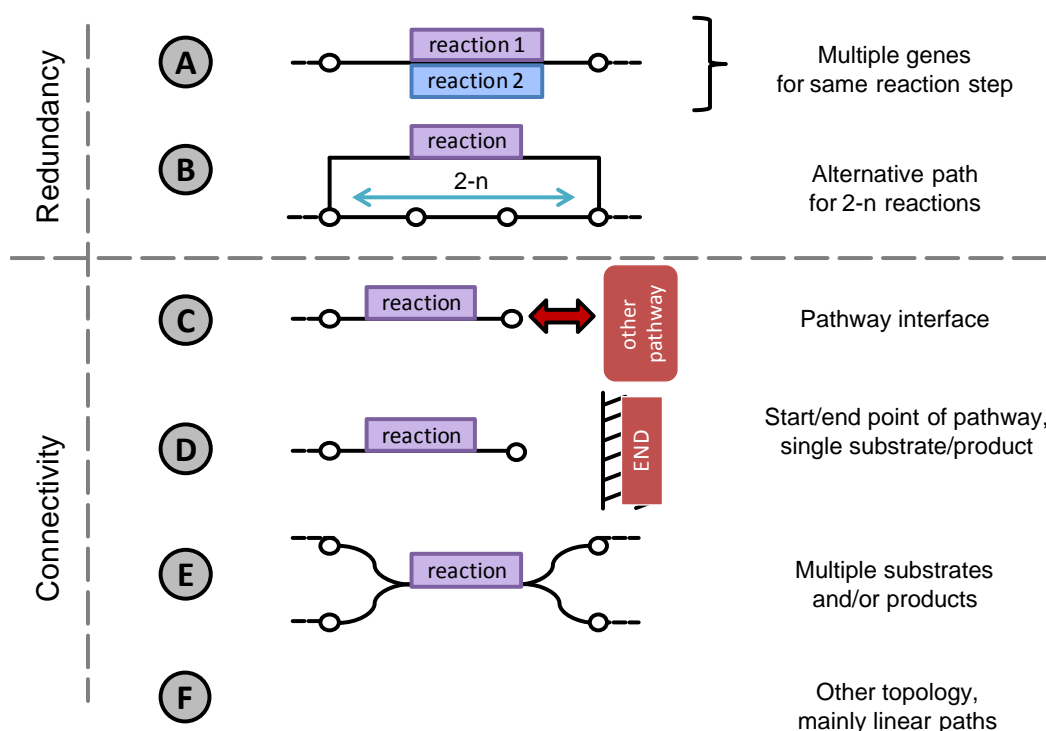
Figure 9-6. Definition of 6 classes of local topological motifs in metabolic pathways. *The class depends on the redundancy and connectivity of the reactions (and associated genes) in the network.*

We constructed an initial set of 41 human metabolic pathways, containing a total of 875 different reactions (figure 9-7). Half of the reactions do not present a particular topology and belongs to the F class (figure 9-7 A) and all pathways have at least one reaction of class F. As might be expected, class C is observed in 95% of the pathways, confirming the high inter-pathway connectivity of metabolic pathways (Caetano-Anolles et al., 2009). All topologies, except class B, are observed in more than 50% of the pathways. Interestingly, the absolute number of reactions describing specific local topologies does not correlate with this pathway repartition (figure 9-7 B). This indicates that the topologies that we defined are not homogeneously shared over the KEGG pathways. For example, catabolism-related pathways are characterized by many pathway interfaces (topology C), i.e.

reactions producing metabolites that are central to all metabolic pathways (sugars, pyruvate, acetyl-Coa, etc.). In contrast, anabolism-related pathways are more enriched with D topologies, as many cofactors are required for the transfer of chemical groups during synthesis of complex metabolites (e.g. transfer of a methyl group). In total, >40% of the reactions have a class C topology, >25% of the reactions have a class A topology, while each of the other topologies are associated with <13% of the reactions. Nevertheless, 58% of the reactions are associated with a specific topology (i.e. topologies A to E), giving a good coverage for evolutionary knowledge extraction for these classes. Moreover, we observe balanced sample sizes that can be used to calculate a potential enrichment of evolutionary history outliers.

A.

Proportion of pathways with at least one reaction with local topology

B.

Proportion of reactions associated with local topology
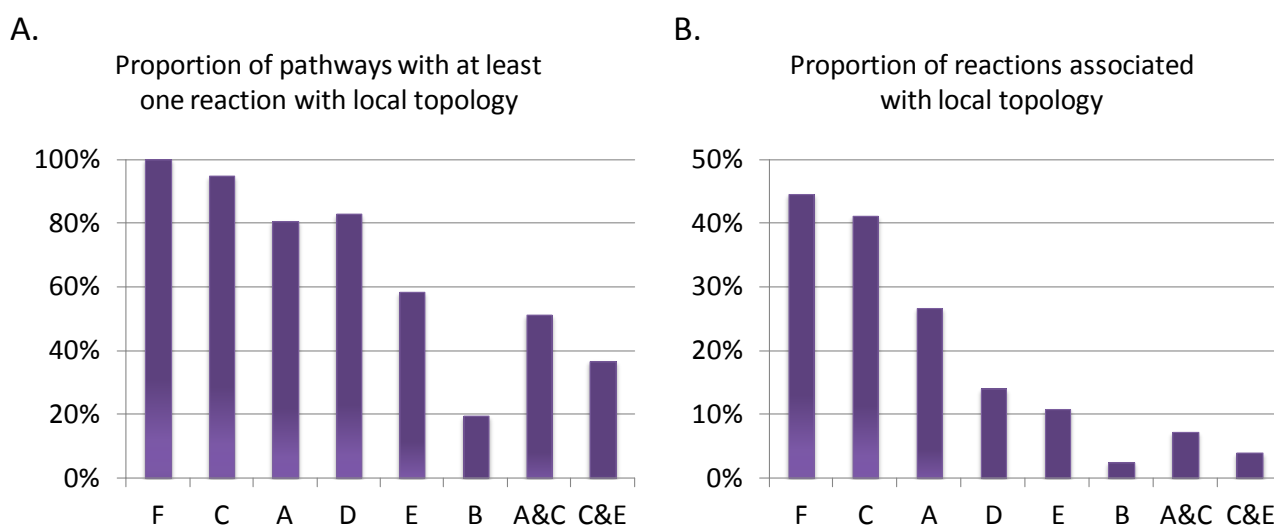
Figure 9-7. Repartition of metabolic reactions associated with topological classes.
X-axis corresponds to topological classes defined in figure 9-6.

To introduce an evolutionary component to the topology of the metabolic networks, we cross-referenced our outlier EvoluCode data with the topological data in the 41 metabolic maps. Since several genes can be associated with a single reaction, in particular if the reaction is catalyzed by a protein complex, we assigned the gene with the lowest LOF score to each reaction, i.e. the gene with the most cohesive EvoluCode. Thus, a metabolic reaction and its local topology is associated with the outlier status of the corresponding gene. We observed that 671 reactions were associated with cohesive genes and 204 reactions with outlier genes. We studied the repartition of all outlier EvoluCodes in the different topological classes (figure 9-8).

The results show that the cohesiveness of a gene in its network context depends on the local topological structure. We observe a clear enrichment of outlier genes in topologies D (27% of outlier genes) and C (22%). This can be compared to the class F, which contains only 8% of outliers despite representing 42% of the metabolic pathway genes. Thus, the smallest proportion of outliers was found at the nodes for which no particular topology was described (class F). In contrast, more outliers were found at the start/end points of a pathway (class D), and at the interface between pathways, the so called network 'hubs' (class C). The correlation that we observe between gene conservation and local network topology may be due to specific selection pressures, as proposed

previously (Yamada and Bork, 2009). Interestingly, the A&C combination of topology classes is associated with a similar enrichment. This may highlight a particular evolutionary pressure associated with multiple enzymes that catalyze a reaction to produce a metabolite that can be used in several pathways. We hypothesize that this particular topology reflects a network hot-spot for metabolite regulation.

A.

Repartition of outliers
in 6 topological classes

B.

Topological classification of outliers normalised
by total number of reactions
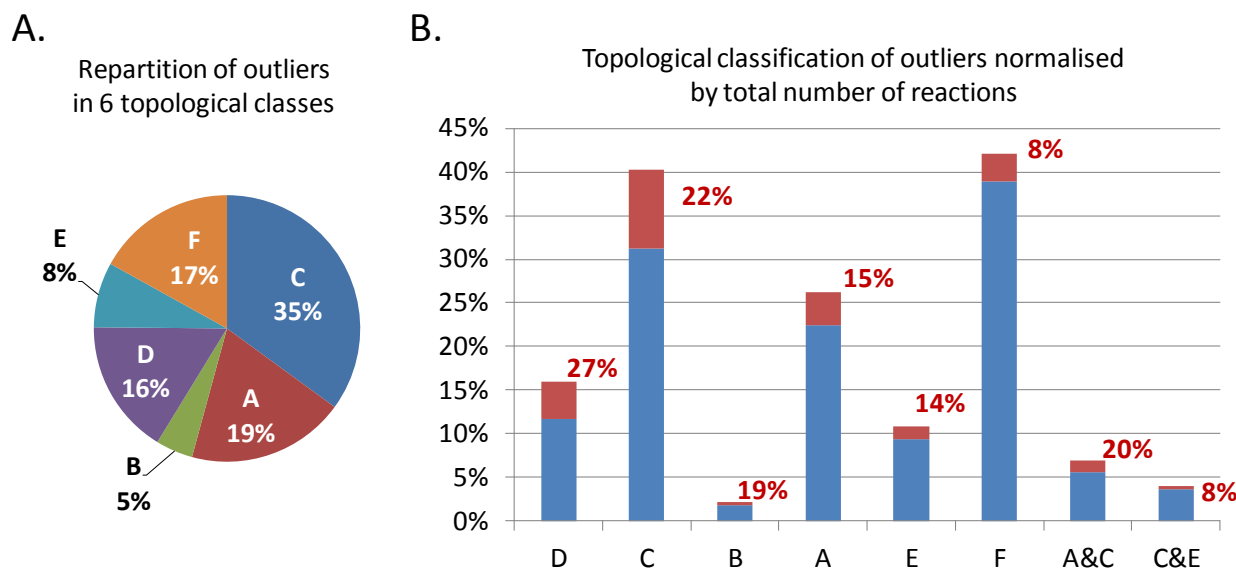
Figure 9-8. Topological inventory of outlier reactions in human metabolic pathways. *(A) Distribution of outlier genes in the 6 topological classes. (B) Distribution of non-outlier genes (blue) and the proportion of outliers (red) in each class. Capital letters correspond to the proportion of genes belonging to a class. X-axis corresponds to topological classes defined in figure 9-6.*

These results demonstrate that local network topology is correlated with evolutionary history. Furthermore, the outliers in the evolutionary maps can be used to pinpoint evolutionary 'hot spots', where differential evolution is observed. Numerous in-depth analyses could be performed on these maps to uncover the evolutionary mechanisms that gave rise to these hot spots. However, a key feature of metabolic networks oriented our research in another direction. Metabolic pathways are highly interconnected and numerous genes participate in different metabolisms. In fact, this observation can be extended to all KEGG pathways and the following section will describe an analysis of EvoluCodes from an inter-pathway point of view.

9.2.2 Towards an integrative view of evolutionary phenomena at the cellular level

Individual pathways often function in a coordinated fashion and understanding the interactions or crosstalk between pathways is important for deciphering complex cellular processes, such as the appropriate physiological responses to internal or external stimuli. The evolutionary cohesiveness of a gene is context-dependent, i.e. a gene may be defined as cohesive in one of these pathways and as

an outlier in another. Such cases of differential evolutionary conservation may indicate important events, such as gene duplications, rearrangements or losses, and the subsequent gain or loss of interactions in the network. To investigate these high-level processes of evolution, we identified all genes involved in the crosstalk between 155 KEGG pathway maps, reflected by the fact that all the genes in the set were present in at least 3 maps. For each pair of KEGG maps, we calculated the proportion of outlier genes observed in the overlapping set of genes shared between the two systems (figure 9-9).
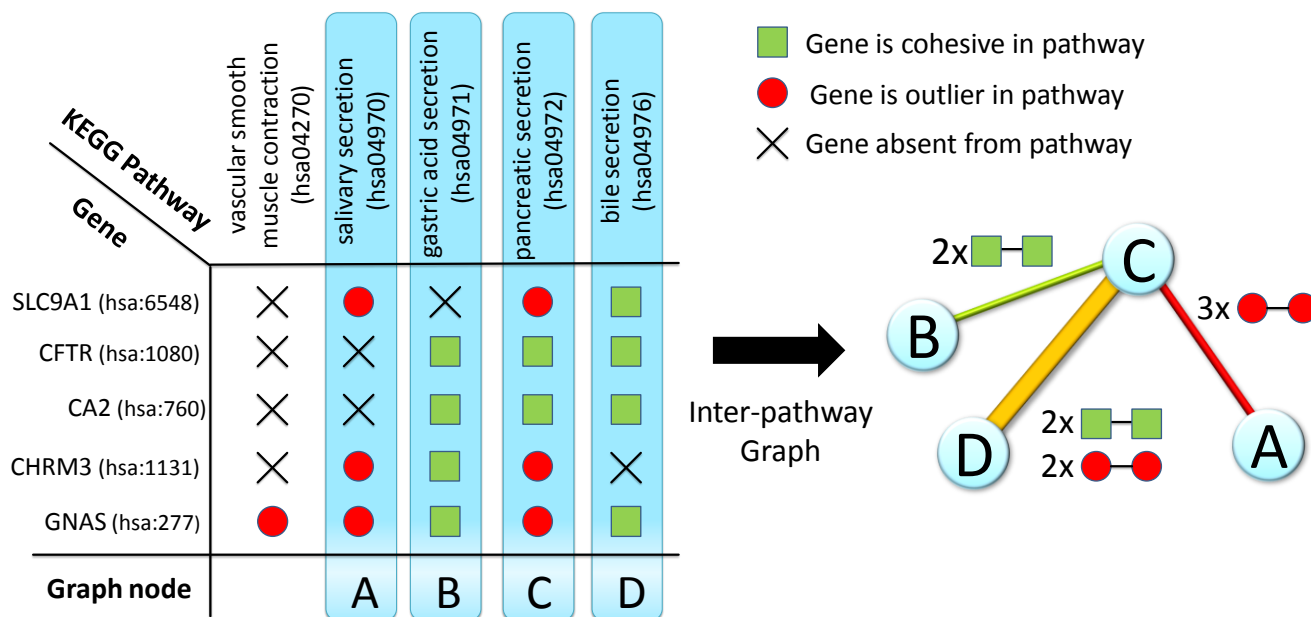
Figure 9-9. Characterization of crosstalk between pathways. *The crosstalk between 2 systems is characterized by the proportion of shared outlier genes, indicated by a color gradient from green (all cohesive) to red (all outlier).*

We then constructed a global map of the relationships between the 155 maps, representing the evolutionary behaviour of these pathways during vertebrate evolution (Figure 9-10). The exploration of this map provides a powerful and visual means of highlighting important events in the evolution of human biological systems. Two examples are highlighted in Figure 9-10 and described below.

Evolutionary map of human pathways
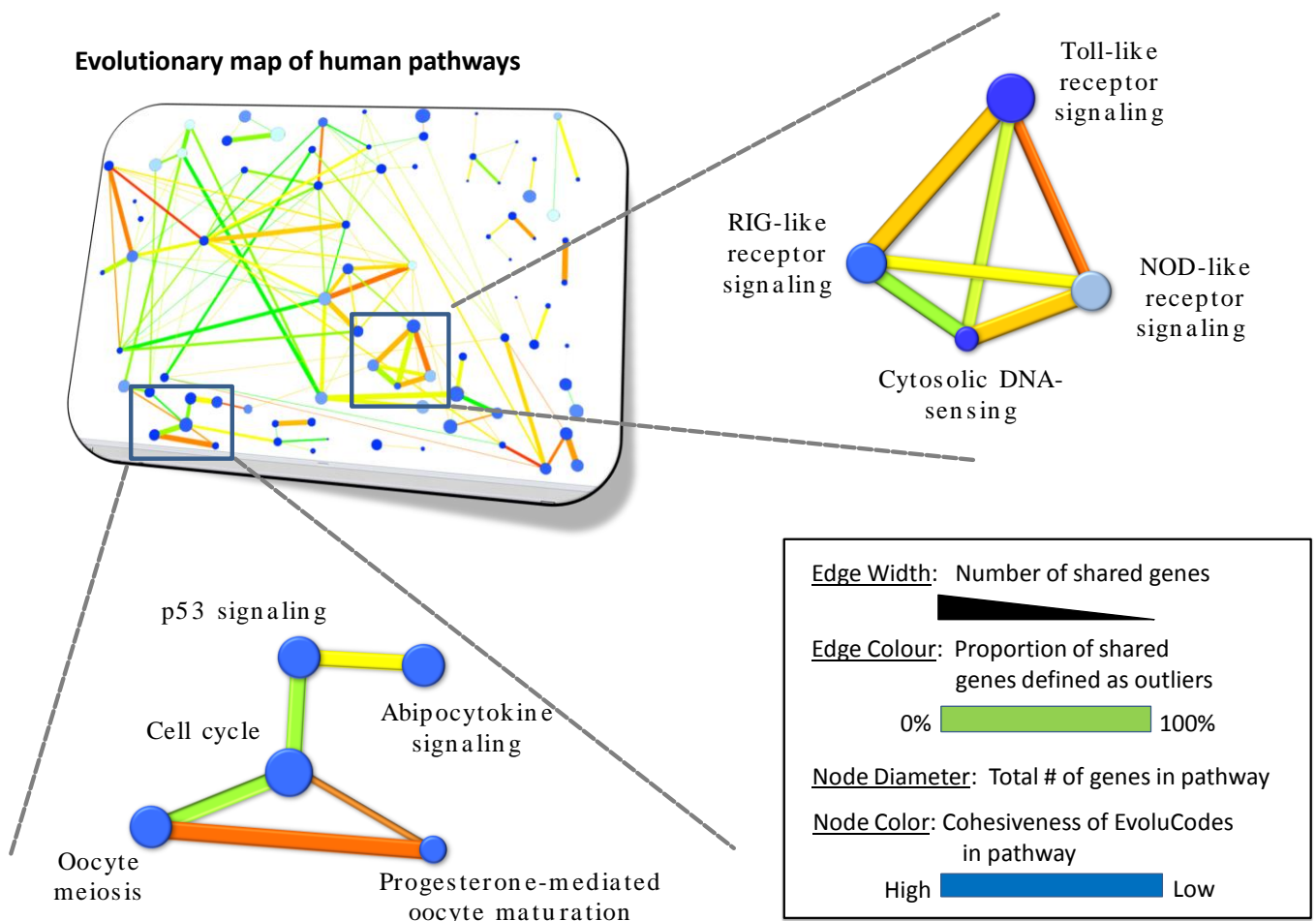
Figure 9-10. Integrative map of vertebrate evolutionary histories at the cellular level. *An integrated evolutionary map of selected human pathways show the number and cohesiveness of the gene evolutionary histories, associated with individual pathways (nodes) and pathway crosstalk (edges).*

Cell cycle and oocyte meiosis and maturation pathways

The cell cycle and oocyte meiosis pathways are well conserved in most vertebrates. This is reflected in the fact that, in the global evolutionary map (Figure 9-10), we observe that the genes involved in the crosstalk between the cell cycle and oocyte meiosis are generally cohesive with the other genes in these pathways. In contrast, the exact nature of oocyte maturation varies in different species, since the females of some species produce thousands of eggs at a time, while in others, females produce relatively few mature eggs (Vasudevan et al., 2006). These differences are reflected in the higher proportion of EvoluCode outliers in the crosstalk with the progesterone-mediated oocyte maturation pathway (Figure 9-10 and Figure 9-11). A number of functional specificities are highlighted by the following outliers:

- PMYT1_HUMAN is a cdc2-inhibitory kinase, which acts as a negative regulator of entry into mitosis during the cell cycle. Inspection of the EvoluCode indicates a more divergent sequence family than is typical for this conserved pathway. This might be a result of the different functions of Myt1, which is also implicated in control of entry into meiosis, either alone (as in Xenopus) or in concert with Wee1 (as in mouse oocytes) (Gaffre et al., 2011).

- CDK2_HUMAN is a highly studied cyclin-dependent kinase that functions in the cell cycle in S phase progression (Liu and Kipreos, 2000). It also plays a role in the regulation of progesterone receptor (PR) signaling (Moore et al., 2007). Although the EvoluCode shows high sequence conservation in most vertebrates studied here, a perturbation is highlighted in Monodelphis domestica (opossum), Ornithorhynchus anatinus (platypus) and Gallus gallus (chicken) with lower sequence identity and fewer inparalog/co-ortholog relationships, reflecting the different CDK complements of these species.

KEGG Identifier	Uniprot Identifier	Cell cycle (hsa04110)	Oocyte meiosis (hsa04114)	Progesterone-mediated oocyte maturation (hsa04914)
hsa:9088	PMYT1_HUMAN	1	0	0
hsa:1017	CDK2_HUMAN	0	0	1
hsa:699	BUB1_HUMAN	0	0	0
hsa:5347	PLK1_HUMAN	0	0	0
hsa:995	MPIP3_HUMAN	0	1	-1
hsa:9126	SMC3_HUMAN	0	0	-1
hsa:891	CCNB1_HUMAN	0	0	-1
hsa:9700	ESPL1_HUMAN	0	0	-1
hsa:4085	MD2L1_HUMAN	0	0	-1
hsa:991	CDC20_HUMAN	0	0	-1
hsa:898	CCNE1_HUMAN	0	0	-1
hsa:8454	CUL1_HUMAN	0	0	-1
hsa:64506	CPEB1_HUMAN	-1	1	0
hsa:3480	IGF1R_HUMAN	-1	1	0
hsa:3630	INS_HUMAN	-1	1	0
hsa:4342	CCNB2_HUMAN	-1	0	1
hsa:9133	MOS_HUMAN	-1	0	1
hsa:5604	MP2K1_HUMAN	-1	0	0
hsa:5241	PRGR_HUMAN	-1	0	0
hsa:993	MPIP1_HUMAN	0	-1	0
hsa:51343	FZR_HUMAN	0	-1	1

PMYT1_HUMAN

CDK2_HUMAN

Figure 9-11. Cellular level analysis of KEGG pathways involved in the cell cycle or oocyte meiosis and maturation. *Cohesiveness of genes shared by at least 2 of the 3 pathways: genes with cohesive EvoluCodes for a given pathway are shown in green, genes with outlier EvoluCodes are highlighted in red, genes shown in grey are not present in the pathway.*

Innate immune system

The innate immune system relies on pattern recognition receptors (PRRs) that recognize different pathogens, such as viruses or bacteria, and that trigger intracellular signalling cascades ultimately culminating in the expression of proinflammatory molecules (Mogensen, 2009). Toll-like receptors (TLR) are membrane-bound PRRs, located either at the cell surface where they mainly recognize bacterial products, or in intracellular compartments where they are involved in recognition of nucleic acids. Cytosolic PRRs, including RIG-I-like receptors (RLR) and NOD-like receptors (NLR), mainly recognize intracellular RNA. Finally, cytoplasmic localization of DNA by cytosolic DNA sensors seems to be involved in mounting a response to both bacteria and DNA viruses. The EvoluCodes in these pathways are more variable than those described in the previous section, reflecting known species-specific immunities. The outlier genes involved in the crosstalk between these pathways are shown in figure 9-12 and two examples are described below:

- IL1B_HUMAN (interleukin-1 beta, IL-1β) is a cytokine that plays a crucial role in mediation and amplification of the innate immune response to bacterial pathogens. It is produced as an inactive precursor, pro-IL-1β, in response to molecular motifs carried by pathogens. After processing, the active IL-1β molecule is secreted. The EvoluCode shows low conservation in terms of sequence identity, domain organization and synteny and IL1B_HUMAN is an outlier in the 3 innate immune system pathways where it is present. In fact, it has been hypothesized that IL-1 β may have arisen by a reverse transcriptase mediated duplication of the alpha gene (Clark et al., 1986).

- RIPK1_HUMAN is a receptor interacting protein, which plays a crucial role in the cellular response to TLR and RLR signals, switching between cell survival through RIP1 activation of NF-κB and cell death induced by caspase-8 cleavage of RIP1 (Festjens et al., 2007). The EvoluCode shows normal sequence conservation in vertebrate evolution, while synteny is only observed in mammals and not in fishes. This evolutionary scenario is considered to be unusual in the RLR signalling pathway, but is cohesive in the TLR signalling pathway. This is in agreement with a recent study that highlighted the acquisition of several fish-specific immune system components in the TLR signalling pathway (Xiang et al., 2012). In fact, RIP1 is known to play a different role in TLR signalling in the fish lineage (Rebl et al., 2010).

KEGG Identifier	Uniprot Identifier	Toll-like (hsa4620)	NOD-like (hsa4621)	RIG-I-like (hsa4622)	DNA sensors (hsa4623)
hsa:1147	IKKA_HUMAN	1	0	0	0
hsa:8517	NEMO_HUMAN	0	0	1	-1
hsa:6885	M3K7_HUMAN	1	1	0	-1
hsa:3576	IL8_HUMAN	1	0	0	-1
hsa:7124	TNFA_HUMAN	1	0	0	-1
hsa:841	CASP8_HUMAN	0	0	0	-1
hsa:7189	TRAF6_HUMAN	0	0	0	-1
hsa:3551	IKKB_HUMAN	0	0	0	-1
hsa:3553	IL1B_HUMAN	1	1	-1	1
hsa:6352	CCL5_HUMAN	0	1	-1	0
hsa:5970	TF65_HUMAN	0	1	-1	0
hsa:3569	IL6_HUMAN	0	1	-1	0
hsa:4792	IKBA_HUMAN	0	0	-1	0
hsa:3665	IRF7_HUMAN	0	-1	0	1
hsa:8737	RIPK1_HUMAN	0	-1	1	0
hsa:3627	CXL10_HUMAN	0	-1	0	0
hsa:3661	IRF3_HUMAN,	0	-1	0	0
hsa:29110	TBK1_HUMAN	0	-1	0	0
hsa:9641	IKKE_HUMAN	0	-1	0	0
hsa:3456	IFNB_HUMAN	0	-1	0	0
hsa:10454	TAB1_HUMAN	0	0	-1	-1
hsa:8772	FADD_HUMAN	0	-1	1	-1
hsa:7187	TRAF3_HUMAN	1	-1	0	-1
hsa:6300	MK12_HUMAN	0	-1	0	-1
hsa:29108	ASC_HUMAN	-1	0	-1	0
hsa:834	CASP1_HUMAN	-1	0	-1	0
hsa:3606	IL18_HUMAN	-1	0	-1	0
hsa:340061	TM173_HUMAN	-1	-1	0	0
hsa:57506	MAVS_HUMAN	-1	-1	0	0
hsa:23586	DDX58_HUMAN	-1	-1	0	0

IL1B_HUMAN
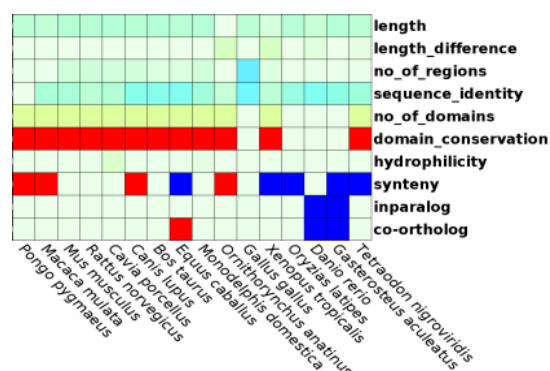
RIPK1_HUMAN
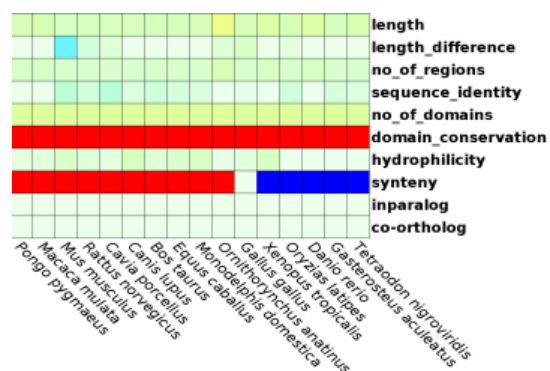
Figure 9-12. Cellular level analysis of KEGG pathways in the innate immume system.
Cohesiveness of genes shared by at least 2 of the 4 pathways: genes with cohesive EvoluCodes for a given pathway are shown in green, genes with outlier EvoluCodes are highlighted in red, genes shown in grey are not present in the pathway.

9.3   Conclusion

Unravelling the evolutionary history of biological pathways will contribute to our fundamental understanding of how biological processes vary with time, but describing evolutionary scenarios at the system level remains challenging. By combining the EvoluCodes with pathway information, we were able to extract large-scale evolutionary messages for the human metabolic systems and their topologies. We then extended our analyses to an inter-pathway exploration, providing original evolutionary maps to describe major innovations in vertebrate systems. Such an approach is ideally adapted to the evolutionary systems biology philosophy: biological data from many biological levels is integrated and innovative high-throughput methodologies are used to explore the evolutionary phenomena that led to the extant states of human biological systems.

10 CONCLUSION & PERSPECTIVES

Evolutionary studies have a long history, and have experienced two major paradigm shifts. The first one was the discovery of the DNA structure in 1953, a discovery that initiated the whole field of molecular biology. The second one is the current expansion of systems biology. After sixty years of gene and genome evolutionary studies, we now have the opportunity to open the black box that was the intermediate between genomes and phenotypes. Achieving a system-level description of molecular evolution will be essential to fully understanding the evolutionary histories that modelled extant organisms. In addition to a better characterization of protein function, understanding systems evolution will shed light on the relationships between sequence diversity and functional diversity. This idea has already been confirmed in several studies, which concluded that the most important factor in the evolution of function is not amino acid sequence, but rather the cellular context in which proteins act (Nehrt et al., 2011). Integrating systems biology data is the key to describing systems evolution and the emerging field of evolutionary systems biology is now focusing on this task (Loewe, 2009). New techniques are currently being developed, combining systems biology, laboratory evolution experiments or large-scale mutational analyses to understand how evolution shapes organisms (Papp et al., 2011). Nevertheless, this context imposes a real challenge inherent to systems biology: how to integrate information representing multiple biological levels in an evolutionary framework? This context motivated the work of this thesis.

The need for efficient tools for comprehensive evolutionary analyses

The first part of this thesis addressed the problem of orthology inference in the post-genomic era. Despite being well established in molecular biology, large-scale orthology inference is challenging because innovations in genome sequencing now allow the complete sequencing of new genomes every week. So many genomes are a great opportunity for phylogenetic studies, comparative genomics and genome annotations but are also useful for inferring the main systemic innovations that characterize different phyla. Unfortunately, current evolutionary studies are generally limited to a small portion of this information.

One challenge in orthology inference is thus to find a way to manage this data increase, while at the same time maintaining a high data quality. We developed the OrthoInspector algorithm for this purpose. It uses an inparalogy-based approach, focusing first on the detection of recent gene duplication events, in order to subsequently reconstruct complex co-orthologous relationships between multiple organismse. This algorithm and its implementation proved to be efficient for inferring orthology relationships in hundreds of genomes at the same time. We made these data available for the biological community through an online database.

The simple inference of large-scale data is however insufficient. When dealing with hundreds of genomes, it is essential to provide efficient analysis tools for data extraction and a comprehensive visualisation of the corresponding orthology relations. The second objective of OrthoInspector was therefore to address these issues. The software suite integrates tools to construct complex queries based on phylogenetic patterns and several visualisation tools.

These tools have been regularly updated since the first release of OrthoInspector, including the addition of new tools dedicated to the summarizing of phylogenetic profiles and interoperability standards with the incorporation of the OrthoXML format. The second version of OrthoInspector , supported by an updated database containing most of the current eukaryotic complete genomes and the release of a second database containing all available genomes of bacteria and archaea, will be published in the near future. The publication will demonstrate the robustness of the OrthoInspector software for predicting large-scale orthology relationships, relations that are central in functional and evolutionary studies.

Multi-scale integration of evolutionary parameters

The second part of this thesis involved the development of new approaches to study evolution, based on a systems biology philosophy. We created the concept of Evolutionary Barcodes or EvoluCodes, a formalism designed to integrate heterogeneous biological parameters related to multiple biological levels. EvoluCodes summarize multi species variations in a unified framework that can be exploited using standard knowledge extraction tools. EvoluCodes complement classical phylogenetic approaches since they can integrate several biological parameters to describe gene evolutionary histories. Our first implementation of the EvoluCode concerns the human proteome and represents gene evolutionary histories since the appearance of vertebrates. The subsequent analysis demonstrated their potential for discovering new knowledge about vertebrate evolution.

The structure of the EvoluCode can be easily modified or extended. Thus, adding new biological parameters is an interesting option for providing a more comprehensive picture of gene evolutionary histories. For example, the group of Pierre Pontarotti has inferred a large number of genetic events that shaped the chordate genomes (Levasseur et al., 2012b). These predictions are based on high quality phylogenies and the annotated multiple alignments that we constructed using PipeAlign and MACSIMS. It would be interesting to incorporate these data in order to refine the current 'domain_conservation' barcode parameter, which is based only on the sequence alignment. Other EvoluCode parameters to describe the genomic context could also be extracted from the GeCco database (an in-house developed database), which provides information about gene promoters and intergenic regions (GeCco database, manuscript in preparation). Finally, gene expression data would be an important enhancement to the EvoluCode, for example the data generated by Brawand et al. concerning five tissues in various mammals and chicken (Brawand et al., 2011). New parameters should however be carefully selected. Indeed, some parameters such as 'sequence_length' and 'number_domains' are correlated. A new parameter should be included in the EvoluCode only if it provides a valuable amount of new information. A statistical analysis, based on component analysis for example, could help to identify the most informative EvoluCode parameters.

Another area where statistical estimation could improve the EvoluCodes is int the definition of typical and atypical values. During the process of EvoluCode construction, we tested several statistical distributions for the different EvoluCode parameters but they did not provide satisfactory results. Further investigations are required for a more robust description of the variations observed between the different parameters and a more accurate characterization of unusual patterns or behaviour.

The integrative power of the EvoluCode should facilitate the exploitation of evolutionary-based analyses in a wide variety of large-scale applications. For instance, its capacity to detect evolutionary anomalies could be used to improve gene function characterization, by highlighting unusual evolutionary histories that might indicate neo- or sub-functionalization. A crucial aspect of such analyses will be the ability to distinguish true innovations from anomalies due to errors in genome sequencing or prediction of gene introns/exon structures. Indeed, in our study of the effects of such errors on the multiple alignment construction process (publication n°4) and the inference of genetic events (publication n°3), we highlighted the importance of error detection and quality control processes in evolutionary inference.

Another important application is the study of human genetic diseases. In this context, the laboratory has developed a complete infrastructure dedicated to knowledge discovery concerning missense variants implicated in human disease (publication n°6). The system uses a combination of sequence/structure/evolution predicates to characterize mutations in human genes and provides tools for prediction of deleterious mutations based on Inductive Logic Programming Rules. The current evolutionary predicates are extracted from MACSIMS alignments, but an EvoluCode module is currently under development and will be used, not only for mutation prediction, but also for gene prioritisation to identify potential new genes implicated in a disease.

Inferring evolutionary knowledge at the system-level

We developed the first prototype of a tool that can be used to decipher evolutionary knowledge at the cellular level. We applied the tool to construct an evolutionary map of human cellular networks and investigated some important sub-networks. The next step will be to integrate other types of gene and pathway data to facilitate the interpretation of the map. For example, GO annotation could be used to characterize the genes that are shared between the different pathways and a GO enrichment analysis could be performed to highlight functions that correlate with unusual gene evolutionary histories in a given context. Similarly, it might be interesting to investigate potential correlations between alternative transcript data and specific EvoluCode profiles or outlierness, in order to test the hypothesis that genes that have evolved to produce more variants are more likely to be implicated in many interactions and cellular processes and are thus subject to specific evolutionary constraints. So far, we have focused our studies on human network data, but evolutionary maps could also be constructed for other model organisms for which reliable pathways are available, providing new insights into the evolutionary processes that shaped the different animal systems.

These fundamental studies are clearly important, but EvoluCodes could also be exploited in more applied studies. We are currently constructing an evolutionary map dedicated to the human disease pathways referenced in KEGG. Previously, the genes shared by metabolic networks were mapped to disease annotations from the OMIM database (Barabasi et al., 2011) and scores were calculated for each gene to estimate its morbidity, i.e. the amount of perturbations and disease it might induce. Our disease-related evolutionary map will allow us to investigate whether these genes are associated with particular evolutionary histories. Then, by searching for other genes with similar EvoluCodes, we could predict their involvement in human diseases.

Why is understanding evolution at the system level so important?

The study of biological network evolution will increase our understanding of the evolutionary principles that model Life. However, understanding systems evolution is not restricted to theoretical biology and some applied biology fields are beginning to exploit evolutionary principles as a technical tool.

One of the domains that is benefiting from systems evolutionary theory is bioengineering. For example, the artificial gene circuits and network modules generated in synthetic biology can be optimized with adaptive laboratory evolution (ALE). This strategy allows the metabolic engineering of microorganisms by combining genetic variation with the selection of beneficial mutations in an unbiased fashion (Portnoy et al., 2011). Currently, ALE is mainly applied in well-characterized organisms such as *Saccharomyces cerevisiae* and *Escherichia coli*. The description of population evolutionary models are also actively used in synthetic biology (Rothschild, 2010). On the other hand, synthetic engineering of biological systems is also contributing to our comprehension of systems evolution. In particular, experimentally rewired circuits in living cells allow a direct testing of hypotheses in evolutionary systems biology (Davidson et al., 2012). For example, building small genetic regulatory systems can provide insight on the trade-offs that constrain adaptation and can shape the structure of biological networks. The *de novo* construction of genomes could also give new perspectives for recreating ancestral systems in the laboratory (Gibson et al., 2010).

Another domain where evolutionary principles are becoming increasingly important is medicine. The first applications were mainly related to viral and population phylogenies. For example, phylogenies were used to track epidemics in human populations (Pybus and Rambaut, 2009). They were also used to determine whether viral outbreaks were due to new circulating vaccine-derived polioviruses (Kew et al., 2004). At the system level, artificial positive selection has also been used to understand the pathogenic mechanisms of infection. This strategy was for example successfully applied to HIV (Bozek and Lengauer, 2010) and influenza studies (Li et al., 2011). But the most promising field for reuniting medicine and evolution is the field of personalized medicine, where the study of evolutionary variation will be essential to understanding the biological network diversity that is observed in humans (Chen et al., 2012).

Such system level exploitations will finally provide Evolution the credibility that it is so often denied by its detractors. Maybe the next important revolution in Evolution will be the widespread use of its principles to improve food production, to solve ecological problems or to develop new therapeutic strategies. After shifting from theories to concrete applications, it will be difficult to deny 150 years of evolutionary research.

LIST OF REFERENCES

Abagyan, R.A., and Batalov, S. (1997). Do aligned sequences share the same fold? J Mol Biol *273*, 355-368.

Abascal, F., and Valencia, A. (2003). Automatic annotation of protein function based on family identification. Proteins *53*, 683-692.

Abeln, S., and Deane, C.M. (2005). Fold usage on genomes and protein fold evolution. Proteins *60*, 690-700.

Adam, M., Murali, B., Glenn, N.O., and Potter, S.S. (2008). Epigenetic inheritance based evolution of antibiotic resistance in bacteria. Bmc Evolutionary Biology *8*.

Aerts, S. (2012). Computational strategies for the genome-wide identification of cis-regulatory elements and transcriptional targets. Curr Top Dev Biol *98*, 121-145.

Alexander, M.R., Wheatley, A.K., Center, R.J., and Purcell, D.F. (2010). Efficient transcription through an intron requires the binding of an Sm-type U1 snRNP with intact stem loop II to the splice donor. Nucleic Acids Res.

Altenhoff, A.M., and Dessimoz, C. (2009). Phylogenetic and functional assessment of orthologs inference projects and methods. PLoS Comput Biol *5*, e1000262.

Altenhoff, A.M., and Dessimoz, C. (2012). Inferring orthology and paralogy. Methods Mol Biol *855*, 259-279.

Altenhoff, A.M., Schneider, A., Gonnet, G.H., and Dessimoz, C. (2011). OMA 2011: orthology inference among 1000 complete genomes. Nucleic Acids Research *39*, D289-D294.

Altenhoff, A.M., Studer, R.A., Robinson-Rechavi, M., and Dessimoz, C. (2012). Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. PLoS Comput Biol *8*, e1002514.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J Mol Biol *215*, 403-410.

Altschul, S.F., and Koonin, E.V. (1998). Iterated profile searches with PSI-BLAST - a tool for discovery in protein databases. Trends in Biochemical Sciences *23*, 444-447.

Andersson, J.O. (2005). Lateral gene transfer in eukaryotes. Cell Mol Life Sci *62*, 1182-1197.

Angiuoli, S.V., and Salzberg, S.L. (2011). Mugsy: fast multiple alignment of closely related whole genomes. Bioinformatics *27*, 334-342.

Aoki-Kinoshita, K.F. (2008). An introduction to bioinformatics for glycomics research. Plos Computational Biology *4*.

Apweiler, R., and constortium, U. (2012). Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Res *40*, D71-75.

Aravind, L., Anantharaman, V., and Venancio, T.M. (2009). Apprehending multicellularity: regulatory networks, genomics, and evolution. Birth Defects Res C Embryo Today *87*, 143-164.

Armougom, F., Moretti, S., Poirot, O., Audic, S., Dumas, P., Schaeli, B., Keduas, V., and Notredame, C. (2006). Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. Nucleic Acids Res *34*, W604-608.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T.*, et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet *25*, 25-29.

Babu, M.M. (2010a). Early Career Research Award Lecture. Structure, evolution and dynamics of transcriptional regulatory networks. Biochem Soc Trans *38*, 1155-1178.

Babu, M.M. (2010b). Structure, evolution and dynamics of transcriptional regulatory networks. Biochemical Society Transactions *38*, 1155-1178.

Baitaluk, M., Kozhenkov, S., Dubinina, Y., and Ponomarenko, J. (2012). IntegromeDB: an integrated system and biological search engine. BMC Genomics *13*, 35.

Baker, M. (2012). Proteomics: The interaction map. Nature *484*, 271-275.

Baker, M.E. (1997). Steroid receptor phylogeny and vertebrate origins. Mol Cell Endocrinol *135*, 101-107.

Bandyopadhyay, S., Sharan, R., and Ideker, T. (2006). Systematic identification of functional orthologs based on protein network comparison. Genome Res *16*, 428-435.

Bapteste, E., O'Malley, M.A., Beiko, R.G., Ereshefsky, M., Gogarten, J.P., Franklin-Hall, L., Lapointe, F.J., Dupre, J., Dagan, T., Boucher, Y.*, et al.* (2009). Prokaryotic evolution and the tree of life are two different things. Biol Direct *4*, 34.

Barabasi, A.L., and Albert, R. (1999). Emergence of scaling in random networks. Science *286*, 509-512.

Barabasi, A.L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. Nat Rev Genet *12*, 56-68.

Barabasi, A.L., and Oltvai, Z.N. (2004). Network biology: understanding the cell's functional organization. Nat Rev Genet *5*, 101-113.

Bard, J. (2010). A systems biology view of evolutionary genetics: network-driven processes incorporate much more variation than evolutionary genetics can handle. This variation is hard to formalise but allows fast change. Bioessays *32*, 559-563.

Bard, N., Bolze, R., Caron, E., Desprez, F., Heymann, M., Friedrich, A., Moulinier, L., Nguyen, N.H., Poch, O., and Toursel, T. (2010). Decrypthon grid - grid resources dedicated to neuromuscular disorders. Stud Health Technol Inform *159*, 124-133.

Baroni-Urbani, C. (1980). A statistical table for the degree of coexistence between two species. Oecologia, 287-289.

Bartlett, J.M., and Stirling, D. (2003). A short history of the polymerase chain reaction. Methods Mol Biol *226*, 3-6.

Bashor, C.J., Horwitz, A.A., Peisajovich, S.G., and Lim, W.A. (2010). Rewiring cells: synthetic biology as a tool to interrogate the organizational principles of living systems. Annu Rev Biophys *39*, 515-537.

Bergsten, J. (2005). A review of long-branch attraction. Cladistics *21*, 163-193.

Bisson, N., James, D.A., Ivosev, G., Tate, S.A., Bonner, R., Taylor, L., and Pawson, T. (2011). Selected reaction monitoring mass spectrometry reveals the dynamics of signaling through the GRB2 adaptor. Nat Biotechnol *29*, 653-658.

Blanchette, M. (2007). Computation and analysis of genomic multi-sequence alignments. Annu Rev Genomics Hum Genet *8*, 193-213.

Blomme, T., Vandepoele, K., De Bodt, S., Simillion, C., Maere, S., and Van de Peer, Y. (2006). The gain and loss of genes during 600 million years of vertebrate evolution. Genome Biol *7*, R43.

Bock, M., Ogishima, S., Tanaka, H., Kramer, S., and Kaderali, L. (2012). Hub-Centered Gene Network Reconstruction Using Automatic Relevance Determination. PLoS One *7*.

Bock, R. (2010). The give-and-take of DNA: horizontal gene transfer in plants. Trends Plant Sci *15*, 11-22.

Boeckmann, B., Robinson-Rechavi, M., Xenarios, I., and Dessimoz, C. (2011). Conceptual framework and pilot study to benchmark phylogenomic databases based on reference gene trees. Briefings in Bioinformatics *12*, 423-435.

Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K.D., Ovcharenko, I., Pachter, L., and Rubin, E.M. (2003). Phylogenetic shadowing of primate sequences to find functional regions of the human genome. Science *299*, 1391-1394.

Boto, L. (2010). Horizontal gene transfer in evolution: facts and challenges. Proc Biol Sci *277*, 819-827.

Boucher, Y., Douady, C.J., Papke, R.T., Walsh, D.A., Boudreau, M.E., Nesbo, C.L., Case, R.J., and Doolittle, W.F. (2003). Lateral gene transfer and the origins of prokaryotic groups. Annu Rev Genet *37*, 283-328.

Boyle, J., Rovira, H., Cavnor, C., Burdick, D., Killcoyne, S., and Shmulevich, I. (2009). Adaptable data management for systems biology investigations. BMC Bioinformatics *10*, 79.

Bozek, K., and Lengauer, T. (2010). Positive selection of HIV host factors and the evolution of lentivirus genes. Bmc Evolutionary Biology *10*.

Bradford, Y., Conlin, T., Dunn, N., Fashena, D., Frazer, K., Howe, D.G., Knight, J., Mani, P., Martin, R., Moxon, S.A.*, et al.* (2011). ZFIN: enhancements and updates to the Zebrafish Model Organism Database. Nucleic Acids Res *39*, D822-829.

Bravo, H.C., and Irizarry, R.A. (2010). Model-based quality assessment and base-calling for second-generation sequencing data. Biometrics *66*, 665-674.

Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csardi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A., Kircher, M.*, et al.* (2011). The evolution of gene expression levels in mammalian organs. Nature *478*, 343-348.

Breunig, M.M., Kriegel, H.P., Ng, R.T., and Sander, J. (2000). LOF: Identifying density-based local outliers. Sigmod Record *29*, 93-104.

Brinkmann, H., and Philippe, H. (2007). The diversity of eukaryotes and the root of the eukaryotic tree. Adv Exp Med Biol *607*, 20-37.

Brower, A. (2010). Should Evolutionary Theory Evolve? Scientist *24*, 14-14.

Brown, D., and Sjolander, K. (2006). Functional classification using phylogenomic inference. Plos Computational Biology *2*, 479-483.

Bru, C., Courcelle, E., Carrre, S., Beausse, Y., Dalmar, S., and Kahn, D. (2005). The ProDom database of protein domain families: more emphasis on 3D. Nucleic Acids Research *33*, D212-D215.

Brudno, M., Do, C.B., Cooper, G.M., Kim, M.F., Davydov, E., Green, E.D., Sidow, A., and Batzoglou, S. (2003). LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. Genome Res *13*, 721-731.

Bylesjo, M., Eriksson, D., Kusano, M., Moritz, T., and Trygg, J. (2007). Data integration in plant biology: the O2PLS method for combined modeling of transcript and metabolite data. Plant J *52*, 1181-1191.

Byrne, K.P., and Wolfe, K.H. (2005). The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. Genome Res *15*, 1456-1461.

Caetano-Anolles, G., Yafremava, L.S., Gee, H., Caetano-Anolles, D., Kim, H.S., and Mittenthal, J.E. (2009). The origin and evolution of modern metabolism. International Journal of Biochemistry & Cell Biology *41*, 285-297.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics *10*, 421.

Cannon, S.B., and Young, N.D. (2003). OrthoParaMap: distinguishing orthologs from paralogs by integrating comparative genome data and gene phylogenies. BMC Bioinformatics *4*, 35.

Carpentier, M., Brouillet, S., and Pothier, J. (2005). YAKUSA: A fast structural database scanning method. Proteins-Structure Function and Bioinformatics *61*, 137-151.

Caspi, R., Altman, T., Dreher, K., Fulcher, C.A., Subhraveti, P., Keseler, I.M., Kothari, A., Krummenacker, M., Latendresse, M., Mueller, L.A.*, et al.* (2012). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res *40*, D742-753.

Cassman, M., and Center, W.T.E. (2007). Systems Biology: International Research and Development (Springer).

Caulfield, T., and McGuire, A.L. (2012). Direct-to-Consumer Genetic Testing: Perceptions, Problems, and Policy Responses. Annual Review of Medicine, Vol 63 *63*, 23-33.

Chang, A.N. (2009). Prioritizing genes for pathway impact using network analysis. Methods Mol Biol *563*, 141-156.

Chen, B.S., and Lin, Y.P. (2011). On the Interplay between the Evolvability and Network Robustness in an Evolutionary Biological Network: A Systems Biology Approach. Evolutionary Bioinformatics *7*, 201-233.

Chen, B.S., and Wu, W.S. (2007). Underlying principles of natural selection in network evolution: systems biology approach. Evol Bioinform Online *3*, 245-262.

Chen, C., Thakkar, S., Knoblock, C., and Shahabi, C. (2003). Automatically annotating and integrating spatial datasets, Vol 2750 (Berlin, ALLEMAGNE, Springer).

Chen, F., Mackey, A.J., Stoeckert, C.J., and Roos, D.S. (2006). OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. Nucleic Acids Research *34*, D363-D368.

Chen, F., Mackey, A.J., Vermunt, J.K., and Roos, D.S. (2007). Assessing Performance of Orthology Detection Strategies Applied to Eukaryotic Genomes. PLoS One *2*.

Chen, R., Mias, G.I., Li-Pook-Than, J., Jiang, L., Lam, H.Y., Miriami, E., Karczewski, K.J., Hariharan, M., Dewey, F.E., Cheng, Y.*, et al.* (2012). Personal omics profiling reveals dynamic molecular and medical phenotypes. Cell *148*, 1293-1307.

Chen, R., and Snyder, M. (2012). Systems biology: personalized medicine for the future? Curr Opin Pharmacol.

Chiu, J.C., Lee, E.K., Egan, M.G., Sarkar, I.N., Coruzzi, G.M., and DeSalle, R. (2006). OrthologID: automation of genome-scale ortholog identification within a parsimony framework. Bioinformatics *22*, 699-707.

Chowbina, S.R., Wu, X., Zhang, F., Li, P.M., Pandey, R., Kasamsetty, H.N., and Chen, J.Y. (2009). HPD: an online integrated human pathway database enabling systems biology studies. BMC Bioinformatics *10 Suppl 11*, S5.

Chung, Y., and Ane, C. (2011). Comparing two Bayesian methods for gene tree/species tree reconstruction: simulations with incomplete lineage sorting and horizontal gene transfer. Syst Biol *60*, 261-275.

Ciccarelli, F.D., Doerks, T., von Mering, C., Creevey, C.J., Snel, B., and Bork, P. (2006). Toward automatic reconstruction of a highly resolved tree of life. Science *311*, 1283-1287.

Clark, B.D., Collins, K.L., Gandy, M.S., Webb, A.C., and Auron, P.E. (1986). Genomic sequence for human prointerleukin 1 beta: possible evolution from a reverse transcribed prointerleukin 1 alpha gene. Nucleic Acids Res *14*, 7897-7914.

Cler, E., Papai, G., Schultz, P., and Davidson, I. (2009). Recent advances in understanding the structure and function of general transcription factor TFIID. Cell Mol Life Sci *66*, 2123-2134.

Coe, C.L., Savage, A., and Bromley, L.J. (1992). Phylogenetic Influences on Hormone Levels across the Primate Order. American Journal of Primatology *28*, 81-100.

Comparot-Moss, S., and Denyer, K. (2009). The evolution of the starch biosynthetic pathway in cereals and other grasses. J Exp Bot *60*, 2481-2492.

Conesa, A., Prats-Montalban, J.M., Tarazona, S., Nueda, M.J., and Ferrer, A. (2010). A multiway approach to data integration in systems biology based on Tucker3 and N-PLS. Chemometrics and Intelligent Laboratory Systems *104*, 101-111.

Council, N.R. (2010). Steps Toward Large-Scale Data Integration in the Sciences:Summary of a Workshop (The National Academies Press).

Covert, M.W., Schilling, C.H., Famili, I., Edwards, J.S., Goryanin, I.I., Selkov, E., and Palsson, B.O. (2001). Metabolic modeling of microbial strains in silico. Trends in Biochemical Sciences *26*, 179-186.

Creevey, C.J., Muller, J., Doerks, T., Thompson, J.D., Arendt, D., and Bork, P. (2011). Identifying Single Copy Orthologs in Metazoa. Plos Computational Biology *7*.

Croft, D., O'Kelly, G., Wu, G.M., Haw, R., Gillespie, M., Matthews, L., Caudy, M., Garapati, P., Gopinath, G., Jassal, B.*, et al.* (2011). Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Research *39*, D691-D697.

Crombach, A., and Hogeweg, P. (2008). Evolution of Evolvability in Gene Regulatory Networks. Plos Computational Biology *4*.

Dagan, T. (2011). Phylogenomic networks. Trends in Microbiology *19*, 483-491.

Dal'Molin, C.G.D., Quek, L.E., Palfreyman, R.W., Brumbley, S.M., and Nielsen, L.K. (2010). AraGEM, a Genome-Scale Reconstruction of the Primary Metabolic Network in Arabidopsis. Plant Physiology *152*, 579-589.

Dang, K.D., Dutt, P.B., and Forsdyke, D.R. (1998). Chargaff difference analysis of the bithorax complex of Drosophila melanogaster. Biochemistry and Cell Biology-Biochimie Et Biologie Cellulaire *76*, 129-137.

Darling, A.E., Mau, B., and Perna, N.T. (2010). progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One *5*, e11147.

Datta, R.S., Meacham, C., Samad, B., Neyer, C., and Sjolander, K. (2009). Berkeley PHOG: PhyloFacts orthology group prediction web server. Nucleic Acids Research *37*, W84-W89.

Davidson, E.A., Windram, O.P., and Bayer, T.S. (2012). Building Synthetic Systems to Learn Nature's Design Principles. Adv Exp Med Biol *751*, 411-429.

Davis, M.J., Shin, C.J., Jing, N., and Ragan, M.A. (2012). Rewiring the dynamic interactome. Mol Biosyst *8*, 2054-2066.

de la Fuente, A. (2010). From 'differential expression' to 'differential networking' - identification of dysfunctional regulatory networks in diseases. Trends Genet *26*, 326-333.

Dehal, P.S., and Boore, J.L. (2006). A phylogenomic gene cluster resource: the Phylogenetically Inferred Groups (PhIGs) database. BMC Bioinformatics *7*, 201.

Delcher, A.L., Salzberg, S.L., and Phillippy, A.M. (2003). Using MUMmer to identify similar regions in large sequence sets. Curr Protoc Bioinformatics *Chapter 10*, Unit 10 13.

DeLuca, T.F., Cui, J., Jung, J.Y., Gabriel, K.C.S., and Wall, D.P. (2012). Roundup 2.0: enabling comparative genomics for over 1800 genomes. Bioinformatics *28*, 715-716.

Demir, E., Cary, M.P., Paley, S., Fukuda, K., Lemer, C., Vastrik, I., Wu, G., D'Eustachio, P., Schaefer, C., Luciano, J.*, et al.* (2010). The BioPAX community standard for pathway data sharing. Nat Biotechnol *28*, 935-942.

Dessimoz, C., Gabaldon, T., Roos, D.S., Sonnhammer, E.L.L., Herrero, J., and Consortium, Q.O. (2012). Toward community standards in the quest for orthologs. Bioinformatics *28*, 900-904.

Deville, Y., Gilbert, D., van Helden, J., and Wodak, S. (2003). An overview of data models for the analysis of biochemical pathways. Computational Methods in Systems Biology, Proceedings *2602*, 174-174.

Dewey, C.N. (2007). Aligning multiple whole genomes with Mercator and MAVID. Methods Mol Biol *395*, 221-236.

Dewey, C.N. (2011). Positional orthology: putting genomic evolutionary relationships into context. Brief Bioinform *12*, 401-412.

Di Tommaso, P., Moretti, S., Xenarios, I., Orobitg, M., Montanyola, A., Chang, J.M., Taly, J.F., and Notredame, C. (2011). T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. Nucleic Acids Res *39*, W13-17.

Dickins, T.E., and Rahman, Q. (2012). The extended evolutionary synthesis and the role of soft inheritance in evolution. Proc Biol Sci *279*, 2913-2921.

Do, C.B., Mahabhashyam, M.S.P., Brudno, M., and Batzoglou, S. (2005). ProbCons: Probabilistic consistency-based multiple sequence alignment. Genome Research *15*, 330-340.

Doolin, M.T., Barbaux, S., McDonnell, M., Hoess, K., Whitehead, A.S., and Mitchell, L.E. (2002). Maternal genetic effects, exerted by genes involved in homocysteine remethylation, influence the risk of spina bifida. American Journal of Human Genetics *71*, 1222-1226.

Dreger, A., Kronfeld, M., Ziller, M.J., Supper, J., Planatscher, H., Magnus, J.B., Oldiges, M., Kohlbacher, O., and Zell, A. (2009). Modeling metabolic networks in C. glutamicum: a comparison of rate laws in combination with various parameter optimization strategies. Bmc Systems Biology *3*.

Dreszer, T.R., Karolchik, D., Zweig, A.S., Hinrichs, A.S., Raney, B.J., Kuhn, R.M., Meyer, L.R., Wong, M., Sloan, C.A., Rosenbloom, K.R.*, et al.* (2012). The UCSC Genome Browser database: extensions and updates 2011. Nucleic Acids Res *40*, D918-923.

Duarte, N.C., Becker, S.A., Jamshidi, N., Thiele, I., Mo, M.L., Vo, T.D., Srivas, R., and Palsson, B.O. (2007). Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proc Natl Acad Sci U S A *104*, 1777-1782.

Dufayard, J.F., Duret, L., Penel, S., Gouy, M., Rechenmann, F., and Perriere, G. (2005). Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. Bioinformatics *21*, 2596-2603.

Dzidic, S., and Bedekovic, V. (2003). Horizontal gene transfer-emerging multidrug resistance in hospital bacteria. Acta Pharmacol Sin *24*, 519-526.

Edsall, J.T. (1956). Configurations of polypeptide chains and protein molecules. J Cell Physiol Suppl *47*, 163-200.

Edwards, A.W.F., and Cavalli-sforza, L.L. (1964). Reconstruction of Evolutionary Trees. Phenetic and phylogenetic classification, 67-76.

Engelman, D.M., Steitz, T.A., and Goldman, A. (1986). Identifying Nonpolar Transbilayer Helices in Amino-Acid-Sequences of Membrane-Proteins. Annual Review of Biophysics and Biophysical Chemistry *15*, 321-353.

Engin, H.B., Keskin, O., Nussinov, R., and Gursoy, A. (2012). A Strategy Based on Protein-Protein Interface Motifs May Help in Identifying Drug Off-Targets. J Chem Inf Model.

Engstrom, P.G., Ho Sui, S.J., Drivenes, O., Becker, T.S., and Lenhard, B. (2007). Genomic regulatory blocks underlie extensive microsynteny conservation in insects. Genome Res *17*, 1898-1908.

Eppig, J.T., Blake, J.A., Bult, C.J., Kadin, J.A., Richardson, J.E., and Grp, M.G.D. (2012). The Mouse Genome Database (MGD): comprehensive resource for genetics and genomics of the laboratory mouse. Nucleic Acids Research *40*, D881-D886.

Feist, A.M., Herrgard, M.J., Thiele, I., Reed, J.L., and Palsson, B.O. (2009). Reconstruction of biochemical networks in microorganisms. Nat Rev Microbiol *7*, 129-143.

Felnagle, E.A., Chaubey, A., Noey, E.L., Houk, K.N., and Liao, J.C. (2012). Engineering synthetic recursive pathways to generate non-natural small molecules. Nat Chem Biol *8*, 518-526.

Festjens, N., Vanden Berghe, T., Cornelis, S., and Vandenabeele, P. (2007). RIP1, a kinase on the crossroads of a cell's decision to live or die. Cell Death Differ *14*, 400-410.

Field, B., and Osbourn, A.E. (2008). Metabolic diversification - Independent assembly of operon-like gene clusters in different plants. Science *320*, 543-547.

Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D., Merregaert, J., Min Jou, W., Molemans, F., Raeymaekers, A., Van den Berghe, A.*, et al.* (1976). Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. Nature *260*, 500-507.

Fitch, W.M. (1970). Distinguishing homologous from analogous proteins. Syst Zool *19*, 99-113.

Flannick, J., Novak, A., Do, C.B., Srinivasan, B.S., and Batzoglou, S. (2009). Automatic parameter learning for multiple local network alignment. J Comput Biol *16*, 1001-1022.

Flannick, J., Novak, A., Srinivasan, B.S., McAdams, H.H., and Batzoglou, S. (2006). Graemlin: general and robust alignment of multiple large interaction networks. Genome Res *16*, 1169-1181.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M.*, et al.* (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science *269*, 496-512.

Fokkens, L., and Snel, B. (2009). Cohesive versus flexible evolution of functional modules in eukaryotes. PLoS Comput Biol *5*, e1000276.

Forslund, K., Pekkari, I., and Sonnhammer, E.L. (2011). Domain architecture conservation in orthologs. BMC Bioinformatics *12*, 326.

Forterre, P., and Philippe, H. (1999). The last universal common ancestor (LUCA), simple or complex? Biol Bull *196*, 373-375; discussion 375-377.

Friedberg, I., Harder, T., Kolodny, R., Sitbon, E., Li, Z.W., and Godzik, A. (2007). Using an alignment of fragment strings for comparing protein structures. Bioinformatics *23*, E219-E224.

Friedrich, A., Ripp, R., Garnier, N., Bettler, E., Deleage, G., Poch, O., and Moulinier, L. (2007). Blast sampling for structural and functional analyses. BMC Bioinformatics *8*, 62.

Frost, A., Elgort, M.G., Brandman, O., Ives, C., Collins, S.R., Miller-Vedam, L., Weibezahn, J., Hein, M.Y., Poser, I., Mann, M.*, et al.* (2012). Functional repurposing revealed by comparing S. pombe and S. cerevisiae genetic interactions. Cell *149*, 1339-1352.

Fu, G.C.L., and Lin, W.C. (2012). Identification of gene-oriented exon orthology between human and mouse. BMC Genomics *13*.

Fulton, D.L., Li, Y.Y., Laird, M.R., Horsman, B.G., Roche, F.M., and Brinkman, F.S. (2006). Improving the specificity of high-throughput ortholog prediction. BMC Bioinformatics *7*, 270.

Gabaldon, T. (2008). Large-scale assignment of orthology: back to phylogenetics? Genome Biology *9*.

Gabaldon, T., Dessimoz, C., Huxley-Jones, J., Vilella, A.J., Sonnhammer, E.L.L., and Lewis, S. (2009). Joining forces in the quest for orthologs. Genome Biology *10*.

Gaffre, M., Martoriati, A., Belhachemi, N., Chambon, J.P., Houliston, E., Jessus, C., and Karaiskou, A. (2011). A critical balance between Cyclin B synthesis and Myt1 activity controls meiosis entry in Xenopus oocytes. Development *138*, 3735-3744.

Gallien, S., Perrodou, E., Carapito, C., Deshayes, C., Reyrat, J.M., Van Dorsselaer, A., Poch, O., Schaeffer, C., and Lecompte, O. (2009). Ortho-proteogenomics: multiple proteomes investigation through orthology and a new MS-based protocol. Genome Res *19*, 128-135.

Galperin, M.Y., and Koonin, E.V. (1998). Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption. In Silico Biol *1*, 55-67.

Gama-Castro, S., Salgado, H., Peralta-Gil, M., Santos-Zavaleta, A., Muniz-Rascado, L., Solano-Lira, H., Jimenez-Jacinto, V., Weiss, V., Garcia-Sotelo, J.S., Lopez-Fuentes, A*., et al.* (2011). RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units). Nucleic Acids Research *39*, D98-D105.

Gandhi, T.K., Zhong, J., Mathivanan, S., Karthick, L., Chandrika, K.N., Mohan, S.S., Sharma, S., Pinkert, S., Nagaraju, S., Periaswamy, B*., et al.* (2006). Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. Nat Genet *38*, 285-293.

Garcia-Moreno, J., and Mindell, D.P. (2000). Rooting a phylogeny with homologous genes on opposite sex chromosomes (gametologs): a case study using avian CHD. Mol Biol Evol *17*, 1826-1832.

Gauges, R., Rost, U., Sahle, S., and Wegner, K. (2006). A model diagram layout extension for SBML. Bioinformatics *22*, 1879-1885.

Gavin, A.C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.M., Cruciat, C.M*., et al.* (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature *415*, 141-147.

Gazave, E., Lapebie, P., Richards, G.S., Brunet, F., Ereskovsky, A.V., Degnan, B.M., Borchiellini, C., Vervoort, M., and Renard, E. (2009). Origin and evolution of the Notch signalling pathway: an overview from eukaryotic genomes. Bmc Evolutionary Biology *9*.

Gharib, W.H., and Robinson-Rechavi, M. (2011). When orthologs diverge between human and mouse. Briefings in Bioinformatics *12*, 436-441.

Ghersi, D., and Sanchez, R. (2011). Beyond structural genomics: computational approaches for the identification of ligand binding sites in protein structures. J Struct Funct Genomics *12*, 109-117.

Gibson, D.G., Glass, J.I., Lartigue, C., Noskov, V.N., Chuang, R.Y., Algire, M.A., Benders, G.A., Montague, M.G., Ma, L., Moodie, M.M*., et al.* (2010). Creation of a bacterial cell controlled by a chemically synthesized genome. Science *329*, 52-56.

Gibson, T.A., and Goldberg, D.S. (2011). Improving evolutionary models of protein interaction networks. Bioinformatics *27*, 376-382.

Giot, L., Bader, J.S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y.L., Ooi, C.E., Godwin, B., Vitols, E*., et al.* (2003). A protein interaction map of Drosophila melanogaster. Science *302*, 1727-1736.

Goel, A., Li, S.S., and Wilkins, M.R. (2011). Four-dimensional visualisation and analysis of protein-protein interaction networks. Proteomics *11*, 2672-2682.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M*., et al.* (1996). Life with 6000 genes. Science *274*, 546, 563-547.

Goh, K.I., Cusick, M.E., Valle, D., Childs, B., Vidal, M., and Barabasi, A.L. (2007). The human disease network. Proc Natl Acad Sci U S A *104*, 8685-8690.

Goldberg, A.D., Allis, C.D., and Bernstein, E. (2007). Epigenetics: a landscape takes shape. Cell *128*, 635-638.

Goldstein, R.A. (2008). The structure of protein evolution and the evolution of protein structure. Curr Opin Struct Biol *18*, 170-177.

Goodman, M., Czelusniak, J., Moore, G.W., Romeroherrera, A.E., and Matsuda, G. (1979). Fitting the Gene Lineage into Its Species Lineage, a Parsimony Strategy Illustrated by Cladograms Constructed from Globin Sequences. Systematic Zoology *28*, 132-163.

Goodsell, D.S. (2005). Representing structural information with RasMol. Curr Protoc Bioinformatics *Chapter 5*, Unit 5 4.

Goodstadt, L., and Ponting, C.P. (2006). Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. PLoS Comput Biol *2*, e133.

Goulas, T., Arolas, J.L., and Gomis-Ruth, F.X. (2011). Structure, function and latency regulation of a bacterial enterotoxin potentially derived from a mammalian adamalysin/ADAM xenolog. Proc Natl Acad Sci U S A *108*, 1856-1861.

Grehan, J.R., and Schwartz, J.H. (2009). Evolution of the second orangutan: phylogeny and biogeography of hominid origins. Journal of Biogeography *36*, 1823-1844.

Gu, J. (2011). Evolutionary systems biology. Curr Genomics *12*, 379.

Guimera, R., Stouffer, D.B., Sales-Pardo, M., Leicht, E.A., Newman, M.E.J., and Amaral, L.A.N. (2010). Origin of compartmentalization in food webs. Ecology *91*, 2941-2951.

Gupta, R.S., and Griffiths, E. (2002). Critical issues in bacterial phylogeny. Theoretical Population Biology *61*, 423-434.

Guttman, M., Garber, M., Levin, J.Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M.J., Gnirke, A., Nusbaum, C.*, et al.* (2010). Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat Biotechnol *28*, 503-510.

Haegeman, A., Jones, J.T., and Danchin, E.G. (2011). Horizontal gene transfer in nematodes: a catalyst for plant parasitism? Mol Plant Microbe Interact *24*, 879-887.

Hahn, M.W. (2007). Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution. Genome Biol *8*, R141.

Halachev, M.R., Loman, N.J., and Pallen, M.J. (2011). Calculating orthologs in bacteria and Archaea: a divide and conquer approach. PLoS One *6*, e28388.

Hardison, R.C. (2003). Comparative genomics. PLoS Biol *1*, E58.

Hasegawa, H., and Holm, L. (2009). Advances and pitfalls of protein structural alignment. Current Opinion in Structural Biology *19*, 341-348.

Hasin, Y., Olender, T., Khen, M., Gonzaga-Jauregui, C., Kim, P.M., Urban, A.E., Snyder, M., Gerstein, M.B., Lancet, D., and Korbel, J.O. (2008). High-resolution copy-number variation map reflects human olfactory receptor diversity and evolution. PLoS Genet *4*, e1000249.

He, M., Wang, Y., and Li, W. (2009). PPI Finder: A Mining Tool for Human Protein-Protein Interactions. PLoS One *4*.

Heger, A., and Ponting, C.P. (2007). Evolutionary rate analyses of orthologs and paralogs from 12 Drosophila genomes. Genome Res *17*, 1837-1849.

Heimberg, A.M., Cowper-Sal-lari, R., Semon, M., Donoghue, P.C., and Peterson, K.J. (2010). microRNAs reveal the interrelationships of hagfish, lampreys, and gnathostomes and the nature of the ancestral vertebrate. Proc Natl Acad Sci U S A *107*, 19379-19383.

Heinicke, S., Livstone, M.S., Lu, C., Oughtred, R., Kang, F., Angiuoli, S.V., White, O., Botstein, D., and Dolinski, K. (2007). The Princeton Protein Orthology Database (P-POD): A Comparative Genomics Analysis Tool for Biologists. PLoS One *2*.

Hellmuth, M., Hernandez-Rosales, M., Huber, K.T., Moulton, V., Stadler, P.F., and Wieseke, N. (2012). Orthology relations, symbolic ultrametrics, and cographs. J Math Biol.

Henikoff, J.G., Greene, E.A., Taylor, N., Henikoff, S., and Pietrokovski, S. (2002). Using the blocks database to recognize functional domains. Curr Protoc Bioinformatics *Chapter 2*, Unit 2 2.

Hertel, J., Lindemeyer, M., Missal, K., Fried, C., Tanzer, A., Flamm, C., Hofacker, I.L., and Stadler, P.F. (2006). The expansion of the metazoan microRNA repertoire. BMC Genomics *7*, 25.

Ho, M.R., Chen, C.H., and Lin, W.C. (2010). Gene-oriented ortholog database: a functional comparison platform for orthologous loci. Database-the Journal of Biological Databases and Curation.

Hoeppner, M.P., White, S., Jeffares, D.C., and Poole, A.M. (2009). Evolutionarily Stable Association of Intronic snoRNAs and microRNAs with Their Host Genes. Genome Biology and Evolution *1*, 420-428.

Hubbard, T.J., Aken, B.L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T.*, et al.* (2007). Ensembl 2007. Nucleic Acids Res *35*, D610-617.

Huelsenbeck, J.P., Bollback, J.P., and Levine, A.M. (2002). Inferring the root of a phylogenetic tree. Syst Biol *51*, 32-43.

Huerta-Cepas, J., Capella-Gutierrez, S., Pryszcz, L.P., Denisov, I., Kormes, D., Marcet-Houben, M., and Gabaldon, T. (2011). PhylomeDB v3.0: an expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. Nucleic Acids Research *39*, D556-D560.

Huerta-Cepas, J., Dopazo, H., Dopazo, J., and Gabaldon, T. (2007). The human phylome. Genome Biology *8*.

Huttenhofer, A., Brosius, J., and Bachellerie, J.P. (2002). RNomics: identification and function of small, non-messenger RNAs. Current Opinion in Chemical Biology *6*, 835-843.

Huynen, M.A., and Bork, P. (1998). Measuring genome evolution. Proceedings of the National Academy of Sciences of the United States of America *95*, 5849-5856.

Hwang, D., Rust, A.G., Ramsey, S., Smith, J.J., Leslie, D.M., Weston, A.D., Atauri, P.D., Aitchison, J.D., Hood, L., Siegel, A.F.*, et al.* (2005). A data integration methodology for systems biology. Proceedings of the National Academy of Sciences of the United States of America *102*, 17296-17301.

Hyduke, D.R., and Palsson, B.O. (2010). Towards genome-scale signalling network reconstructions. Nat Rev Genet *11*, 297-307.

Ideker, T., Galitski, T., and Hood, L. (2001). A new approach to decoding life: systems biology. Annu Rev Genomics Hum Genet *2*, 343-372.

Ideker, T., and Krogan, N.J. (2012). Differential network biology. Mol Syst Biol *8*, 565.

Irimia, M., Tena, J.J., Alexis, M., Fernandez-Minan, A., Maeso, I., Bogdanovic, O., de la Calle-Mustienes, E., Roy, S.W., Gomez-Skarmeta, J.L., and Fraser, H.B. (2012). Extensive conservation of ancient microsynteny across metazoans due to cis-regulatory constraints. Genome Res.

Ivanic, J., Yu, X., Wallqvist, A., and Reifman, J. (2009). Influence of protein abundance on high-throughput protein-protein interaction detection. PLoS One *4*, e5815.

Iyer, L.M., Balaji, S., Koonin, E.V., and Aravind, L. (2006). Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. Virus Res *117*, 156-184.

Jablonka, E., and Raz, G. (2009). Transgenerational epigenetic inheritance: prevalence, mechanisms, and implications for the study of heredity and evolution. Q Rev Biol *84*, 131-176.

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bulletin de la Société Vaudoise des Sciences Naturelles *37*.

Jager, S., Gulbahce, N., Cimermancic, P., Kane, J., He, N., Chou, S., D'Orso, I., Fernandes, J., Jang, G., Frankel, A.D.*, et al.* (2011). Purification and characterization of HIV-human protein complexes. Methods *53*, 13-19.

Janga, S.C., Diaz-Mejia, J.J., and Moreno-Hagelsieb, G. (2011). Network-based function prediction and interactomics: The case for metabolic enzymes. Metabolic Engineering *13*, 1-10.

Jensen, L.J., Julien, P., Kuhn, M., von Mering, C., Muller, J., Doerks, T., and Bork, P. (2008). eggNOG: automated construction and annotation of orthologous groups of genes. Nucleic Acids Res *36*, D250-254.

Jeong, H., Mason, S.P., Barabasi, A.L., and Oltvai, Z.N. (2001). Lethality and centrality in protein networks. Nature *411*, 41-42.

Jia, Y.Z., Wong, T.K.F., Song, Y.Q., Yiu, S.M., and Smith, D.K. (2010). Refining orthologue groups at the transcript level. BMC Genomics *11*.

Jiang, Z. (2008). Protein function predictions based on the phylogenetic profile method. Crit Rev Biotechnol *28*, 233-238.

Jothi, R., Zotenko, E., Tasneem, A., and Przytycka, T.M. (2006). COCO-CL: hierarchical clustering of homology relations based on evolutionary correlations. Bioinformatics *22*, 779-788.

Juan, D., Pazos, F., and Valencia, A. (2008). High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. Proceedings of the National Academy of Sciences of the United States of America *105*, 934-939.

Julien, P., Brawand, D., Soumillon, M., Necsulea, A., Liechti, A., Schutz, F., Daish, T., Grutzner, F., and Kaessmann, H. (2012). Mechanisms and evolutionary patterns of mammalian and avian dosage compensation. PLoS Biol *10*, e1001328.

Kaake, R.M., Wang, X.R., and Huang, L. (2010). Profiling of Protein Interaction Networks of Protein Complexes Using Affinity Purification and Quantitative Mass Spectrometry. Molecular & Cellular Proteomics *9*, 1650-1665.

Kaessmann, H. (2010). Origins, evolution, and phenotypic impact of new genes. Genome Res *20*, 1313-1326.

Kalaev, M., Smoot, M., Ideker, T., and Sharan, R. (2008). NetworkBLAST: comparative analysis of protein networks. Bioinformatics *24*, 594-596.

Kalendar, R., Flavell, A.J., Ellis, T.H., Sjakste, T., Moisy, C., and Schulman, A.H. (2011). Analysis of plant diversity with retrotransposon-based molecular markers. Heredity (Edinb) *106*, 520-530.

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research *28*, 27-30.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res *40*, D109-114.

Kasprzyk, A. (2011). BioMart: driving a paradigm change in biological data management. Database-the Journal of Biological Databases and Curation.

Katoh, K., Kuma, K., Toh, H., and Miyata, T. (2005). MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res *33*, 511-518.

Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res *30*, 3059-3066.

Katoh, K., and Toh, H. (2008a). Recent developments in the MAFFT multiple sequence alignment program. Brief Bioinform *9*, 286-298.

Katoh, K., and Toh, H. (2008b). Recent developments in the MAFFT multiple sequence alignment program. Briefings in Bioinformatics *9*, 286-298.

Keator, D.B., Grethe, J.S., Marcus, D., Ozyurt, B., Gadde, S., Murphy, S., Pieper, S., Greve, D., Notestine, R., Bockholt, H.J.*, et al.* (2008). A national human neuroimaging collaboratory enabled by the Biomedical Informatics Research Network (BIRN). IEEE Trans Inf Technol Biomed *12*, 162-172.

Keeling, P.J. (2009). Functional and ecological impacts of horizontal gene transfer in eukaryotes. Curr Opin Genet Dev *19*, 613-619.

Keeling, P.J., Burger, G., Durnford, D.G., Lang, B.F., Lee, R.W., Pearlman, R.E., Roger, A.J., and Gray, M.W. (2005). The tree of eukaryotes. Trends Ecol Evol *20*, 670-676.

Keeling, P.J., and Palmer, J.D. (2008). Horizontal gene transfer in eukaryotic evolution. Nature Reviews Genetics *9*, 605-618.

Kelder, T., van Iersel, M.P., Hanspers, K., Kutmon, M., Conklin, B.R., Evelo, C.T., and Pico, A.R. (2012). WikiPathways: building research communities on biological pathways. Nucleic Acids Res *40*, D1301-1307.

Kelley, B.P., Yuan, B., Lewitter, F., Sharan, R., Stockwell, B.R., and Ideker, T. (2004). PathBLAST: a tool for alignment of protein interaction networks. Nucleic Acids Res *32*, W83-88.

Kendrew, J.C. (1961). The three-dimensional structure of a protein molecule. Sci Am *205*, 96-110.

Kent, W.J., Baertsch, R., Hinrichs, A., Miller, W., and Haussler, D. (2003). Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. Proc Natl Acad Sci U S A *100*, 11484-11489.

Kew, O.M., Wright, P.F., Agol, V.I., Delpeyroux, F., Shimizu, H., Nathanson, N., and Pallansch, M.A. (2004). Circulating vaccine-derived polioviruses: current state of knowledge. Bull World Health Organ *82*, 16-23.

Khaitovich, P., Enard, W., Lachmann, M., and Paabo, S. (2006). Evolution of primate gene expression. Nature Reviews Genetics *7*, 693-702.

Khalil, A.S., and Collins, J.J. (2010). Synthetic biology: applications come of age. Nat Rev Genet *11*, 367-379.

Kim, K., Kim, W., and Kim, S. (2011). ReMark: an automatic program for clustering orthologs flexibly combining a Recursive and a Markov clustering algorithms. Bioinformatics *27*, 1731-1733.

Kim, T.M., and Park, P.J. (2011). Advances in analysis of transcriptional regulatory networks. Wiley Interdisciplinary Reviews-Systems Biology and Medicine *3*, 21-35.

King, M.C., and Wilson, A.C. (1975). Evolution at two levels in humans and chimpanzees. Science *188*, 107-116.

Kitano, H. (2002). Computational systems biology. Nature *420*, 206-210.

Klamt, S., Haus, U.U., and Theis, F. (2009). Hypergraphs and Cellular Networks. Plos Computational Biology *5*.

Kohonen, T. (1988). Self-organization and associative memory (Springer-Verlag).

Koonin, E.V. (2005). Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet *39*, 309-338.

Koonin, E.V. (2009a). Darwinian evolution in the light of genomics. Nucleic Acids Res *37*, 1011-1034.

Koonin, E.V. (2009b). On the Origin of Cells and Viruses Primordial Virus World Scenario. Natural Genetic Engineering and Natural Genome Editing *1178*, 47-64.

Koonin, E.V., Makarova, K.S., and Aravind, L. (2001). Horizontal gene transfer in prokaryotes: quantification and classification. Annu Rev Microbiol *55*, 709-742.

Koonin, E.V., and Wolf, Y.I. (2006). Evolutionary systems biology: links between gene evolution and function. Current Opinion in Biotechnology *17*, 481-487.

Koonin, E.V., and Wolf, Y.I. (2010). Constraints and plasticity in genome and molecular-phenome evolution. Nat Rev Genet *11*, 487-498.

Kornberg, H.L. (1987). Krebs Citric-Acid Cycle - Half a Century and Still Turning - Introductory. Biochemical Society Symposium, 1-2.

Krishnan, N., Fu, C.X., Pappin, D.J., and Tonks, N.K. (2011). H2S-Induced Sulfhydration of the Phosphatase PTP1B and Its Role in the Endoplasmic Reticulum Stress Response. Science Signaling *4*.

Kristensen, D.M., Wolf, Y.I., Mushegian, A.R., and Koonin, E.V. (2011). Computational methods for Gene Orthology inference. Briefings in Bioinformatics *12*, 379-391.

Kriventseva, E.V., Rahman, N., Espinosa, O., and Zdobnov, E.M. (2008). OrthoDB: the hierarchical catalog of eukaryotic orthologs. Nucleic Acids Res *36*, D271-275.

Krivoruchko, A., Siewers, V., and Nielsen, J. (2011). Opportunities for yeast metabolic engineering: Lessons from synthetic biology. Biotechnol J *6*, 262-276.

Kruger, F.A., and Overington, J.P. (2012). Global analysis of small molecule binding to related protein targets. PLoS Comput Biol *8*, e1002333.

Ku, C.S., Cooper, D.N., Polychronakos, C., Naidoo, N., Wu, M.C., and Soong, R. (2012). Exome sequencing: Dual role as a discovery and diagnostic tool. Annals of Neurology *71*, 5-14.

Kuchaiev, O., Milenkovic, T., Memisevic, V., Hayes, W., and Przulj, N. (2010). Topological network alignment uncovers biological function and phylogeny. J R Soc Interface *7*, 1341-1354.

Kuzniar, A., Lin, K., He, Y., Nijveen, H., Pongor, S., and Leunissen, J.A.M. (2009). ProGMap: an integrated annotation resource for protein orthology. Nucleic Acids Research *37*, W428-W434.

Kuzniar, A., van Ham, R.C.H.J., Pongor, S., and Leunissen, J.A.M. (2008). The quest for orthologs: finding the corresponding gene across genomes. Trends in Genetics *24*, 539-551.

Laird, P.W. (2010). Principles and challenges of genome-wide DNA methylation analysis. Nature Reviews Genetics *11*, 191-203.

Lamberts, S.W., and Uitterlinden, A.G. (2009). Genetic testing in clinical practice. Annu Rev Med *60*, 431-442.

Lander, E.S. (2011). Initial impact of the sequencing of the human genome. Nature *470*, 187-197.

Larhlimi, A., Blachon, S., Selbig, J., and Nikoloski, Z. (2011). Robustness of metabolic networks: A review of existing definitions. Biosystems *106*, 1-8.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., *et al.* (2007). Clustal W and Clustal X version 2.0. Bioinformatics *23*, 2947-2948.

Larkum, A.W., Ross, I.L., Kruse, O., and Hankamer, B. (2012). Selection, breeding and engineering of microalgae for bioenergy and biofuel production. Trends Biotechnol *30*, 198-205.

Le, S., Josse, J., and Husson, F. (2008). FactoMineR: An R package for multivariate analysis. Journal of Statistical Software *25*, 1-18.

Lechner, M., Findeiss, S., Steiner, L., Marz, M., Stadler, P.F., and Prohaska, S.J. (2011). Proteinortho: detection of (co-)orthologs in large-scale analysis. BMC Bioinformatics *12*, 124.

Lecompte, O., Poch, O., and Laporte, J. (2008). Ptdlns5P regulation through evolution: roles in membrane trafficking? Trends in Biochemical Sciences *33*, 453-460.

Lecompte, O., Thompson, J.D., Plewniak, F., Thierry, J., and Poch, O. (2001). Multiple alignment of complete sequences (MACS) in the post-genomic era. Gene *270*, 17-30.

Lee, D., Smallbone, K., Dunn, W.B., Murabito, E., Winder, C.L., Kell, D.B., Mendes, P., and Swainston, N. (2012). Improving metabolic flux predictions using absolute gene expression data. BMC Syst Biol *6*, 73.

Lee, Y., Sultana, R., Pertea, G., Cho, J., Karamycheva, S., Tsai, J., Parvizi, B., Cheung, F., Antonescu, V., White, J.*, et al.* (2002). Cross-referencing eukaryotic genomes: TIGR orthologous gene alignments (TOGA). Genome Research *12*, 493-502.

Lehner, B., Crombie, C., Tischler, J., Fortunato, A., and Fraser, A.G. (2006). Systematic mapping of genetic interactions in Caenorhabditis elegans identifies common modifiers of diverse signaling pathways. Nat Genet *38*, 896-903.

Letunic, I., and Bork, P. (2007). Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics *23*, 127-128.

Levasseur, A., Paganini, J., Dainat, J., Thompson, J.D., Poch, O., Pontarotti, P., and Gouret, P. (2012a). The chordate proteome history database. Evol Bioinform Online *8*, 437-447.

Levasseur, A., Paganini, J., Dainat, J., Thompson, J.D., Poch, O., Pontarotti, P., and Gouret, P. (2012b). The Chordate Proteome History Database. Evolutionary Bioinformatics *8*, 437.

Levasseur, A., Pontarotti, P., Poch, O., and Thompson, J.D. (2008). Strategies for reliable exploitation of evolutionary concepts in high throughput biology. Evol Bioinform Online *4*, 121-137.

Levine, M., and Davidson, E.H. (2005). Gene regulatory networks for development. Proc Natl Acad Sci U S A *102*, 4936-4942.

Levskaya, A., Weiner, O.D., Lim, W.A., and Voigt, C.A. (2009). Spatiotemporal control of cell signalling using a light-switchable protein interaction. Nature *461*, 997-1001.

Li, L., Stoeckert, C.J., Jr., and Roos, D.S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res *13*, 2178-2189.

Li, S.M., Armstrong, C.M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P.O., Han, J.D.J., Chesneau, A., Hao, T.*, et al.* (2004). A map of the interactome network of the metazoan C-elegans. Science *303*, 540-543.

Li, S.P., Tseng, J.J., and Wang, S.C. (2005). Reconstructing gene regulatory networks from time-series microarray data. Physica a-Statistical Mechanics and Its Applications *350*, 63-69.

Li, W.F., Shi, W.F., Qiao, H.J., Ho, S.Y.W., Luo, A.R., Zhang, Y.Z., and Zhu, C.D. (2011). Positive selection on hemagglutinin and neuraminidase genes of H1N1 influenza viruses. Virology Journal *8*.

Liberles, D.A., Teichmann, S.A., Bahar, I., Bastolla, U., Bloom, J., Bornberg-Bauer, E., Colwell, L.J., de Koning, A.P., Dokholyan, N.V., Echave, J.*, et al.* (2012). The interface of protein structure, protein biophysics, and molecular evolution. Protein Sci *21*, 769-785.

Lima-Mendez, G., and van Helden, J. (2009). The powerful law of the power law and other myths in network biology. Mol Biosyst *5*, 1482-1493.

Linard, B., Thompson, J.D., Poch, O., and Lecompte, O. (2011). OrthoInspector: comprehensive orthology analysis and visual exploration. BMC Bioinformatics *12*.

Liu, J., and Kipreos, E.T. (2000). Evolution of cyclin-dependent kinases (CDKs) and CDK-activating kinases (CAKs): differential conservation of CAKs in yeast and metazoa. Mol Biol Evol *17*, 1061-1074.

Loewe, L. (2009). A framework for evolutionary systems biology. Bmc Systems Biology *3*.

Lozada-Chavez, I., Janga, S.C., and Collado-Vides, J. (2006). Bacterial regulatory networks are extremely flexible in evolution. Nucleic Acids Res *34*, 3434-3445.

Lynch, M., and Force, A. (2000). The probability of duplicate gene preservation by subfunctionalization. Genetics *154*, 459-473.

Ma, H., Sorokin, A., Mazein, A., Selkov, A., Selkov, E., Demin, O., and Goryanin, I. (2007). The Edinburgh human metabolic network reconstruction and its functional analysis. Mol Syst Biol *3*, 135.

Maere, S., De Bodt, S., Raes, J., Casneuf, T., Van Montagu, M., Kuiper, M., and Van de Peer, Y. (2005). Modeling gene and genome duplications in eukaryotes. Proc Natl Acad Sci U S A *102*, 5454-5459.

Mahmood, K., Konagurthu, A.S., Song, J., Buckle, A.M., Webb, G.I., and Whisstock, J.C. (2010). EGM: encapsulated gene-by-gene matching to identify gene orthologs and homologous segments in genomes. Bioinformatics *26*, 2076-2084.

Mao, X., Cai, T., Olyarchuk, J.G., and Wei, L. (2005). Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. Bioinformatics *21*, 3787-3793.

Marco, D. (2011). Metagenomics: Current Innovations and Future Trends (Caister Academic Press).

Matsuya, A., Sakate, R., Kawahara, Y., Koyanagi, K.O., Sato, Y., Fujii, Y., Yamasaki, C., Habara, T., Nakaoka, H., Todokoro, F*., et al.* (2008). Evola: Ortholog database of all human genes in H-InvDB with manual curation of phylogenetic trees. Nucleic Acids Research *36*, D787-D792.

Maury, S., Trap-Gentil, M.V., Hebrard, C., Weyens, G., Delaunay, A., Barnes, S., Lefebvre, M., and Joseph, C. (2012). Genic DNA methylation changes during in vitro organogenesis: organ specificity and conservation between parental lines of epialleles. Physiol Plant.

McGinnis, S., and Madden, T.L. (2004). BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Research *32*, W20-W25.

McPherson, J.D., Marra, M., Hillier, L., Waterston, R.H., Chinwalla, A., Wallis, J., Sekhon, M., Wylie, K., Mardis, E.R., Wilson, R.K*., et al.* (2001). A physical map of the human genome. Nature *409*, 934-941.

Melendez-Hevia, E. (2009). From the RNA world to the DNA-protein world: clues to the origin and early evolution of life in the ribosome. J Biosci *34*, 825-827.

Menon, R., and Farina, C. (2011). Shared Molecular and Functional Frameworks among Five Complex Human Disorders: A Comparative Study on Interactomes Linked to Susceptibility Genes. PLoS One *6*.

Mi, H.Y., Dong, Q., Muruganujan, A., Gaudet, P., Lewis, S., and Thomas, P.D. (2010). PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. Nucleic Acids Research *38*, D204-D210.

Michalak, P. (2008). Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. Genomics *91*, 243-248.

Mika, S., and Rost, B. (2006). Protein-protein interactions more conserved within species than across species. Plos Computational Biology *2*, 698-709.

Milinkovitch, M.C., Helaers, R., Depiereux, E., Tzika, A.C., and Gabaldon, T. (2010). 2x genomes--depth does matter. Genome Biol *11*, R16.

Millard, B.L., Niepel, M., Menden, M.P., Muhlich, J.L., and Sorger, P.K. (2011). Adaptive informatics for multifactorial and high-content biological data. Nature Methods *8*, 487-U2255.

Misawa, N. (2011). Pathway engineering for functional isoprenoids. Curr Opin Biotechnol *22*, 627-633.

Missiuro, P.V., Liu, K.S., Zou, L.H., Ross, B.C., Zhao, G.Y., Liu, J.S., and Ge, H. (2009). Information Flow Analysis of Interactome Networks. Plos Computational Biology *5*.

Mizushima, S.I., Simanouti, T., and et al. (1949). Stable configurations of a polypeptide chain. Nature *164*, 918.

Mogensen, T.H. (2009). Pathogen recognition and inflammatory signaling in innate immune defenses. Clin Microbiol Rev *22*, 240-273, Table of Contents.

Mohn, F., and Schubeler, D. (2009). Genetics and epigenetics: stability and plasticity during cellular differentiation. Trends in Genetics *25*, 129-136.

Moore, N.L., Narayanan, R., and Weigel, N.L. (2007). Cyclin dependent kinase 2 and the regulation of human progesterone receptor activity. Steroids *72*, 202-209.

Moret, B.M.E., Roshan, U., and Warnow, T. (2002). Sequence-length requirements for phylogenetic methods. Algorithms in Bioinformatics, Proceedings *2452*, 343-356.

Morgan, D.K., and Whitelaw, E. (2008). The case for transgenerational epigenetic inheritance in humans. Mammalian Genome *19*, 394-397.

Morgenstern, B. (2007). Alignment of genomic sequences using DIALIGN. Methods Mol Biol *395*, 195-204.

Morgenstern, B., Frech, K., Dress, A., and Werner, T. (1998). DIALIGN: Finding local similarities by multiple sequence alignment. Bioinformatics *14*, 290-294.

Morris, K.V. (2012). Non-Coding RNAs and Epigenetic Regulation of Gene Expression: Drivers of Natural Selection (Caister Academic Press).

Moult, J., Fidelis, K., Rost, B., Hubbard, T., and Tramontano, A. (2005). Critical assessment of methods of protein structure prediction (CASP)--round 6. Proteins *61 Suppl 7*, 3-7.

Muffato, M., and Crollius, H.R. (2008). Paleogenomics in vertebrates, or the recovery of lost genomes from the mist of time. Bioessays *30*, 122-134.

Neal, M.L., and Kerckhoffs, R. (2010). Current progress in patient-specific modeling. Briefings in Bioinformatics *11*, 111-126.

Nehrt, N.L., Clark, W.T., Radivojac, P., and Hahn, M.W. (2011). Testing the ortholog conjecture with comparative functional genomic data from mammals. PLoS Comput Biol *7*, e1002073.

Neuwald, A.F., Liu, J.S., Lipman, D.J., and Lawrence, C.E. (1997). Extracting protein alignment models from the sequence database. Nucleic Acids Research *25*, 1665-1677.

Ng, P.C., Levy, S., Huang, J., Stockwell, T.B., Walenz, B.P., Li, K., Axelrod, N., Busam, D.A., Strausberg, R.L., and Venter, J.C. (2008). Genetic Variation in an Individual Human Exome. Plos Genetics *4*.

Nguyen, H.N., Wicker, N., Kieffer, D., and Poch, O. (2009). A new projection method for biological semantic map generation. J Biomedical Science and Engineering *3*, **13-19**.

Nothnagel, M., Herrmann, A., Wolf, A., Schreiber, S., Platzer, M., Siebert, R., Krawczak, M., and Hampe, J. (2011). Technology-specific error signatures in the 1000 Genomes Project data. Hum Genet *130*, 505-516.

O'Connell, L.A., and Hofmann, H.A. (2011). Genes, hormones, and circuits: An integrative approach to study the evolution of social behavior. Frontiers in Neuroendocrinology *32*, 320-335.

O'Donovan, C., Apweiler, R., and Bairoch, A. (2001). The human proteomics initiative (HPI). Trends Biotechnol *19*, 178-181.

Oberst, A., Bender, C., and Green, D.R. (2008). Living with death: the evolution of the mitochondrial pathway of apoptosis in animals. Cell Death Differ *15*, 1139-1146.

Ochman, H., Lawrence, J.G., and Groisman, E.A. (2000). Lateral gene transfer and the nature of bacterial innovation. Nature *405*, 299-304.

Ogata, H., Fujibuchi, W., Goto, S., and Kanehisa, M. (2000). A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters. Nucleic Acids Res *28*, 4021-4028.

Ohno, S., Wolf, U., and Atkin, N.B. (1968). Evolution from fish to mammals by gene duplication. Hereditas *59*, 169-187.

Ooi, H.S., Schneider, G., Lim, T.T., Chan, Y.L., Eisenhaber, B., and Eisenhaber, F. (2010). Biomolecular pathway databases. Methods Mol Biol *609*, 129-144.

Osbourn, A.E., and Field, B. (2009). Operons. Cell Mol Life Sci *66*, 3755-3775.

Ostlund, G., Schmitt, T., Forslund, K., Kostler, T., Messina, D.N., Roopra, S., Frings, O., and Sonnhammer, E.L.L. (2010). InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Research *38*, D196-D203.

Page, R.D.M. (1994). Maps between Trees and Cladistic-Analysis of Historical Associations among Genes, Organisms, and Areas. Systematic Biology *43*, 58-77.

Panchenko, A., and Przytycka, T.M. (2008). Protein-protein Interactions and Networks: Identification, Computer Analysis, and Prediction (Springer).

Papp, B., Notebaart, R.A., and Pal, C. (2011). Systems-biology approaches for predicting genomic evolution. Nat Rev Genet *12*, 591-602.

Park, D., Singh, R., Baym, M., Liao, C.S., and Berger, B. (2011). IsoBase: a database of functionally related proteins across PPI networks. Nucleic Acids Res *39*, D295-300.

Parks, D.H., and Beiko, R.G. (2010). Identifying biologically relevant differences between metagenomic communities. Bioinformatics *26*, 715-721.

Patrick, T. (2000). Darwin et la science de l'évolution. Gallimard découvertes.

Patti, G.J., Yanes, O., and Siuzdak, G. (2012). Innovation: Metabolomics: the apogee of the omics trilogy. Nat Rev Mol Cell Biol *13*, 263-269.

Paulsen, I., and von Haeseler, A. (2006). INVHOGEN: a database of homologous invertebrate genes. Nucleic Acids Research *34*, D349-D353.

Pazos, F., and Valencia, A. (2008). Protein co-evolution, co-adaptation and interactions. EMBO J *27*, 2648-2655.

Pei, J., Kim, B.H., and Grishin, N.V. (2008). PROMALS3D: a tool for multiple protein sequence and structure alignments. Nucleic Acids Res *36*, 2295-2300.

Pellegrini, M. (2012). Using phylogenetic profiles to predict functional relationships. Methods Mol Biol *804*, 167-177.

Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., and Yeates, T.O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A *96*, 4285-4288.

Penel, S., Arigon, A.M., Dufayard, J.F., Sertier, A.S., Daubin, V., Duret, L., Gouy, M., and Perriere, G. (2009). Databases of homologous gene families for comparative genomics. BMC Bioinformatics *10*.

Penkett, C.J., Morris, J.A., Wood, V., and Bahler, J. (2006). YOGY: a web-based, integrated database to retrieve protein orthologs and associated Gene Ontology terms. Nucleic Acids Research *34*, W330-W334.

Penloglou, G., Chatzidoukas, C., and Kiparissides, C. (2012). Microbial production of polyhydroxybutyrate with tailor-made properties: an integrated modelling approach and experimental validation. Biotechnol Adv *30*, 329-337.

Pennisi, E. (2008). Evolution. Modernizing the modern synthesis. Science *321*, 196-197.

Perkins, J.R., Diboun, I., Dessailly, B.H., Lees, J.G., and Orengo, C. (2010). Transient Protein-Protein Interactions: Structural, Functional, and Network Properties. Structure *18*, 1233-1243.

Peterson, M.E., Chen, F., Saven, J.G., Roos, D.S., Babbitt, P.C., and Sali, A. (2009). Evolutionary constraints on structural similarity in orthologs and paralogs. Protein Sci *18*, 1306-1315.

Pevsner, J. (2009a). Bioinformatics and Functional Genomics (Wiley-Blackwell).

Pevsner, J. (2009b). Bioinformatics and functional genomics. Wiley-Blackwell Press.

Pfeiffer, T., Soyer, O.S., and Bonhoeffer, S. (2005). The evolution of connectivity in metabolic networks. PLoS Biol *3*, e228.

Philippe, H., and Douady, C.J. (2003). Horizontal gene transfer and phylogenetics. Curr Opin Microbiol *6*, 498-505.

Pigliucci, M. (2007). Do we need an extended evolutionary synthesis? Evolution *61*, 2743-2749.

Plewniak, F., Bianchetti, L., Brelivet, Y., Carles, A., Chalmel, F., Lecompte, O., Mochel, T., Moulinier, L., Muller, A., Muller, J.*, et al.* (2003). PipeAlign: A new toolkit for protein family analysis. Nucleic Acids Res *31*, 3829-3832.

Pollard, K.S., Salama, S.R., Lambert, N., Lambot, M.A., Coppens, S., Pedersen, J.S., Katzman, S., King, B., Onodera, C., Siepel, A.*, et al.* (2006). An RNA gene expressed during cortical development evolved rapidly in humans. Nature *443*, 167-172.

Ponting, C.P., Oliver, P.L., and Reik, W. (2009). Evolution and functions of long noncoding RNAs. Cell *136*, 629-641.

Pop, M., and Salzberg, S.L. (2008). Bioinformatics challenges of new sequencing technology. Trends Genet *24*, 142-149.

Portnoy, V.A., Bezdan, D., and Zengler, K. (2011). Adaptive laboratory evolution--harnessing the power of biology for metabolic engineering. Curr Opin Biotechnol *22*, 590-594.

Postlethwait, J.H. (2007). The zebrafish genome in context: ohnologs gone missing. J Exp Zool B Mol Dev Evol *308*, 563-577.

Powell, S., Szklarczyk, D., Trachana, K., Roth, A., Kuhn, M., Muller, J., Arnold, R., Rattei, T., Letunic, I., Doerks, T.*, et al.* (2012). eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. Nucleic Acids Research *40*, D284-D289.

Prosdocimi, F., Linard, B., Pontarotti, P., Poch, O., and Thompson, J.D. (2012). Controversies in modern evolutionary biology: the imperative for error detection and quality control. BMC Genomics *13*, 5.

Pryszcz, L.P., Huerta-Cepas, J., and Gabaldon, T. (2011). MetaPhOrs: orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score. Nucleic Acids Research *39*.

Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J.*, et al.* (2012a). The Pfam protein families database. Nucleic Acids Res *40*, D290-301.

Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J.*, et al.* (2012b). The Pfam protein families database. Nucleic Acids Research *40*, D290-D301.

Pybus, O.G., and Rambaut, A. (2009). Evolutionary analysis of the dynamics of viral infectious disease. Nature Reviews Genetics *10*, 540-550.

Raffelsberger, W., Krause, Y., Moulinier, L., Kieffer, D., Morand, A.L., Brino, L., and Poch, O. (2008). RReportGenerator: automatic reports from routine statistical analysis using R. Bioinformatics *24*, 276-278.

Ramachandran, G.N., Ramakrishnan, C., and Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. J Mol Biol *7*, 95-99.

Raman, K. (2010). Construction and analysis of protein-protein interaction networks. Autom Exp *2*, 2.

Rascol, V.L., Pontarotti, P., and Levasseur, A. (2007). Ancestral animal genomes reconstruction. Curr Opin Immunol *19*, 542-546.

Rebl, A., Goldammer, T., and Seyfert, H.M. (2010). Toll-like receptor signaling in bony fish. Vet Immunol Immunopathol *134*, 139-150.

Reumann, S. (2011). Toward a definition of the complete proteome of plant peroxisomes: Where experimental proteomics must be complemented by bioinformatics. Proteomics *11*, 1764-1779.

Rohwer, J.M. (2012). Kinetic modelling of plant metabolic pathways. Journal of Experimental Botany *63*, 2275-2292.

Ros, V.I., and Hurst, G.D. (2009). Lateral gene transfer between prokaryotes and multicellular eukaryotes: ongoing and significant? BMC Biol *7*, 20.

Roth, A.C., Gonnet, G.H., and Dessimoz, C. (2008). Algorithm of OMA for large-scale orthology inference. BMC Bioinformatics *9*, 518.

Rothschild, L.J. (2010). A powerful toolkit for synthetic biology: Over 3.8 billion years of evolution. Bioessays *32*, 304-313.

Rouard, M., Guignon, V., Aluome, C., Laporte, M.A., Droc, G., Walde, C., Zmasek, C.M., Perin, C., and Conte, M.G. (2011). GreenPhylDB v2.0: comparative and functional genomics in plants. Nucleic Acids Research *39*, D1095-D1102.

Rual, J.F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G.F., Gibbons, F.D., Dreze, M., Ayivi-Guedehoussou, N.*, et al.* (2005). Towards a proteome-scale map of the human protein-protein interaction network. Nature *437*, 1173-1178.

Ruan, J., Li, H., Chen, Z.Z., Coghlan, A., Coin, L.J.M., Guo, Y., Heriche, J.K., Hu, Y.F., Kristiansen, K., Li, R.Q.*, et al.* (2008). TreeFam: 2008 update. Nucleic Acids Research *36*, D735-D740.

Ruano-Rubio, V., Poch, O., and Thompson, J.D. (2009). Comparison of eukaryotic phylogenetic profiling approaches using species tree aware methods. BMC Bioinformatics *10*, 383.

Rumpho, M.E., Worful, J.M., Lee, J., Kannan, K., Tyler, M.S., Bhattacharya, D., Moustafa, A., and Manhart, J.R. (2008). Horizontal gene transfer of the algal nuclear gene psbO to the photosynthetic sea slug Elysia chlorotica. Proc Natl Acad Sci U S A *105*, 17867-17871.

Rustici, M., and Lesk, A.M. (1994). Three-dimensional searching for recurrent structural motifs in data bases of protein structures. J Comput Biol *1*, 121-132.

Sacan, A., Toroslu, I.H., and Ferhatosmanoglu, H. (2008). Integrated search and alignment of protein structures. Bioinformatics *24*, 2872-2879.

Salichos, L., and Rokas, A. (2011). Evaluating Ortholog Prediction Algorithms in a Yeast Model Clade. PLoS One *6*.

Sanger, F., and Coulson, A.R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol *94*, 441-448.

Sankoff, D. (1999). Genome rearrangement with gene families. Bioinformatics *15*, 909-917.

Sayers, E.W., Barrett, T., Benson, D.A., Bolton, E., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., Dicuccio, M., Federhen, S.*, et al.* (2012). Database resources of the National Center for Biotechnology Information. Nucleic Acids Res *40*, D13-25.

Scannell, D.R., Frank, A.C., Conant, G.C., Byrne, K.P., Woolfit, M., and Wolfe, K.H. (2007). Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication. Proceedings of the National Academy of Sciences of the United States of America *104*, 8397-8402.

Schaefer, C.F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., and Buetow, K.H. (2009). PID: the Pathway Interaction Database. Nucleic Acids Research *37*, D674-D679.

Schmitt, T., Messina, D.N., Schreiber, F., and Sonnhammer, E.L. (2011). Letter to the editor: SeqXML and OrthoXML: standards for sequence and orthology information. Brief Bioinform *12*, 485-488.

Schwartz, S., Kent, W.J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R.C., Haussler, D., and Miller, W. (2003). Human-mouse alignments with BLASTZ. Genome Res *13*, 103-107.

Schwikowski, B., Uetz, P., and Fields, S. (2000). A network of protein-protein interactions in yeast. Nat Biotechnol *18*, 1257-1261.

Seal, R.L., Gordon, S.M., Lush, M.J., Wright, M.W., and Bruford, E.A. (2011). genenames.org: the HGNC resources in 2011. Nucleic Acids Research *39*, D514-D519.

Secrier, M., Pavlopoulos, G.A., Aerts, J., and Schneider, R. (2012). Arena3D: visualizing time-driven phenotypic differences in biological systems. BMC Bioinformatics *13*.

Seret, M.L., and Baret, P.V. (2011). IONS: Identification of Orthologs by Neighborhood and Similarity- an Automated Method to Identify Orthologs in Chromosomal Regions of Common Evolutionary Ancestry and its Application to Hemiascomycetous Yeasts. Evol Bioinform Online *7*, 123-133.

Shapira, S.D., Gat-Viks, I., Shum, B.O., Dricot, A., de Grace, M.M., Wu, L., Gupta, P.B., Hao, T., Silver, S.J., Root, D.E.*, et al.* (2009). A physical and regulatory map of host-influenza interactions reveals pathways in H1N1 infection. Cell *139*, 1255-1267.

Sharan, R., Suthram, S., Kelley, R.M., Kuhn, T., McCuine, S., Uetz, P., Sittler, T., Karp, R.M., and Ideker, T. (2005). Conserved patterns of protein interaction in multiple species. Proc Natl Acad Sci U S A *102*, 1974-1979.

Shea, N., Pen, I., and Uller, T. (2011). Three epigenetic information channels and their different roles in evolution. Journal of Evolutionary Biology *24*, 1178-1187.

Shealy, P., and Valafar, H. (2012). Multiple structure alignment with msTALI. BMC Bioinformatics *13*, 105.

Shi, G., Zhang, L., and Jiang, T. (2010). MSOAR 2.0: Incorporating tandem duplications into ortholog assignment based on genome rearrangement. BMC Bioinformatics *11*, 10.

Shi, G.Q., Peng, M.C., and Jiang, T. (2011). MultiMSOAR 2.0: An Accurate Tool to Identify Ortholog Groups among Multiple Genomes. PLoS One *6*.

Shih, Y.K., and Parthasarathy, S. (2012). Scalable global alignment for multiple biological networks. BMC Bioinformatics *13 Suppl 3*, S11.

Shin, J.W., Park, S.H., and Kang, Y.G. (2012). Potential of engineering methodologies for the application to pharmaceutical research. Arch Pharm Res *35*, 299-309.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J.*, et al.* (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol *7*, 539.

Sigrist, C.J., Cerutti, L., de Castro, E., Langendijk-Genevaux, P.S., Bulliard, V., Bairoch, A., and Hulo, N. (2010). PROSITE, a protein domain database for functional characterization and annotation. Nucleic Acids Res *38*, D161-166.

Simillion, C., Vandepoele, K., and Van de Peer, Y. (2004). Recent developments in computational approaches for uncovering genomic homology. Bioessays *26*, 1225-1235.

Singh, R., Xu, J., and Berger, B. (2008). Global alignment of multiple protein interaction networks with application to functional orthology detection. Proc Natl Acad Sci U S A *105*, 12763-12768.

Singh, R., Xu, J.B., and Berger, B. (2007). Pairwise global alignment of protein interaction networks by matching neighborhood topology. Research in Computational Molecular Biology, Proceedings *4453*, 16-31.

Sjolander, K., Datta, R.S., Shen, Y., and Shoffner, G.M. (2011). Ortholog identification in the presence of domain architecture rearrangement. Brief Bioinform *12*, 413-422.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J.*, et al.* (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology *25*, 1251-1255.

Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.L., and Ideker, T. (2011). Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics *27*, 431-432.

Soares, E.V., and Soares, H.M. (2012). Bioremediation of industrial effluents containing heavy metals using brewing cells of Saccharomyces cerevisiae as a green technology: a review. Environ Sci Pollut Res Int *19*, 1066-1083.

Springer, M.S., Meredith, R.W., Janecka, J.E., and Murphy, W.J. (2011). The historical biogeography of Mammalia. Philos Trans R Soc Lond B Biol Sci *366*, 2478-2502.

Srinivasan, B.S., Novak, A.F., Flannick, J.A., Batzoglou, S., and McAdams, H.H. (2006). Integrated protein interaction networks for 11 microbes. Research in Computational Molecular Biology, Proceedings *3909*, 1-14.

Stobbe, M.D., Houten, S.M., Jansen, G.A., van Kampen, A.H.C., and Moerland, P.D. (2011). Critical assessment of human metabolic pathway databases: a stepping stone for future integration. Bmc Systems Biology *5*.

Stobbe, M.D., Houten, S.M., van Kampen, A.H., Wanders, R.J., and Moerland, P.D. (2012). Improving the description of metabolic networks: the TCA cycle as example. FASEB J.

Storm, C.E., and Sonnhammer, E.L. (2002). Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. Bioinformatics *18*, 92-99.

Storm, C.E.V., and Sonnhammer, E.L.L. (2003). Comprehensive analysis of orthologous protein domains using the HOPS database. Genome Research *13*, 2353-2362.

Straus, C., Vasilakos, K., Wilson, R.J.A., Oshima, T., Zelter, M., Derenne, J.P., Similowski, T., and Whitelaw, W.A. (2003). A phylogenetic hypothesis for the origin of hiccough. Bioessays *25*, 182-188.

Sundin, G.W. (2007). Genomic insights into the contribution of phytopathogenic bacterial Plasmids to the evolutionary history of their hosts. Annual Review of Phytopathology *45*, 129-151.

Tafforeau, L., Rabourdin-Combe, C., and Lotteau, V. (2012). Virus-human cell interactomes. Methods Mol Biol *812*, 103-120.

Tang, H. (2007). Genome assembly, rearrangement, and repeats. Chem Rev *107*, 3391-3406.

Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N.*, et al.* (2003). The COG database: an updated version includes eukaryotes. BMC Bioinformatics *4*, 41.

Tatusov, R.L., Koonin, E.V., and Lipman, D.J. (1997). A genomic perspective on protein families. Science *278*, 631-637.

Tautz, D. (1998). Evolutionary biology. Debatable homologies. Nature *395*, 17, 19.

Taylor, J.S., and Raes, J. (2004). Duplication and divergence: the evolution of new genes and old ideas. Annu Rev Genet *38*, 615-643.

Tekaia, F., and Yeramian, E. (2012). SuperPartitions: detection and classification of orthologs. Gene *492*, 199-211.

Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L., Angiuoli, S.V., Crabtree, J., Jones, A.L., Durkin, A.S.*, et al.* (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". Proc Natl Acad Sci U S A *102*, 13950-13955.

Thomas, D., Becker, A., and Surdin-Kerjan, Y. (2000). Reverse methionine biosynthesis from S-adenosylmethionine in eukaryotic cells. J Biol Chem *275*, 40718-40724.

Thomas, P.D., Wood, V., Mungall, C.J., Lewis, S.E., Blake, J.A., and Consortium, G.O. (2012). On the Use of Gene Ontology Annotations to Assess Functional Similarity among Orthologs and Paralogs: A Short Report. Plos Computational Biology *8*.

Thompson, J.D., Gibson, T.J., and Higgins, D.G. (2002). Multiple sequence alignment using ClustalW and ClustalX. Curr Protoc Bioinformatics *Chapter 2*, Unit 2 3.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994). Clustal-W - Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice. Nucleic Acids Research *22*, 4673-4680.

Thompson, J.D., Holbrook, S.R., Katoh, K., Koehl, P., Moras, D., Westhof, E., and Poch, O. (2005a). MAO: a multiple alignment ontology for nucleic acid and protein sequences. Nucleic Acids Research *33*, 4164-4171.

Thompson, J.D., Holbrook, S.R., Katoh, K., Koehl, P., Moras, D., Westhof, E., and Poch, O. (2005b). MAO: a Multiple Alignment Ontology for nucleic acid and protein sequences. Nucleic Acids Res *33*, 4164-4171.

Thompson, J.D., Linard, B., Lecompte, O., and Poch, O. (2011). A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives. PLoS One *6*, e18093.

Thompson, J.D., Muller, A., Waterhouse, A., Procter, J., Barton, G.J., Plewniak, F., and Poch, O. (2006). MACSIMS: multiple alignment of complete sequences information management system. BMC Bioinformatics *7*, 318.

Thompson, J.D., Plewniak, F., Ripp, R., Thierry, J.C., and Poch, O. (2001). Towards a reliable objective function for multiple sequence alignments. J Mol Biol *314*, 937-951.

Thompson, J.D., Prigent, V., and Poch, O. (2004). LEON: multiple aLignment Evaluation Of Neighbours. Nucleic Acids Res *32*, 1298-1307.

Thompson, J.D., Thierry, J.C., and Poch, O. (2003). RASCAL: rapid scanning and correction of multiple sequence alignments. Bioinformatics *19*, 1155-1161.

Timmis, J.N., Ayliffe, M.A., Huang, C.Y., and Martin, W. (2004). Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nat Rev Genet *5*, 123-135.

Towfic, F., VanderPlas, S., Oliver, C.A., Couture, O., Tuggle, C.K., West Greenlee, M.H., and Honavar, V. (2010). Detection of gene orthology from gene co-expression and protein interaction networks. BMC Bioinformatics *11 Suppl 3*, S7.

Trachana, K., Larsson, T.A., Powell, S., Chen, W.H., Doerks, T., Muller, J., and Bork, P. (2011). Orthology prediction methods: A quality assessment using curated protein families. Bioessays *33*, 769-780.

Tranchevent, L.C., Barriot, R., Yu, S., Van Vooren, S., Van Loo, P., Coessens, B., De Moor, B., Aerts, S., and Moreau, Y. (2008). ENDEAVOUR update: a web resource for gene prioritization in multiple species. Nucleic Acids Res *36*, W377-384.

Tranchevent, L.C., Capdevila, F.B., Nitsch, D., De Moor, B., De Causmaecker, P., and Moreau, Y. (2011). A guide to web tools to prioritize candidate genes. Brief Bioinform *12*, 22-32.

Uchiyama, I. (2006). Hierarchical clustering algorithm for comprehensive orthologous-domain classification in multiple genomes. Nucleic Acids Res *34*, 647-658.

Uchiyama, I., Higuchi, T., and Kawai, M. (2010). MBGD update 2010: toward a comprehensive resource for exploring microbial genome diversity. Nucleic Acids Research *38*, D361-D365.

Ulitsky, I., Shkumatava, A., Jan, C.H., Sive, H., and Bartel, D.P. (2011). Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. Cell *147*, 1537-1550.

Van Bel, M., Proost, S., Wischnitzki, E., Movahedi, S., Scheerlinck, C., Van de Peer, Y., and Vandepoele, K. (2012). Dissecting Plant Genomes with the PLAZA Comparative Genomics Platform. Plant Physiology *158*, 590-600.

van Dam, T.J., Zwartkruis, F.J., Bos, J.L., and Snel, B. (2011). Evolution of the TOR pathway. J Mol Evol *73*, 209-220.

van der Heijden, R.T.J.M., Snel, B., van Noort, V., and Huynen, M.A. (2007). Orthology prediction at scalable resolution by phylogenetic tree analysis. BMC Bioinformatics *8*.

Vankadavath, R.N., Hussain, A.J., Bodanapu, R., Kharshiing, E., Basha, P.O., Gupta, S., Sreelakshmi, Y., and Sharma, R. (2009). Computer aided data acquisition tool for high-throughput phenotyping of plant populations. Plant Methods *5*.

Vasudevan, S., Seli, E., and Steitz, J.A. (2006). Metazoan oocyte and early embryo development program: a progression through translation regulatory cascades. Genes Dev *20*, 138-146.

Velankar, S., Alhroub, Y., Best, C., Caboche, S., Conroy, M.J., Dana, J.M., Fernandez Montecelo, M.A., van Ginkel, G., Golovin, A., Gore, S.P.*, et al.* (2012). PDBe: Protein Data Bank in Europe. Nucleic Acids Res *40*, D445-452.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A.*, et al.* (2001). The sequence of the human genome. Science *291*, 1304-1351.

Vidal, M., Cusick, M.E., and Barabasi, A.L. (2011). Interactome networks and human disease. Cell *144*, 986-998.

Vogel, C., Bashton, M., Kerrison, N.D., Chothia, C., and Teichmann, S.A. (2004). Structure, function and evolution of multidomain proteins. Curr Opin Struct Biol *14*, 208-216.

von Bertalanffy, L., and Woodger, J.H. (1933). Modern theories of development: an introduction to theoretical biology (Oxford university press, H. Milford).

Wall, D.P., Fraser, H.B., and Hirsh, A.E. (2003). Detecting putative orthologs. Bioinformatics *19*, 1710-1711.

Wallace, I.M., O'Sullivan, O., Higgins, D.G., and Notredame, C. (2006a). M-Coffee: combining multiple sequence alignment methods with T-Coffee. Nucleic Acids Research *34*, 1692-1699.

Wallace, I.M., O'Sullivan, O., Higgins, D.G., and Notredame, C. (2006b). M-Coffee: combining multiple sequence alignment methods with T-Coffee. Nucleic Acids Res *34*, 1692-1699.

Wang, J., Zhang, J., Zheng, H., Li, J., Liu, D., Li, H., Samudrala, R., Yu, J., and Wong, G.K. (2004). Mouse transcriptome: neutral evolution of 'non-coding' complementary DNAs. Nature *431*, 1 p following 757; discussion following 757.

Wang, L.S., Leebens-Mack, J., Kerr Wall, P., Beckmann, K., dePamphilis, C.W., and Warnow, T. (2011). The impact of multiple protein sequence alignment on phylogenetic estimation. IEEE/ACM Trans Comput Biol Bioinform *8*, 1108-1119.

Wang, X.J., Wei, X.M., Thijssen, B., Das, J., Lipkin, S.M., and Yu, H.Y. (2012). Three-dimensional reconstruction of protein networks provides insight into human genetic disease. Nature Biotechnology *30*, 159-164.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet *10*, 57-63.

Wapinski, I., Pfeffer, A., Friedman, N., and Regev, A. (2007). Automatic genome-wide reconstruction of phylogenetic gene trees. Bioinformatics *23*, i549-558.

Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M., and Barton, G.J. (2009). Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics *25*, 1189-1191.

Waterhouse, R.M., Zdobnov, E.M., Tegenfeldt, F., Li, J., and Kriventseva, E.V. (2011). OrthoDB: the hierarchical catalog of eukaryotic orthologs in 2011. Nucleic Acids Research *39*, D283-D288.

Watts, D.J., and Strogatz, S.H. (1998a). Collective dynamics of 'small-world' networks. Nature *393*, 440-442.

Watts, D.J., and Strogatz, S.H. (1998b). Collective dynamics of 'small-world' networks. Nature *393*, 440-442.

Weber, W., and Fussenegger, M. (2012). Emerging biomedical applications of synthetic biology. Nat Rev Genet *13*, 21-35.

Weiss, K.M., and Buchanan, A.V. (2011). Is Life Law-Like? Genetics *188*, 761-771.

Wenk, M.R. (2005). The emerging field of lipidomics. Nature Reviews Drug Discovery *4*, 594-610.

Weston, A.D., and Hood, L. (2004). Systems biology, proteomics, and the future of health care: toward predictive, preventative, and personalized medicine. J Proteome Res *3*, 179-196.

Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S.*, et al.* (2007). Database resources of the National Center for Biotechnology Information. Nucleic Acids Research *35*, D5-D12.

Whisstock, J.C., and Lesk, A.M. (2003). Prediction of protein function from protein sequence and structure. Q Rev Biophys *36*, 307-340.

Wicker, N., Perrin, G.R., Thierry, J.C., and Poch, O. (2001). Secator: a program for inferring protein subfamilies from phylogenetic trees. Mol Biol Evol *18*, 1435-1441.

Wilczynski, B., and Furlong, E.E. (2010). Challenges for modeling global gene regulatory networks during development: insights from Drosophila. Dev Biol *340*, 161-169.

Wilkinson, L. (2012a). Exact and Approximate Area-Proportional Circular Venn and Euler Diagrams. Ieee Transactions on Visualization and Computer Graphics *18*, 321-331.

Wilkinson, L. (2012b). Exact and approximate area-proportional circular Venn and Euler diagrams. IEEE Trans Vis Comput Graph *18*, 321-331.

Wilm, A., Higgins, D.G., and Notredame, C. (2008). R-Coffee: a method for multiple alignment of non-coding RNA. Nucleic Acids Res *36*, e52.

Wolf, Y.I., Carmel, L., and Koonin, E.V. (2006). Unifying measures of gene function and evolution. Proc Biol Sci *273*, 1507-1515.

Wolfe, K. (2000). Robustness - it's not where you think it is. Nature Genetics *25*, 3-4.

Wolstenholme, J.T., Rissman, E.F., and Connelly, J.J. (2011). The role of Bisphenol A in shaping the brain, epigenome and behavior. Horm Behav *59*, 296-305.

Worth, C.L., Gong, S., and Blundell, T.L. (2009). Structural and functional constraints in the evolution of protein families. Nat Rev Mol Cell Biol *10*, 709-720.

Wu, D.D., Irwin, D.M., and Zhang, Y.P. (2009). Molecular evolution of the keratin associated protein gene family in mammals, role in the evolution of mammalian hair (vol 8, pg 241, 2008). Bmc Evolutionary Biology *9*.

Wu, M., Chatterji, S., and Eisen, J.A. (2012). Accounting for alignment uncertainty in phylogenomics. PLoS One *7*, e30288.

Xiang, L.X., He, D., Dong, W.R., Zhang, Y.W., and Shao, J.Z. (2012). Deep sequencing-based transcriptome profiling analysis of bacteria-challenged Lateolabrax japonicus reveals insight into the immune-relevant genes in marine fish. BMC Genomics *11*, 472.

Xie, X.H., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., Lander, E.S., and Kellis, M. (2005). Systematic discovery of regulatory motifs in human promoters and 3 ' UTRs by comparison of several mammals. Nature *434*, 338-345.

Xu, F., Zhao, C., Li, Y., Li, J., Deng, Y., and Shi, T. (2011). Exploring virus relationships based on virus-host protein-protein interaction network. BMC Syst Biol *5 Suppl 3*, S11.

Xu, J., and Li, Y. (2006). Discovering disease-genes by topological features in human protein-protein interaction network. Bioinformatics *22*, 2800-2805.

Yamada, T., and Bork, P. (2009). Evolution of biomolecular networks: lessons from metabolic and protein interactions. Nat Rev Mol Cell Biol *10*, 791-803.

Yamamoto, S., Asanuma, T., Takagi, T., and Fukuda, K.I. (2004). The molecule role ontology: an ontology for annotation of signal transduction pathway molecules in the scientific literature. Comp Funct Genomics *5*, 528-536.

Yang, J.M., and Tung, C.H. (2006). Protein structure database search and evolutionary classification. Nucleic Acids Research *34*, 3646-3659.

Yang, Z., and Rannala, B. (2012). Molecular phylogenetics: principles and practice. Nat Rev Genet *13*, 303-314.

Yu, C.G., Zavaljevski, N., Desai, V., and Reifman, J. (2011). QuartetS: a fast and accurate algorithm for large-scale orthology detection. Nucleic Acids Research *39*.

Yu, H., Braun, P., Yildirim, M.A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N.*, et al.* (2008). High-quality binary protein interaction map of the yeast interactome network. Science *322*, 104-110.

Yu, H.Y., Luscombe, N.M., Lu, H.X., Zhu, X.W., Xia, Y., Han, J.D.J., Bertin, N., Chung, S., Vidal, M., and Gerstein, M. (2004). Annotation transfer between genomes: Protein-protein interologs and protein-DNA regulogs. Genome Research *14*, 1107-1118.

Yu, N., Seo, J., Rho, K., Jang, Y., Park, J., Kim, W.K., and Lee, S. (2012). hiPathDB: a human-integrated pathway database with facile visualization. Nucleic Acids Res *40*, D797-802.

Yuan, Y.P., Eulenstein, O., Vingron, M., and Bork, P. (1998). Towards detection of orthologues in sequence databases. Bioinformatics *14*, 285-289.

Zaslaver, A., Baugh, L.R., and Sternberg, P.W. (2011). Metazoan Operons Accelerate Recovery from Growth-Arrested States. Cell *145*, 981-992.

Zeeberg, B.R., Feng, W., Wang, G., Wang, M.D., Fojo, A.T., Sunshine, M., Narasimhan, S., Kane, D.W., Reinhold, W.C., Lababidi, S*., et al.* (2003). GoMiner: a resource for biological interpretation of genomic and proteomic data. Genome Biol *4*, R28.

Zhang, Y., Szustakowski, J., and Schinke, M. (2009). Bioinformatics analysis of microarray data. Methods Mol Biol *573*, 259-284.

Zhang, Z., and Gerstein, M. (2003). Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements. J Biol *2*, 11.

Zhang, Z., and Zhang, J. (2009). A big world inside small-world networks. PLoS One *4*, e5686.

Zhu, X., Gerstein, M., and Snyder, M. (2007). Getting connected: analysis and principles of biological networks. Genes Dev *21*, 1010-1024.

Zimek, A., and Weber, K. (2005). Terrestrial vertebrates have two keratin gene clusters; striking differences in teleost fish. European Journal of Cell Biology *84*, 623-635.

Zmasek, C.M., and Eddy, S.R. (2001). A simple algorithm to infer gene duplication and speciation events on a gene tree. Bioinformatics *17*, 821-828.

Zmasek, C.M., and Eddy, S.R. (2002). RIO: Analyzing proteomes by automated phylogenomics using resampled inference of orthologs. BMC Bioinformatics *3*.

Zurita, M., and Merino, C. (2003). The transcriptional complexity of the TFIIH complex. Trends Genet *19*, 578-584.

ANNEXES

Annex I : publication n°3

Annex II : publication n°4

Annex III : publication n°5

Annex IV : publication n°6

Annex V : awards

Annex I :

Publication n° 3

**BMC
Genomics**

RESEARCH ARTICLE   **Open Access**

Controversies in modern evolutionary biology: the imperative for error detection and quality control

Francisco Prosdocimi[1,2†], Benjamin Linard[1†], Pierre Pontarotti[3], Olivier Poch[1] and Julie D Thompson[1*]

Abstract

Background: The data from high throughput genomics technologies provide unique opportunities for studies of complex biological systems, but also pose many new challenges. The shift to the genome scale in evolutionary biology, for example, has led to many interesting, but often controversial studies. It has been suggested that part of the conflict may be due to errors in the initial sequences. Most gene sequences are predicted by bioinformatics programs and a number of quality issues have been raised, concerning DNA sequencing errors or badly predicted coding regions, particularly in eukaryotes.

Results: We investigated the impact of these errors on evolutionary studies and specifically on the identification of important genetic events. We focused on the detection of asymmetric evolution after duplication, which has been the subject of controversy recently. Using the human genome as a reference, we established a reliable set of 688 duplicated genes in 13 complete vertebrate genomes, where significantly different evolutionary rates are observed. We estimated the rates at which protein sequence errors occur and are accumulated in the higher-level analyses. We showed that the majority of the detected events (57%) are in fact artifacts due to the putative erroneous sequences and that these artifacts are sufficient to mask the true functional significance of the events.

Conclusions: Initial errors are accumulated throughout the evolutionary analysis, generating artificially high rates of event predictions and leading to substantial uncertainty in the conclusions. This study emphasizes the urgent need for error detection and quality control strategies in order to efficiently extract knowledge from the new genome data.

Keywords: gene duplication, asymmetric evolution, gene prediction, error detection, quality control

Background

High throughput genomics technologies are now providing the raw data for genome-level or systems-level studies [1]. At the same time, the avalanche of data also poses many new challenges. The shift to genome scale studies in evolutionary biology, for instance, has led to many interesting, but often controversial studies. Many branches in the Tree of Life are still the subject of intense discussions, and simply adding more sequences has not resolved the inconsistencies [2]. In prokaryotes,

phylogenetic incongruencies are often assumed to be the result of lateral gene transfers, but the frequency of these events has been challenged recently [3,4]. In eukaryotes, the ancestral relationships between the major eukaryotic kingdoms [5-8], as well as many more recent clades such as fish or mammalian [9-11], are also hotly debated. It has been suggested that at least some of the conflicting results from evolutionary analyses are due to differences in the models and methodologies used to test the original hypotheses, e.g. [12,13], as well as errors in the input sequences [2].

High throughput biological datasets are notoriously incomplete [14-16], noisy and inconsistent and DNA or protein sequences are no exception. The DNA sequences produced by next generation sequencing

* Correspondence: Julie.Thompson@igbmc.fr
† Contributed equally
[1]Department of Integrated Structural Biology, IGBMC (Institut de Génétique et de Biologie Moléculaire et Cellulaire) CNRS/INSERM/Université de Strasbourg, 1 rue Laurent Fries, Illkirch, F-67404, France
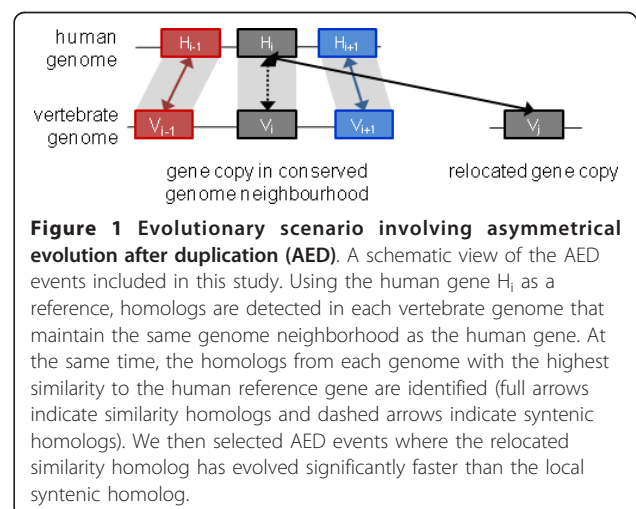Full list of author information is available at the end of the article

(NGS) technologies or low-coverage assemblies pose particular problems [17,18]. A number of recent studies have investigated the rate of errors in these new genome sequences and their impact on the accuracy of downstream analyses [19-22]. In the context of proteome studies, the DNA sequencing errors are further confounded by inaccuracies in the delineation of the protein-coding genes. Coding regions are mostly predicted by automatic methods, but the relationship between genes, transcripts and proteins is complex and automated genome annotation is not completely accurate. Thus, ten years after the publication of the human genome, the exact number of human protein-coding genes is still unknown [23]. Furthermore, recent analyses have shown that, even for those genes that have been identified, the complete exon/intron structure is correctly predicted for only about 50-60% of them [24-26]. In eukaryotic genomes, the situation is also complicated by widespread alternative splicing events, which affects more than 92-94% of multi-exon human genes [27].

To what extent do these quality issues affect our understanding of the evolutionary events shaping modern organisms? Although sequence errors are essentially ignored in most genome-scale analyses, some studies have addressed certain aspects of this question. For example, Hubisz and coworkers [19] investigated the impact of DNA sequencing errors in low-coverage genome assemblies on inferred rates and patterns of insertion/deletion and substitution on the mammalian phylogeny. Schneider et al. [28] showed that the estimated amount of positively selected genes in genome scale analyses may be inflated by the presence of unreliable sequences.

Here, we have investigated the impact of erroneous protein sequences, resulting from either DNA sequencing errors or inaccurate prediction of exon/intron structures, on evolutionary analyses and the detection of important genetic events. We concentrated specifically on duplication events, which are known to be an important source of functional diversity [29-32] and where there has been a great deal of debate about the long term fate of duplicated genes. Two main models have been proposed for the evolution of novel gene function associated with gene duplication. The neofunctionalization model predicts the evolution of a new function in one of the duplicates, with accelerated evolution of the deconstrained copy compared with the copy that retains the ancestral function. The subfunctionalization model implies the division of the ancestral functions among the duplicates and does not make any prediction about the symmetry or asymmetry of sequence evolution. Although individual cases of both modes of evolution have been reported, the relative frequency of the different scenarios in nature is still not clear [12,33,34].

To some extent, the evolutionary fate of duplicated genes depends on the duplication mechanism. After tandem duplications or large-scale (e.g. whole-chromosome or whole-genome) duplications, both gene copies retain the same genome context. In contrast, after segmental duplications or retrotranspositions, one of the gene copies retains the ancestral genome position while the other copy is relocated elsewhere. It is generally expected that the gene copy that retains the genome context will be more conserved, and thus will be more likely to retain the ancestral functions [35]. The hypothesis is that newly duplicated genes that have been transposed to new chromosomal locations experience a new genomic and epigenetic environment, modifying the expression and/or function of the genes.

In this work, we have searched for duplication events that contradict this hypothesis, in order to quantify the effect of protein sequence errors on our ability to accurately identify unusual evolutionary histories. The goal was not to identify an exhaustive list of duplications, but to establish a reliable test set of events that could be used for the error analysis. Using the well-studied human genome as a reference, we identified 114,680 homologs in 13 high coverage vertebrate genomes from the Ensembl [36] database that were located in a region with local synteny (Figure 1). We then identified 688 cases where another homolog of the reference human gene was found elsewhere in the vertebrate genome with significantly higher sequence similarity than the syntenic homolog. In other words, we identified 688 gene triplets, composed of one human reference gene and two corresponding gene copies from another vertebrate genome (the local "syntenic homolog" and the remote "highest similarity homolog"), that might indicate putative asymmetric evolution after duplication (AED) events where the less similar gene copy retained

Figure 1 Evolutionary scenario involving asymmetrical evolution after duplication (AED). A schematic view of the AED events included in this study. Using the human gene H_i as a reference, homologs are detected in each vertebrate genome that maintain the same genome neighborhood as the human gene. At the same time, the homologs from each genome with the highest similarity to the human reference gene are identified (full arrows indicate similarity homologs and dashed arrows indicate syntenic homologs). We then selected AED events where the relocated similarity homolog has evolved significantly faster than the local syntenic homolog.
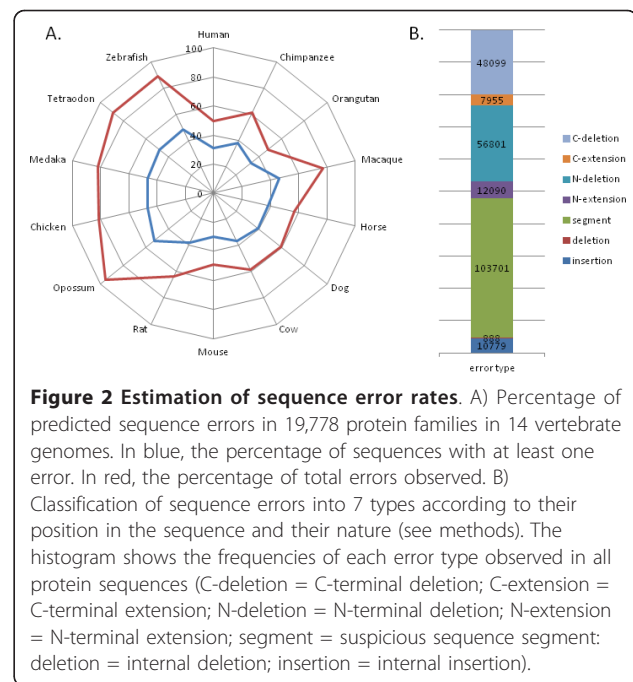
the ancestral gene-neighbourhood. To determine what proportion of these putative AED events may be due to erroneous protein sequences (resulting from either DNA sequencing errors or badly predicted protein coding regions), we identified potential sequence errors in the gene triplets and showed that the majority (57%) of detected AED events are in fact false positives. A Gene Ontology (GO) functional analysis highlighted a number of GO categories that are over-represented in the true positive gene set, which were masked before filtering of the erroneous sequences.

Results

Estimation of sequence error rates

We predicted protein sequence errors, resulting from genome sequencing errors and exon/intron prediction errors, in the 14 high coverage vertebrate genomes (Table 1) from the Ensembl database, using a previously published method [37]. First, we constructed multiple sequence alignments (MSAs) for each of the 19,778 human protein sequences defined by the Human Proteome Initiative (HPI) and their potential vertebrate homologs. The sequences in the alignments were then clustered into more similar subgroups and errors were predicted if discrepancies were observed between one sequence and its close neighbours, for example between human-chimpanzee or between fish genomes. The error detection protocol was thus used to identify lineage-specific insertions, deletions or sequence segments, which are inconsistent with the conservation information in the MSA. Finally, we calculated the rate of sequence errors found in all 19,778 MSAs (Figure 2A). The MSAs contained a total of 344,437 protein sequences and 240,313 potential sequence errors, giving an estimated

Figure 2 Estimation of sequence error rates. A) Percentage of predicted sequence errors in 19,778 protein families in 14 vertebrate genomes. In blue, the percentage of sequences with at least one error. In red, the percentage of total errors observed. B) Classification of sequence errors into 7 types according to their position in the sequence and their nature (see methods). The histogram shows the frequencies of each error type observed in all protein sequences (C-deletion = C-terminal deletion; C-extension = C-terminal extension; N-deletion = N-terminal deletion; N-extension = N-terminal extension; segment = suspicious sequence segment; deletion = internal deletion; insertion = internal insertion).

sequence error rate of at least 0.7 errors per sequence. The total number of sequences with at least one potential error was 142,836. Thus, on average 41% of sequences were predicted to be erroneous.

The observed error rates were not homogeneous across the different species. Lower rates were observed for the human and mouse proteomes, with 30-31% erroneous sequences, as might be expected for these well studied organisms. Among the non-human primate proteomes considered here, lower error rates were estimated for the orangutan (*Pongo pygmaeus*), compared

Table 1 Ensembl genomes used in this study

Genome identifier	Organism	No. of genes	No. of proteins
ENSP	'Human','Homo sapiens'	21971	60953
ENSPTR	'Chimpanzee','Pan troglodytes'	19829	39256
ENSPPY	'Orangutan','Pongo pygmaeus'	20068	29256
ENSMMU	'Macaque','Macaca mulatta'	21905	42370
ENSECA	'Horse','Equus caballus'	20322	28128
ENSCAF	'Dog','Canis familiaris'	19305	29804
ENSBTA	'Cow','Bos taurus'	21036	29517
ENSMUS	'Mouse','Mus musculus'	23873	43630
ENSRNO	'Rat','Rattus norvegicus'	22503	37672
ENSMOD	'Opossum','Monodelphis domestica'	19471	34132
ENSGAL	'Chicken','Gallus gallus'	16736	22945
ENSORL	'Medaka','Oryzias latipes'	19686	25174
ENSTNI	'Tetraodon','Tetraodon nigroviridis'	19602	23909
ENSDAR	'Zebrafish','Danio rerio'	21322	35967

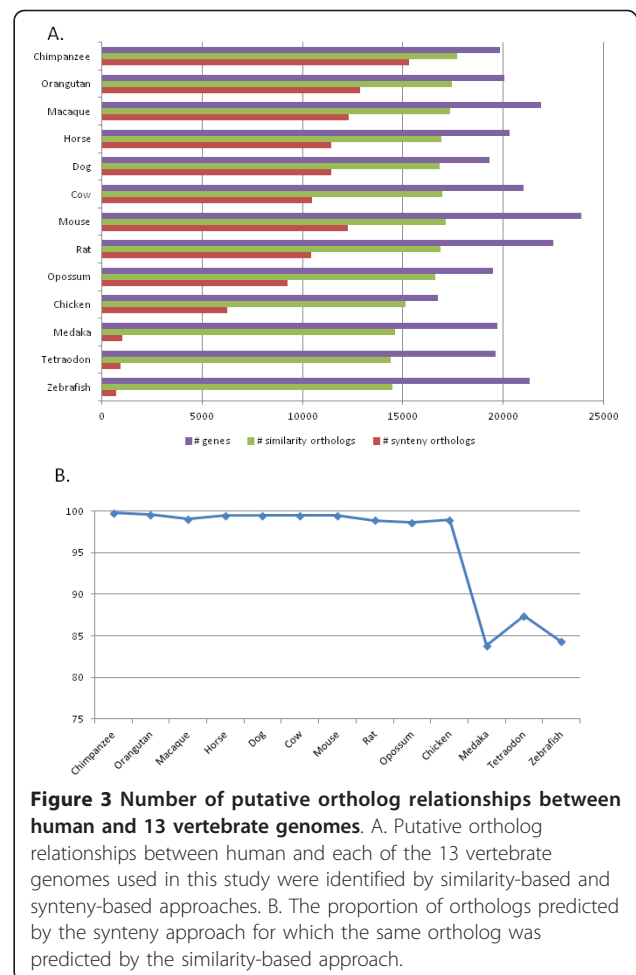Protein sequences were obtained from the Ensembl database version 51.

to the chimpanzee (*Pan troglodytes*) and especially the Rhesus macaque (*Macaca mulatta*). The relatively high error rate for the macaque is not surprising since the macaque genome in Ensembl version 51 is a preliminary assembly using whole genome shotgun (WGS) reads from small and medium insert clones. On the other hand, the relative error rates in chimpanzee and orangutan are more surprising. Both the chimpanzee and orangutan genomes have been sequenced to 6x coverage, but in a recent study of primate genome assembly quality, the chimpanzee genome assembly was estimated to be of higher quality [38].

Nevertheless, the same study found that about 70% of inferred errors in the orangutan genome were clustered in the 3.2% of the assembly that is of low quality, implying that > 96% of the assembly could be considered of high fidelity. We found the highest error rates in the opossum, chicken and fish proteomes, with > 45% erroneous sequences. Although these genomes have all been sequenced to high coverage, the lack of a well annotated reference genome from a closely related model organism may result in lower quality protein sequence prediction.

The predicted protein sequence errors were then characterized according to two different factors: (i) the nature of the error, i.e. insertion, deletion or suspicious segment and (ii) the position in the sequence, i.e. at the N/C-terminus or within the sequence. Figure 2B shows the proportion of the different errors observed. The most commonly found error was the presence of a suspicious sequence segment, possibly representing a mispredicted exon. At the N- and C- termini, deletions were observed more frequently than extensions. Although this may be due in part to the protocol used to detect sequence errors, it may also reflect the difficulty of predicting the first and last coding exons. In contrast, internal insertions were more common than internal deletions, suggesting that more internal errors were due to the over-prediction of introns as coding sequences, rather than the under-prediction of exons.

Comparison of similarity and synteny based homologs

Putative orthologs were predicted for each of the 19,778 human proteins based on the MSAs of the human reference sequences and related sequences from the 13 vertebrate genomes. Two different approaches were implemented. First, the sequences from each organism with the smallest evolutionary distance were identified based on pairwise alignments extracted from the MSAs, and denoted "highest similarity homologs". Second, "syntenic homologs" were defined based on the local gene order conservation. The genome coverage achieved by the two methods is shown in Figure 3 and Table S1 in Additional file 1. The highest similarity homologs covered 80% of the 265,658 genes in the 13 vertebrate

Figure 3 Number of putative ortholog relationships between human and 13 vertebrate genomes. A. Putative ortholog relationships between human and each of the 13 vertebrate genomes used in this study were identified by similarity-based and synteny-based approaches. B. The proportion of orthologs predicted by the synteny approach for which the same ortholog was predicted by the similarity-based approach.

genomes, ranging from 89% in chimpanzee to 68% in zebrafish. As expected, a smaller proportion (43%) of homologs was found with locally conserved synteny, including 77% of chimpanzee genes and only 3% of zebrafish. Although our definition of locally syntenic regions is relatively stringent, we observe a comparable coverage to other existing methods. For example, we found 51% of mouse genes to be syntenic with human, compared to 59% using the method developed by [39]. Other more refined methods have been developed, such as Syntenator [40], that use less stringent criteria to define conserved syntenic regions. By allowing more gene mismatches and gene insertions/deletions, Syntenator aligned 79% of mouse genes with human.

We then investigated whether the gene that is most similar on the sequence level is also the gene that shares the same gene-neighbourhood (Figure 3 and Table S2 in Additional file 1). Of the 212,409 similarity homologs identified in the 13 vertebrate genomes, 113,517 were found in locally syntenic regions. In mammals, this represents 69% of the highest similarity homologs. This is less than that estimated in a previous study [41],

where 97.5% of Inparanoid orthologs in human, mouse, rat and dog were found in syntenic regions, most likely due to our stricter definition of local synteny. On the other hand, only 1% of the identified syntenic homologs (1,157 out of 114,680) were not identified by the similarity-based approach. As expected, a generally higher level of disagreement was observed for more divergent genome pairs. Nevertheless, in human-chicken comparisons, the synteny method identified the same homolog as the similarity approach in 98.8% of the cases. Fewer consistencies were observed in human-fish comparisons (84-87% of syntenic homologs were also the highest similarity homologs), possibly due in part to the whole genome duplications in the fish lineage, resulting in a larger number of paralogs.

Asymmetric evolution events

We then examined in more detail the 1,157 gene triplets (consisting of the human reference sequence and the two homologs representing putative orthologs in one of the 13 vertebrate genomes), where the syntenic homolog was not the same as the highest similarity homolog. To avoid including chance outcomes caused by very similar rates of sequence evolution of these homologs relative to the human sequence, we identified significantly different rates of evolution at the 95% confidence level (see Methods). Of the 1,157 gene triplets, a total of 688 corresponded to evolutionary scenarios where the syntenic homolog (i.e. the gene copy with the shared genome neighbourhood) evolved significantly faster (Table 2). A complete list of the 688 gene triplets is available in Table S3 in Additional file 1. The alternative scenario

for asymmetric evolution where the remote copy evolved faster than the synteny copy is not detected by our protocol. since in this case the homologs defined by similarity and synteny would be the same.

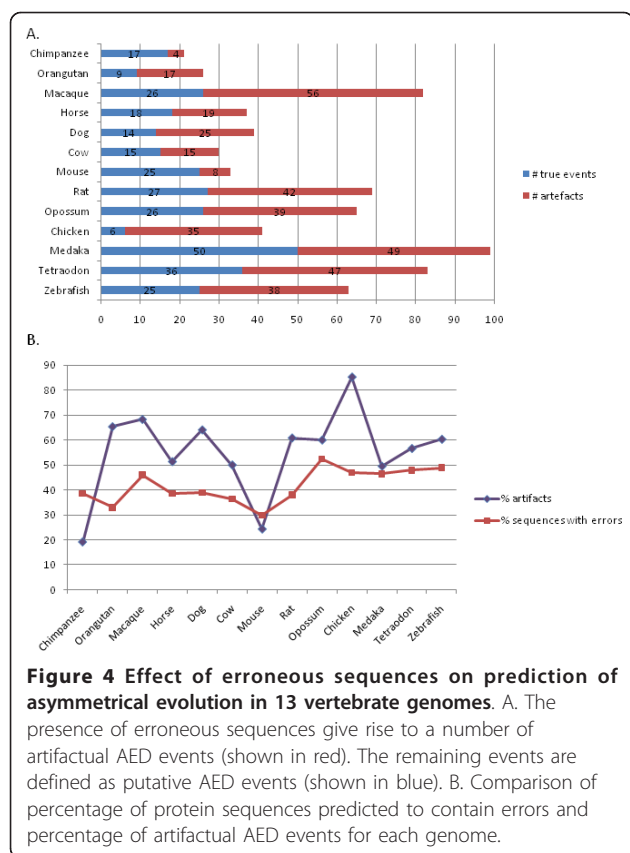### Effect of erroneous sequences on prediction of asymmetrical evolution

The 688 gene triplets identified above, consisting of the human reference sequence, the highest similarity homolog and the synteny homolog, constitute a reliable test set representing potential asymmetrical evolution events. To study the impact of errors on the prediction of AED events, we identified erroneous sequences in this test set. Figure 4A shows the number of events that are assumed to be artifacts since at least one of the sequences was predicted to be erroneous, as well as the number of remaining 'true' events. Of the 688 gene triplets, only 294 (43%) do not contain erroneous sequences and may correspond to true events, while a total of 394 (57%) are putative artifacts.

As might be expected, the proportion of artifactual events varies with the different genomes studied, depending on the percentage of erroneous sequence detected (Figure 4B). For example, 19% of chimpanzee and 24% of mouse predicted events are due to artifacts, while this figure increases significantly for the draft macaque and chicken genomes (69% and 88% respectively). It is interesting to note that a larger proportion of artifacts are observed in the orangutan genome than in the chimpanzee, even though the orangutan genome is predicted to contain less sequence errors than the chimpanzee (see above).

Table 2 Number of syntenic homologs with significantly faster evolutionary rates compared to the remote similarity homolog

Genome identifier	No. of syntenic homologs	No. of inconsistencies: syntenic versus highest similarity homologs	Significant asymmetric evolution events (AED)
Human	15295	37	21
Chimpanzee	12881	54	26
Orangutan	12286	121	82
Macaque	11447	59	37
Horse	11443	64	39
Dog	10486	59	30
Cow	12276	70	33
Mouse	10439	117	69
Rat	9261	126	65
Opossum	6231	65	41
Chicken	1027	166	99
Medaka	907	114	83
Tetraodon	701	111	63
Total	114680	1157	688

These may indicate putative asymmetric evolution after duplication (AED) events where the less similar gene copy retained the ancestral gene-neighbourhood.

Figure 4 Effect of erroneous sequences on prediction of asymmetrical evolution in 13 vertebrate genomes. A. The presence of erroneous sequences give rise to a number of artifactual AED events (shown in red). The remaining events are defined as putative AED events (shown in blue). B. Comparison of percentage of protein sequences predicted to contain errors and percentage of artifactual AED events for each genome.

In order to validate the putative protein sequence errors leading to artifactual AED events, we investigated the 413 predicted sequence errors in the human reference sequences and their syntenic homologs. The results of the analysis are shown in Table 3 and examples of the different errors detected are provided in Additional file 2. The majority (59%) of the erroneous sequences resulted from DNA sequencing or assembly errors, characterised by the presence of 'N' characters in the DNA sequences. For the remaining 171 protein sequence errors, we searched for the missing protein fragments in the corresponding DNA sequences. For errors involving missing segments (i.e. internal insertion, N/C-terminal extensions or suspicious segments), 89 of the 148 missing segments were detected and we therefore concluded that the error was due to an inaccurate gene structure prediction. In the case of sequence errors corresponding to inserted segments (internal insertions, N/C-terminal insertions), 16 of the 23 inserted segments were conserved in closely related organisms, although 5 of them had one or more stop codons. Finally, we manually verified the transcript evidence in Ensembl for all 23 insertions in gene sequences with no genome errors, as well as for the 59 unconserved deletions. Of these, 62 protein errors were not supported by any transcript information

and 9 errors were due to the alternative splicing variants reported for homologous genes. Only 11 (2.7%) of the 413 putative protein sequence errors were identified as false positive predictions, since a transcript was found corresponding to the affected sequence segment.

Detailed analysis of sequence errors leading to artifactual AED events

To investigate whether the sequence errors leading to artifactual events were enriched for a particular type, we classified the errors into 7 types as described above. We then calculated the proportion of the different error types found in the gene triplets corresponding to the 688 predicted AED events (Figure 5). In the human reference sequences, only 32 errors were predicted, as might be expected since the human genes have been very widely studied. The majority (24 out of 32) of the human sequence errors were found at the N/C termini, with the exception of a small number of internal sequence segments that were labeled as being suspicious.

When all the sequences in the gene triplets were pooled, no significant enrichment was observed in the frequency distribution of the different error types causing artifactual events, compared to the background distribution observed in all the sequences (as shown in Figure 2). The goodness-of-fit was measured using a likelihood ratio chi-square statistic (chi-square = 3.12, p-value = 0.79). Nevertheless, different error types were observed when the syntenic and highest similarity homologs were considered separately. For example, artifactual events were observed more frequently if the syntenic homolog, i.e. the gene copy that retained the genome neighbourhood after duplication, contained suspicious segments. In contrast, N- and C-deletions in the highest similarity homolog, i.e. the gene copy that was relocated, were more likely to cause artifacts.

Figure 6 shows an example of an artifactual event observed in the gene triplet corresponding to [Swiss-prot: COPG_HUMAN] and the two homologs from macaque (the full length alignment is provided in Figure S1 in Additional file 1). The COPG protein forms part of the coatomer complex, involved in protein transport between the endoplasmic reticulum and the Golgi. The macaque syntenic homolog [Ensembl:ENSMMUP00000017291] contains a suspicious segment and an exon deletion that artificially increase its evolutionary distance to human, due to a low quality segment in the genome sequence (indicated by 'N' characters in the gene sequence). Consequently, another macaque protein [Ensembl:ENSMMUP00000006382] is identified as the highest similarity homolog of human COPG, resulting in an artifactual AED event prediction. In fact, [Ensembl:ENSMMUP00000006382] is the ortholog of [Uniprot:COPG2_HUMAN].

Table 3 Validation of putative protein sequence errors

	Putative protein errors[a]	Genome errors[b]		Exon conservation[c]			Transcript evidence		% FP error[g]
		Yes	No	Yes	No	No	Splicing variants[e]	FP error prediction[f]	
Suspicious segment	223	161	62	43	19	12	3	4	1.8
Deletion	7	1	6	6	0	0	0	0	0.0
N-deletion	68	26	42	19	23	18	2	3	4.4
C-deletion	64	26	38	21	17	16	0	1	1.6
Deletion sub-total	362	214	148	89	59	46	5	8	2.9
	Putative protein errors	Genome errors		Intron conservation[d]			Transcript evidence		% FP error
		Yes	No	Yes (stop)	No	No	Splicing variants	FP error prediction	
Insertion	22	15	7	6 (1)	1	5	2	0	0.0
N-extension	18	7	11	7 (3)	4	7	1	3	16.7
C-extension	11	6	5	3 (1)	2	4	1	0	0.0
Insertion sub-total	51	28	23	16 (5)	7	16	4	3	5.9
Total	413	242	171	100	14	62	9	11	2.7

Putative errors were estimated by analyzing the corresponding gene sequences. [a]The total number of protein sequence errors included in the analysis. [b]The number of errors resulting from genome sequencing or assembly errors. [c]The number of missing segments detected in the corresponding gene sequences. [d]The number of errors resulting from alternative splicing variants reported for homologous genes. [e]The number of inserted sequence segments detected in the gene sequences of homologous proteins. The number of these inserted sequence segments with at least one stop codon is given in brackets. [f]The number of errors supported by transcript evidence, i.e. false positive (FP) error predictions. [g]The percentage of the total number of putative errors that were invalidated by the analysis.
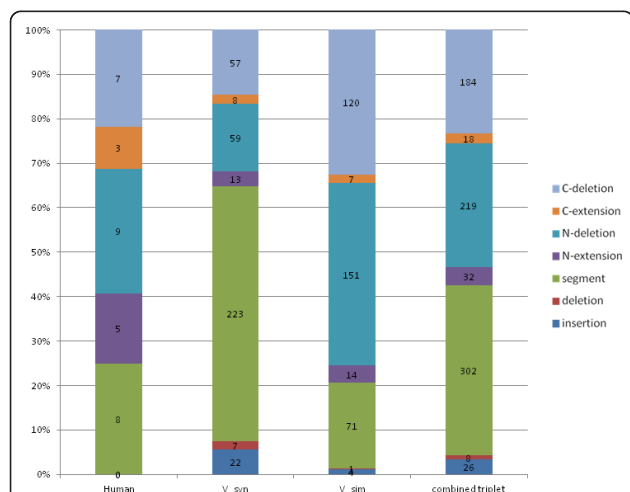
Figure 5 Characterization of sequence errors in predicted asymmetrical evolution events. Errors are classified into 7 types according to their position in the sequence and their nature (see methods). The proportions of the different classes found in the human reference sequences, the syntenic homolog (V_syn) and the highest similarity homolog (V_sim) are shown, as well as the proportions observed in the pooled sequences in the gene triplets. (C-deletion = C-terminal deletion; C-extension = C-terminal extension; N-deletion = N-terminal deletion; N-extension = N-terminal extension; segment = suspicious sequence segment: deletion = internal deletion; insertion = internal insertion).

The orthology prediction method used in the Ensembl project, based on a phylogenetic gene tree approach, finds the correct 1-to-1 orthology relationship between the human and macaque COPG proteins. Unfortunately, many other orthology databases are less successful. For example, in the Inparanoid database (inparanoid.sbc.su. se), the Ensembl human COPG and macaque COPG2 sequences are in the same orthologous cluster, while no human ortholog is found for the macaque COPG sequence.

Functional analysis of asymmetrical evolution events

In order to investigate the effect of filtering the erroneous sequences on the subsequent functional analysis of asymmetrical evolution events, we conducted a gene ontology (GO) term enrichment analysis. Specifically, we investigated the 688 AED events detected in this work, where the local syntenic homolog was observed to evolve more rapidly than the relocated highest similarity homolog. At this stage, we excluded 81 events where the human reference sequence had more than one exon, but the relocated homolog had only one exon, since they are likely to be non-functional pseudogenes. For comparison purposes, we used two gene lists: (i) gene list 1 corresponding to the remaining 607 detected

Figure 6 An example of an artifactual AED event. Part of the multiple sequence alignment of the human COPG protein sequence [Ensembl: ENSP00000325002] and putative orthologs in the macaque genome. The suspicious segment is boxed in grey. For the Ensembl macaque sequences, exons are colored alternately in black and blue. Residues overlapping splice sites are shown in red.

events, including both artifactual and putative true events and (ii) gene list 2 corresponding to 250 putative true events only (Table S4 in Additional file 1). The two gene lists were then analyzed for enrichment of GO terms using the AmiGO [42] web server, using the complete set of human genes as the background set and default parameters (Tables S5-6 in Additional file 1). The results of the AmiGO analyses were also submitted to the GO-Module [43] web server, in order to reduce the complexity and identify 'key' GO terms (Table 4).

Gene list 1 was enriched in 24 key GO terms, including a number of vertebrate specializations (e.g. anatomical structure development), but also some fundamental eukaryotic processes (e.g. regulation of metabolic processes, gene expression, axon guidance). For example, the term 'RNA biosynthetic process' is found with a P-value of 5E-16, involving 101 (20%) of the 607 genes in the list. However, only 6 of these 24 key GO terms are associated with the true events in gene list 2. Thus, the remaining 18 (75%) enriched GO terms are probably false positives resulting from the artifactual events. Furthermore, and perhaps more importantly, important key GO terms associated with the true events are not enriched in gene list 1, notably neurogenesis related functions. After filtering of gene triplets with erroneous sequences, gene set 2 was enriched in 10 key terms, including neuron differentiation functions, and response to the environment.

Figure 7 shows an example of a true AED event detected in the hepatoma-derived growth factor (HDGF) protein family. The HDGF and HDGF-like family members are characterized by a conserved PWWP domain in the N-terminal region. In human, the HDGF protein [Ensembl:ENSP00000349878] exhibits growth factor properties and has been implicated in organ development and tissue differentiation of the intestine, kidney, liver, and cardiovascular system. In addition, the role of HDGF in cancer biology has recently become a focus of

research, since HDGF was found to be over-expressed in a large number of different tumor types (genecards. org). Whereas some family members, such as HDGF and HDGFL2, are expressed in a wide range of tissues, the expression of others is very restricted. For example, HDGFL1 and HDGFL4 are only expressed in testis, although their precise functions are still unknown. We observed an EAD event in several organisms, including mouse and rat. For example, mouse HDGFL1 [Ensembl:ENSMUSP00000057557] on chromosome 13 is syntenic with human HDGFL1 [Ensembl: ENSP00000230012] on chromosome 6, but mouse HDGF [Ensembl:ENSMUSP00000005017] shares higher sequence similarity with human HDGFL1 (58% identity versus 53%). Although mouse HDGFL1 is specifically expressed in testis, like human HDGFL1, the human and mouse proteins are more divergent in the C-terminal region and probably have different functions. In fact, mouse HDGFL1 lacks the caspase cleavage site identified in mouse HDGF, as well as a number of conserved residues that are known to be phosphorylated (genecards.org).

Discussion

Several recent studies have highlighted the prevalence of errors in genes predicted from genome sequences [24-26,44], particularly in eukaryotic genes. The situation is further complicated by the fact that multiple transcript variants are often expressed by the same gene. Nevertheless, orthology and paralogy, which are fundamental concepts for most evolutionary analyses, are generally defined at the gene level. Many systems, including Ensembl compara [45], simply select the longest transcripts to represent a gene, although there is no guarantee that the longest predicted transcripts in different organisms are equivalent. Some authors have specifically addressed these issues by defining relationships at the transcript level [46,47] or by using processed

Table 4 GO term enrichment analysis for artifactual and putative AED events

GO enrichment for all events			GO enrichment for true events only		
GO ID	GO biological process	P-value	GO ID	GO biological process	P-value
0032501	multicellular organismal process	4.E-43	0032501	multicellular organismal process	2.E-13
0048856	anatomical structure development	2.E-32	0050896	**response to stimulus**	9.E-12
0065007	biological regulation	4.E-26	0048856	anatomical structure development	3.E-09
0080090	**regulation of primary metabolic process**	6.E-21	0042060	wound healing	2.E-07
0071842	cellular component organization at cellular level	3.E-20	0050789	regulation of biological process	1.E-06
0060255	**regulation of macromolecule metabolic process**	5.E-19	0071842	cellular component organization at cellular level	2.E-06
0051171	**regulation of nitrogen compound metabolic process**	5.E-19	0007596	blood coagulation	4.E-06
0032774	**RNA biosynthetic process**	5.E-16	0022008	**neurogenesis**	5.E-05
2000112	**regulation of cellular macromolecule biosynthetic process**	7.E-16	0006928	**cellular component movement**	6.E-05
0006139	**nucleobase, nucleoside, nucleotide and nucleic acid metabolic process**	1.E-15	0030182	**neuron differentiation**	4.E-04
0010467	**gene expression**	4.E-13			
0042060	wound healing	4.E-09			
0007596	blood coagulation	2.E-08			
0006810	**transport**	2.E-08			
0007166	**cell surface receptor linked signaling pathway**	3.E-06			
0007411	**axon guidance**	5.E-06			
0007601	**visual perception**	2.E-05			
0016477	**cell migration**	5.E-05			
0030168	**platelet activation**	1.E-04			
0006195	**purine nucleotide catabolic process**	1.E-04			
0009207	**purine ribonucleoside triphosphate catabolic process**	5.E-04			
0016568	**chromatin modification**	6.E-04			
0006915	**apoptosis**	8.E-04			
0060173	**limb development**	9.E-04			

Comparison of GO term enrichment analysis for (i) gene list 1 corresponding to 607 predicted asymmetrical evolution events, including both artifactual and putative true events and (ii) 25O true events obtained after filtering the erroneous sequences. GO terms for biological processes were found with P < 10-4 using AmiGO and then filtered with GO-Module (only key terms are shown). Terms that are specific to only one gene list are highlighted in bold.

transcription units, i.e. a combination of all overlapping sequence variants in the genomic region [48]. Nevertheless, these remain partial solutions only and do not resolve all problems.
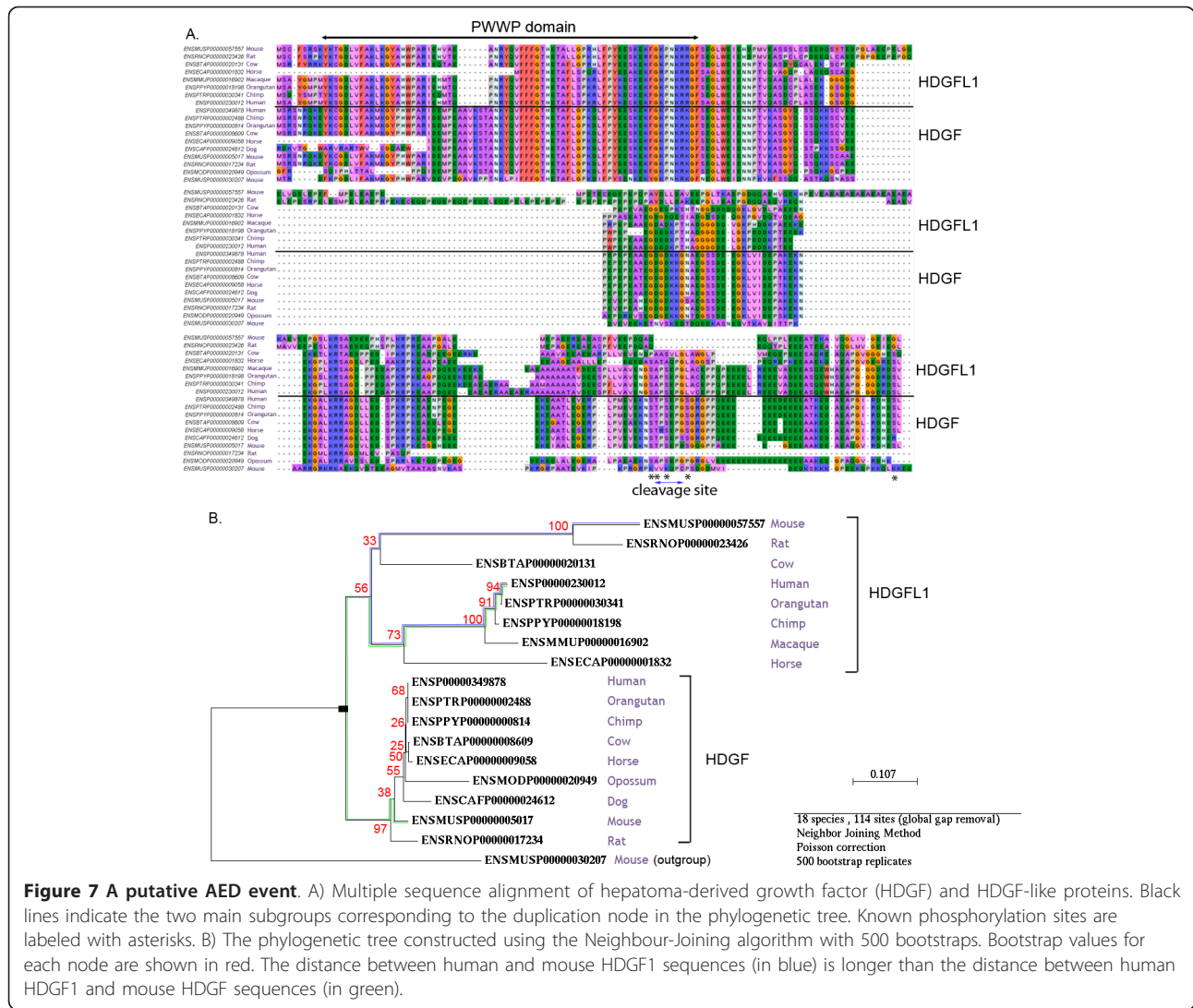
These quality issues may lead to inaccurate or erroneous conclusions if they are integrated indiscriminately in downstream evolutionary or functional analyses. As an example, when annotating a new genome, gene structure data is often transferred from the genome of a closely related species, e.g., many chimpanzee genes in the Ensembl database were predicted based on comparisons with human transcript data. These gene sequences were then used to perform genome-wide scans for positive selection [49]. Although more positively selected genes were identified in chimpanzees compared to human, it has been suggested that the majority of the signals may be due to errors in the original sequences or in the gene alignments [50]. Thus, we have a vicious

circle, where the gene sequences that provide the starting point for most evolutionary analyses are themselves generally predicted based on evolutionary information.

Protein sequence error rates

We detected erroneous protein sequences based on discrepancies in the conservation of vertebrate protein MSAs. The sequence errors may result from (i) DNA sequencing errors, (ii) badly predicted introns/exons, (iii) different splicing variants predicted in different organisms. We estimated the frequency of erroneous sequences to be at least 41%, although some genomes are more error-prone than others, depending on factors such as sequencing coverage or the availability of a well annotated genome from a closely related organism.

In this study, we only considered sequences from the Ensembl database and we used cross-comparisons between species to identify discrepancies. However,

Figure 7 A putative AED event. A) Multiple sequence alignment of hepatoma-derived growth factor (HDGF) and HDGF-like proteins. Black lines indicate the two main subgroups corresponding to the duplication node in the phylogenetic tree. Known phosphorylation sites are labeled with asterisks. B) The phylogenetic tree constructed using the Neighbour-Joining algorithm with 500 bootstraps. Bootstrap values for each node are shown in red. The distance between human and mouse HDGF1 sequences (in blue) is longer than the distance between human HDGF1 and mouse HDGF sequences (in green).

Ensembl may produce predictions that are consistent across organisms, i.e. may reproduce the same errors in different genomes or propagate intron/exon structures. Thus, our estimate of the average sequence error rate is probably conservative. Another recent study [51] showed that the Ensembl compara sequence prediction method correctly identified only 55% of coding transcripts exactly.

Identification of evolutionary events

Our main goal was to determine to what extent these erroneous sequences affect subsequent evolutionary analyses. We focused on a specific event: gene duplication and the evolutionary fate of paralogs, since gene duplication is often assumed to be the most important source of new functions.

Since duplication events where the local copy has evolved more rapidly may indicate unusual evolutionary scenarios, innovations or adaptations, we specifically searched for examples of such asymmetric evolution events. Our approach involved the identification of reliable AED events that could be used as a test set for estimating the impact of sequence errors. We therefore designed a stringent protocol where we included only high coverage genomes and used the well studied human genome as a reference. We then identified putative orthologs in 13 vertebrate genomes, based on either sequence similarity or local synteny conservation. The similarity-based method used a very simple model of sequence evolution, in order to avoid bias towards one particular model. Nevertheless, this model clearly oversimplifies the complex evolutionary processes involved, and in the future, it would be interesting to investigate the effect of a more realistic model of sequence evolution on AED detection, once sequencing/annotation errors have been removed. We also used a strict

definition of local synteny, which led to lower genome coverage in the ortholog prediction step. For the detection of asymmetric evolution, we used a simple measure of amino acid divergence and specified a high significance threshold that would ensure only reliable predictions. Nevertheless, 688 putative AED events were identified that were then used to perform an in-depth investigation of the effect of sequence errors.

Impact of sequence errors

We compared the syntenic and highest similarity homologs and identified cases where significantly faster evolutionary rates were observed in the syntenic homolog, i.e. the gene copy that retained the genome neighbourhood after duplication, compared to the relocated highest similarity homolog. Initially, 688 AED events were identified, of which 81 similarity homologs were potential retropseudogenes with a reduced exonic map. The majority (57%) of the remaining detected events corresponded to erroneous sequences and only 250 represented putative true AED events. Thus, we conclude that care should be taken when performing genome-wide scans to search for genes with unusual patterns, since outlying genes are more likely to be due to artifacts in the input sequences than the result of true evolutionary events. Furthermore, our in-depth study revealed some of the mechanisms by which errors in the input sequences are propagated during the event prediction. For example, a badly predicted internal segment in one of the homologs results in an increased evolutionary distance to the human reference sequence, while a loss in the more variable N/C-terminal regions artificially reduces the distance. These observations provide guidelines for future error detection and correction strategies that will hopefully allow us to reduce the impact of the sequencing errors.

In asymmetric evolution, one duplicate evolves or degrades faster than the other and often becomes functionally or conditionally specialized. In this context, the accurate detection of the 'functional' homologs, i.e. protein pairs that play functionally equivalent roles [52], is critical. We have shown that orthology assignment and the detection of important genetic events are severely impacted by the high proportion of errors in the initial set of protein sequences, even in high coverage genomes. The errors in the initial data are accumulated and amplified in the higher-level analyses. Our estimated rate of 41% erroneous protein sequences leads to 57% errors in AED event prediction and, in the subsequent Gene Ontology (GO) functional analysis, 75% of the enriched terms are in fact false positives.

The false positive terms in the functional analysis can be very costly to investigate experimentally and a reduction in the false discovery rate is clearly desirable. They

are also sufficient to mask some of the true functional enrichments. After filtering the artifactual events corresponding to erroneous sequences, the remaining AED events were enriched in a number of GO categories, including neuron differentiation and response to external stimuli. Interestingly, human-specific duplicates evolving under adaptive natural selection also include genes involved in neuronal and cognitive functions, as well as response to inflammation or stress [53]. Similarly, gene families involved in copy number variations (CNVs) are enriched for similar categories, including interactions with the environment, neurophysiological processes and brain development [54]. A recent study suggested that the relationship between CNVs and positive selection may play an important role in the emergence and evolution of species-specific traits in primates [55]. Genes in many of these categories are thus thought to be important in evolutionary adaptation and to be particular targets of natural selection.

Conclusions

Up to half of all protein sequences in today's genome databases contain erroneous insertions, deletions or suspicious segments. The high error rates have profound implications, not only for the analysis of protein functions, interaction networks, biochemical pathways or disease phenotypes, but also for our understanding of life's evolution.

The putative sequence errors identified here lead to a significant number of false positives in the detection of asymmetric evolution events, which, if ignored, are sufficient to obscure their true functional significance. We have looked at one important event, asymmetric evolution after duplication, but the effect of protein sequence errors is likely to be similar for other types of events. This might explain many of the contradictions observed in many recent evolutionary studies, aggravating the effects of differences in source data, methodology and planning of experiments [12].

Exploitation of the new genome data is clearly challenging, due to the size of the data sets, their complexity and the high level of noise, and the situation is not likely to improve with low coverage genomes becoming the norm. As a consequence, data cleaning tools and robust statistical analyses will be essential for its reliable interpretation. With as many as 50% erroneous sequences, the simple removal of this data will result in the loss of too much information. It will be necessary to validate and correct the sequence errors and ideally, propagate these corrections to the public databases. Some recent efforts have been undertaken to address these issues [19,26,47], but additional work will be essential to reduce the impact of error and to extract the true meaning hidden in the data.

The alternative is an escalating process where systematic errors are accumulated at each level of the analysis, generating artificially high rates of unusual event predictions and eventually leading to an 'error catastrophe', where the noise overwhelms the true signal.

Methods

Protein sequence data sets

Human protein coding genes were retrieved from the Human Proteome Initiative (HPI) and Swiss-prot databases [56], resulting in a total of 19,778 human sequences. Each gene was then used as a query for a BlastP [57] search in a database consisting of the proteomes of 14 vertebrates (Table 1) with almost complete genomes from the Ensembl (version 51) database [36]. The Ensembl human protein sequence with the highest similarity to the HPI query was designated as the reference protein sequence. For each of the 19,778 human reference sequences, potential orthologs were then identified using two different, complementary approaches: sequence similarity and local synteny.

Putative orthologs based on sequence similarity

For each human reference sequence, a modified version of the PipeAlign [58] protein analysis pipeline was used to construct a multiple sequence alignment (MSA) for all sequences detected by the BlastP search with E < 10^{-3} (maximum sequences = 500). PipeAlign integrates several steps, including post-processing of the BlastP results, construction of a MSA of the full-length sequences with DbClustal [59], verification of the MSA with RASCAL [60] and removal of unrelated sequences with LEON [61]. In this modified version, DbClustal was replaced by the MAFFT [62] program, since the computational speed of MAFFT is better suited to high throughput projects. The MSAs obtained from this pipeline were then annotated with structural and functional information using MACSIMS [63], an information management system that combines knowledge-based methods with complementary *ab initio* sequence-based predictions. MACSIMS integrates several types of data in the alignment, in particular Gene Ontology annotations, functional annotations and keywords from Swissprot, and functional/structural domains from the Pfam database [64].

Based on the MSA, the evolutionary pairwise distance, d, between any two sequences was defined as the number of amino acid substitutions per site under the assumption that the number of amino acid substitutions at each site follows the Poisson distribution. Thus:

$$d = -\ln(1-p)$$

where d is the pairwise distance and p is the proportion of different amino acids aligned (dissimilarity).

Then, for each human reference sequence, H_i, the sequences from the 13 vertebrate organisms with the highest similarity (i.e. the smallest distance) to H_i were identified and denoted Vn_Sim_i, where Vn refers to one of the 13 vertebrate organisms (Figure S2A in Additional file 1).

Putative orthologs based on local synteny

The chromosomal localization of all genes coding for protein sequences was obtained from the Ensembl database. Locally developed software was used to identify regions on the human chromosomes where local synteny was conserved between the human genome and each of the other 13 vertebrate genomes. The chromosomes in each genome are thus represented as a linear sequence of genes. For each human reference sequence, the local syntenic homolog was defined as outlined in (Figure S2B in Additional file 1). For the coding gene, H_i, at position i on the human genome, its neighbours (H_{i-1} and H_{i+1}) were identified. For each of the 13 vertebrate genomes, the sequences with the highest similarity to H_{i-1} and H_{i+1} were selected from the MSA as described above, and denoted Vn_Sim_{i-1} and Vn_Sim_{i+1} respectively, where Vn refers to one of the 13 vertebrate genomes. A local synteny homolog, Vn_Syn_i exists for H_i and genome Vn if: (i) homologs were found in Vn for H_{i-1} and H_{i+1}, (ii) the separation between the highest similarity homologs, denoted Vn_Sim_{i-1} and Vn_Sim_{i+1}, on the genome was less than 5 genes and (iii) a homolog of H_i was found on the genome between Vn_Sim_{i-1} and Vn_Sim_{i+1}. The homolog of H_i localized between V_Sim_{i-1} and V_Sim_{i+1} with the highest similarity (smallest evolutionary distance) to the human reference sequence was then defined as the syntenic homolog.

Genes with ambiguous genomic locations, such as scaffolds etc., were discarded since the synteny relationship could not be reliably established. In addition, local or tandem duplications were excluded since the genome contexts of the two gene copies were similar. Although tandem duplicates should be adjacent to each other on one chromosome, extensive gene inversions may insert irrelevant genes into the tandem arrays. We therefore used a stringent threshold and excluded cases where Vn_Sim_i and Vn_Syn_i were separated on the genomes by less than 10 genes.

Automatic detection of potential sequence errors

For each MSA corresponding to a human reference sequence, an automatic protocol was used to detect sequence discrepancies that may indicate gene prediction errors. Different types of prediction error were
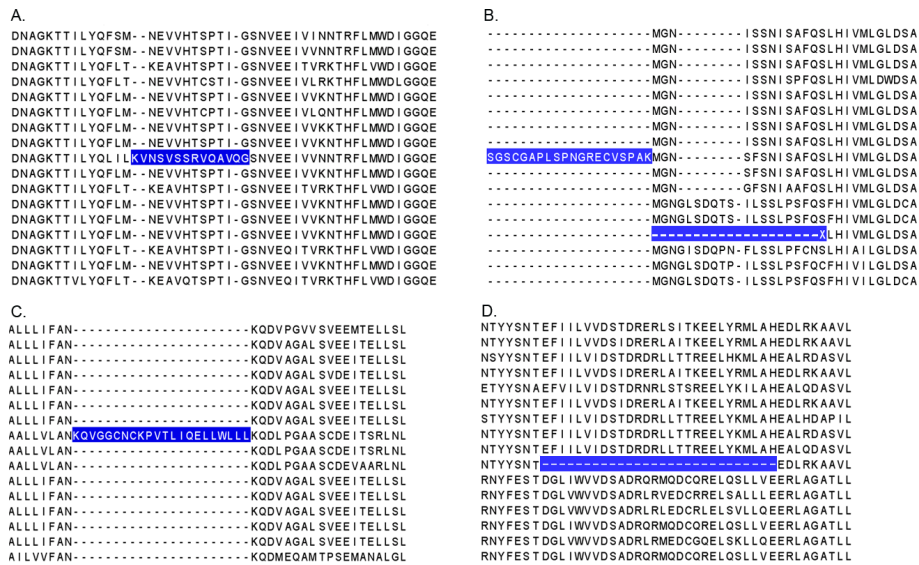
Figure 8 Detection of potential sequence errors. Examples of sequence discrepancies (highlighted in blue) that are identified in the subfamily alignments. A) Potential mispredicted exons, resulting in suspicious sequence segments, are identified based on the conserved blocks in the subfamily alignment. B) Potential start and stop site errors are predicted based on the distribution of the positions of the N/C-terminal residues. C) Identification of a potential inserted intron, based on the presence of a single sequence with the insertion in a given subfamily. D) Identification of a potential missing exon, based on the presence of a single sequence with a deletion in a given subfamily.

considered, such as excluding coding exons, including introns as part of the coding sequence, or wrongly predicting start and termination sites. The protocol is described in detail elsewhere [37]. Briefly, the sequences in the MSA were first clustered into more related subfamilies then, for each subfamily, sequences with potential errors were identified using an empirical rule-based approach. (i) Badly predicted exons were identified using the RASCAL algorithm [60] as outliers or 'suspicious' sequence segments (Figure 8A). (ii) Badly predicted start or stop sites were identified by considering the positions of the N/C-terminal residues for each sequence in the subfamily alignment (Figure 8B). Normal values were defined as lying within the lower and upper quartiles of the distribution of terminal positions. Sequences with terminal positions outside this window were annotated as potential deletion/extension errors. (iii) Inserted introns (Figure 8C) were detected if a single sequence contained an insertion of more than 10 residues. (iv) Missing exons (Figure 8D) were detected if a single sequence contained a deletion of more than 10 residues.

Each error was then classified in one of 7 different classes: internal insertions, internal deletions, suspicious sequence segments, extensions at the N- or C-terminus, and deletions at the N- or C-terminus.

Validation of potential sequence errors

The errors in the protein sequences were estimated by analysing the corresponding DNA gene sequences from the Ensembl database. First, if the gene sequence contained a run of 'N' characters, we assumed that the predicted protein sequence error was the result of a DNA sequencing or assembly error. Second, the gene sequences with no 'N' characters were searched for the missing protein sequence fragments. For errors corresponding to internal deletions, deletions at the N- or C- terminus or suspicious sequence segments, the missing protein fragment was extracted from a closely related sequence in the multiple alignment. The protein fragment was then aligned to the gene sequence from the ENSEMBL database using the PairWise software [65]. The fragment was considered to be present in the gene sequence if the percent identity of the protein and translated gene sequences was greater than a given threshold. The threshold used here was specific to the pair of organisms compared and was defined as the lower quartile of the protein sequence identities for the complete proteomes of the two organisms. A similar protocol was used for the errors corresponding to insertions in a given protein sequence, except that, in this case, the protein fragment corresponding to the insertion was aligned to the gene sequence of another closely

related sequence. Finally, the transcript evidence for the protein sequences in the Ensembl database was searched manually for known transcripts and splicing variants.

Prediction of asymmetrical evolutionary rates

It has been suggested that, after a gene duplication event, one duplicate generally maintains the ancestral function while the other is free to evolve and acquire novel functionality. This scenario implies that the protein with conserved functionality will undergo less sequence evolution than the one exploring new functionalities. To determine which of the two homologs described above (highest sequence similarity or syntenic) was more likely to share the same function as the human reference sequence, we estimated the difference between the two evolutionary distances: human reference to similarity homolog and human reference to syntenic homolog. Thus, for each of the 13 vertebrate genomes considered in this study, we have a triplet of homologs, H_i, Vn_sim_i, Vn_syn_i, and we want to estimate the difference Δ between two distances $d(H_i, Vn\_sim_i)$ and $d(H_i, Vn\_syn_i)$.

We used an estimator based on pairwise sequence distances similar to one defined previously, that is relatively fast to compute and has almost the same statistical power as the widely used maximum likelihood estimator [66]. The distance, d, between two sequences is defined as the number of amino acid substitutions per site under the assumption that the number of amino acid substitutions at each site follows the Poisson distribution, as before. The variance σ of the distance d is given by:

$$\sigma^2(d) = p / \left[(1 - p)\, n \right]$$

where p is the proportion of amino acid differences and n is the total number of amino acids compared.

If X has two homologs Y and Z, and Y is the closest homolog to X, an estimator for the difference in evolutionary distances is:

$$\Delta = d(X, Y) - d(X, Z)$$

The variance of the difference can be computed as:

$$\sigma^2(\Delta) = \sigma^2(d(X, Y)) + \sigma^2(d(X, Z)) \\ - 2cov(d(X, Y), d(X, Z))$$

and thus, an upper bound for the variance of the estimator is:

$$\sigma^2(\Delta) = \sigma^2(d(X, Y)) + \sigma^2(d(X, Z))$$

Finally, we assume X,Y are significantly closer than X, Z if:

$$\Delta < -k.\sigma(\Delta)$$

In this work, the parameter k was set to 1.96, reflecting the 95% confidence level. Thus, we would expect 5% of the tested gene triplets to falsely reject the hypothesis of asymmetrical evolution.

Additional material

Additional file 1: Supporting figures and tables. Supporting figures and tables for the manuscript are provided as a PDF file.

Additional file 2: Examples of erroneous protein sequences and their validation. Example text and figures are provided as a PDF file.

Author details
[1]Department of Integrated Structural Biology, IGBMC (Institut de Génétique et de Biologie Moléculaire et Cellulaire) CNRS/INSERM/Université de Strasbourg, 1 rue Laurent Fries, Illkirch, F-67404, France. [2]Medical Biochemistry Department, Federal University of Rio de Janeiro, Avenida Carlos Chagas Filho 373, Rio de Janeiro, 21941-902, Brazil. [3]UMR-CNRS 6632 Evolution Biologique et Modélisation, Université de Provence, 3, Place Victor Hugo, Marseille, 13331, France.

Authors' contributions
FP participated in the design of the study, constructed the multiple alignments and synteny data, and helped draft the manuscript. BL designed and carried out the ortholog predictions and participated in the analysis of the data. PP participated in the design of the study and the genetic event analysis, and helped draft the manuscript. OP participated in the design and coordination of the study and the analysis of the data and helped draft the manuscript. JDT conceived the study, participated in its design and coordination, and helped to analyse the data and to draft the manuscript. All authors read and approved the final manuscript.

References
1. Mardis ER: **A decade's perspective on DNA sequencing technology.** *Nature* 2011, **470**(7333):198-203.
2. Philippe H, Brinkmann H, Lavrov DV, Littlewood DT, Manuel M, Worheide G, Baurain D: **Resolving difficult phylogenetic questions: why more sequences are not enough.** *PLoS Biol* 2011, **9**(3):e1000602.
3. Soria-Carrasco V, Castresana J: **Estimation of phylogenetic inconsistencies in the three domains of life.** *Mol Biol Evol* 2008, **25**(11):2319-2329.
4. Stiller JW: **Experimental design and statistical rigor in phylogenomics of horizontal and endosymbiotic gene transfer.** *BMC Evol Biol* 2011, **11**(1):259.
5. Koonin EV: **The origin and early evolution of eukaryotes in the light of phylogenomics.** *Genome Biol* 2011, **11**(5):209.
6. Pace NR: **Mapping the tree of life: progress and prospects.** *Microbiol Mol Biol Rev* 2009, **73**(4):565-576.
7. Parfrey LW, Lahr DJ, Knoll AH, Katz LA: **Estimating the timing of early eukaryotic diversification with multigene molecular clocks.** *Proc Natl Acad Sci USA* 2011, **108**(33):13624-13629.
8. Desmond E, Brochier-Armanet C, Forterre P, Gribaldo S: **On the last common ancestor and early evolution of eukaryotes:**

reconstructing the history of mitochondrial ribosomes. *Res Microbiol* 2011, **162**(1):53-70.

9. Negrisolo E, Kuhl H, Forcato C, Vitulo N, Reinhardt R, Patarnello T, Bargelloni L: **Different phylogenomic approaches to resolve the evolutionary relationships among model fish species.** *Mol Biol Evol* 2010, **27**(12):2757-2774.

10. Campbell V, Lapointe FJ: **An application of supertree methods to Mammalian mitogenomic sequences.** *Evol Bioinform Online* 2010, **6**:57-71.

11. Agnarsson I, Kuntner M, May-Collado LJ: **Dogs, cats, and kin: a molecular species-level phylogeny of Carnivora.** *Mol Phylogenet Evol* 2010, **54**(3):726-745.

12. Studer R, Robinson-Rechavi M: **How confident can we be that orthologs are similar, but paralogs differ?** *Trends Genet* 2009, **25**:210-216.

13. Kumar S, Filipski AJ, Battistuzzi FU, Kosakovsky Pond SL, Tamura K: **Statistics and Truth in Phylogenomics.** *Mol Biol Evol* 2011.

14. Sanderson MJ, McMahon MM, Steel M: **Phylogenomics with incomplete taxon coverage: the limits to inference.** *BMC Evol Biol* 2010, **10**:155.

15. Aittokallio T: **Dealing with missing values in large-scale studies: microarray data imputation and beyond.** *Brief Bioinform* 2010, **11**(2):253-264.

16. Berthoumieux S, Brilli M, de Jong H, Kahn D, Cinquemani E: **Identification of metabolic network models from incomplete high-throughput datasets.** *Bioinformatics* 2010, **27**(13):i186-195.

17. Pop M, Salzberg SL: **Bioinformatics challenges of new sequencing technology.** *Trends Genet* 2008, **24**:142-149.

18. Hayden EC: **Genome builders face the competition.** *Nature* 2011, **471**(7339):425.

19. Hubisz M, Lin M, Kellis M, Siepel A: **Error and error mitigation in low-coverage genome assemblies.** *PLOS One* 2011, **6**:e17034.

20. Vilella AJ, Birney E, Flicek P, Herrero J: **Considerations for the inclusion of 2x mammalian genomes in phylogenetic analyses.** *Genome Biol* 2011, **12**(2):401.

21. Hoff KJ: **The effect of sequencing errors on metagenomic gene prediction.** *BMC Genomics* 2009, **10**:520.

22. Milinkovitch M, Helaers R, Depiereux E, Tzika A, Gabaldon T: **2X genomes - depth does matter.** 2010, **11**:R16.

23. Pertea M, Salzberg SL: **Between a chicken and a grape: estimating the number of human genes.** *Genome Biol* 2011, **11**(5):206.

24. Brent MR: **Steady progress and recent breakthroughs in the accuracy of automated genome annotation.** *Nat Rev Genet* 2008, **9**(1):62-73.

25. Harrow J, Nagy A, Reymond A, Alioto T, Patthy L, Antonarakis SE, Guigo R: **Identifying protein-coding genes in genomic sequences.** *Genome Biol* 2009, **10**(1):201.

26. Nagy A, Hegyi H, Farkas K, Tordai H, Kozma E, Banyai L, Patthy L: **Identification and correction of abnormal, incomplete and mispredicted proteins in public databases.** *BMC Bioinformatics* 2008, **9**:353.

27. Hallegger M, Llorian M, Smith CW: **Alternative splicing: global insights.** *Febs J* 2010, **277**(4):856-866.

28. Schneider A, Souvorov A, Sabath N, Landan G, Gonnet GH, Graur D: **Estimates of positive Darwinian selection are inflated by errors in sequencing, annotation, and alignment.** *Genome Biol Evol* 2009, **1**:114-118.

29. Ohno S: *Evolution by gene duplication* Berlin (Germany): Springer Verlag; 1970.

30. Semon M, Wolfe KH: **Consequences of genome duplication.** *Curr Opin Genet Dev* 2007, **17**(6):505-512.

31. Kassahn KS, Dang VT, Wilkins SJ, Perkins AC, Ragan MA: **Evolution of gene function and regulatory control after whole-genome duplication: comparative analyses in vertebrates.** *Genome Res* 2009, **19**(8):1404-1418.

32. Levasseur A, Pontarotti P: **The role of duplications in the evolution of genomes highlights the need for evolutionary-based approaches in comparative genomics.** *Biol Direct* 2011, **6**:11.

33. Durand D, Hoberman R: **Diagnosing duplications–can it be done?** *Trends Genet* 2006, **22**(3):156-164.

34. Conant GC, Wolfe KH: **Turning a hobby into a job: how duplicated genes find new functions.** *Nat Rev Genet* 2008, **9**(12):938-950.

35. Jun J, Ryvkin P, Hemphill E, Nelson C: **Duplication mechanism and disruptions in flanking regions determine the fate of Mammalian gene duplicates.** *J Comput Biol* 2009, **16**(9):1253-1266.

36. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Chen Y, Clapham P, Coates G, Fairley S, Fitzgerald S, *et al*: **Ensembl 2011.** *Nucleic Acids Res* 2011, , **39 Database**: D800-806.

37. Thompson JD, Linard B, Lecompte O, Poch O: **A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives.** *PLoS One* 2011, **6**(3):e18093.

38. Meader S, Hillier L, Locke D, Ponting C, Lunter G: **Genome assembly quality: Assessment and improvement using the neutral indel model.** *Genome Res* 2010, **20**:675-684.

39. Boyer F, Morgat A, Labarre L, Pothier J, Viari A: **Syntons, metabolons and interactons: an exact graph-theoretical approach for exploring neighbourhood between genomic and functional data.** *Bioinformatics* 2005, **21**(23):4209-4215.

40. Rodelsperger C, Dieterich C: **Syntenator: multiple gene order alignments with a gene-specific scoring function.** *Algorithms Mol Biol* 2008, **3**:14.

41. Jun J, Mandoiu I, Nelson C: **Identification of mammalian orthologs using local synteny.** *BMC Genomics* 2009, **10**:630.

42. Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S: **AmiGO: online access to ontology and annotation data.** *Bioinformatics* 2009, **25**(2):288-289.

43. Yang X, Li J, Lee Y, Lussier YA: **GO-Module: functional synthesis and improved interpretation of Gene Ontology patterns.** *Bioinformatics* 2011, **27**(10):1444-1446.

44. Ranwez V, Harispe S, Delsuc F, Douzery EJ: **MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons.** *Plos One* 2011, **6**(9):e22594.

45. Hubbard TJ, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, *et al*: **Ensembl 2007.** *Nucleic Acids Res* 2007, , **35 Database**: D610-617.

46. Zambelli F, Pavesi G, Gissi C, Horner DS, Pesole G: **Assessment of orthologous splicing isoforms in human and mouse orthologous genes.** *BMC Genomics* 2010, **11**:534.

47. Goodstadt L, Ponting CP: **Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human.** *PLoS Comput Biol* 2006, **2**(9):e133.

48. Ho MR, Jang WJ, Chen CH, Ch'ang LY, Lin WC: **Designating eukaryotic orthology via processed transcription units.** *Nucleic Acids Res* 2008, **36**(10):3436-3442.

49. Bakewell MA, Shi P, Zhang J: **More genes underwent positive selection in chimpanzee evolution than in human evolution.** *Proc Natl Acad Sci USA* 2007, **104**(18):7489-7494.

50. Mallick S, Gnerre S, Muller P, Reich D: **The difficulty of avoiding false positives in genome scans for natural selection.** *Genome Res* 2009, **19**(5):922-933.

51. Guigo R, Flicek P, Abril JF, Reymond A, Lagarde J, Denoeud F, Antonarakis S, Ashburner M, Bajic VB, Birney E, *et al*: **EGASP: the human ENCODE Genome Annotation Assessment Project.** *Genome Biol* 2006, **7**(Suppl 1):S2 1-31.

52. Remm M, Storm CE, Sonnhammer EL: **Automatic clustering of orthologs and in-paralogs from pairwise species comparisons.** *J Mol Biol* 2001, **314**(5):1041-1052.

53. Han MV, Demuth JP, McGrath CL, Casola C, Hahn MW: **Adaptive evolution of young gene duplicates in mammals.** *Genome Res* 2009, **19**(5):859-867.

54. de Smith AJ, Walters RG, Froguel P, Blakemore AI: **Human genes involved in copy number variation: mechanisms of origin, functional effects and implications for disease.** *Cytogenet Genome Res* 2008, **123**(1-4):17-26.

55. Gokcumen OO, Babb PL, Iskow R, Zhu Q, Shi X, Mills RE, Ionita-Laza I, Vallender EJ, Clark AG, Johnson WE, *et al*: **Refinement of primate CNV hotspots identifies candidate genomic regions evolving under positive selection.** *Genome Biol* 2011, **12**(5):R52.

56. UniProt: **Ongoing and future developments at the Universal Protein Resource.** *Nucleic Acids Res* 2011, , **39 Database**: D214-219.

57. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**(17):3389-3402.

58. Plewniak F, Thompson JD, Poch O: **Ballast: blast post-processing based on locally conserved segments.** *Bioinformatics* 2000, **16**:750-759.

59. Thompson J, Plewniak F, Thierry J, O P: **DbClustal: rapid and reliable global multiple alignments of protein sequences detected by database searches.** *Nucleic Acids Res* 2000, **28**:2919-2926.

60. Thompson JD, Thierry JC, Poch O: **RASCAL: rapid scanning and correction of multiple sequence alignments.** *Bioinformatics* 2003, **19**:1155-1161.

61. Thompson JD, Prigent V, Poch O: **LEON: multiple aLignment Evaluation Of Neighbours.** *Nucleic Acids Res* 2004, **32**:1298-1307.

62. Katoh K, Toh H: **Recent developments in the MAFFT multiple sequence alignment program.** *Brief Bioinform* 2008, **9**:286-298.

63. Thompson JD, Muller A, Waterhouse A, Procter J, Barton GJ, Plewniak F, Poch O: **MACSIMS: multiple alignment of complete sequences information management system.** *BMC Bioinformatics* 2006, **7**:318.

64. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, *et al*: **The Pfam protein families database.** *Nucleic Acids Res* 2010, , **38 Database:** D211-222.

65. Birney E, Thompson J, Gibson T: **PairWise and SearchWise: finding the optimal alignment in a simultaneous comparison of a protein profile against all DNA translation frames.** *Nucleic Acids Res* 1996, **24(14)**:2730-2739.

66. Dessimoz C, Gil M, Schneider A, Gonnet G: **Fast estimation of the difference between two PAM/JTT evolutionary distances in triplets of homologous sequences.** *BMC Bioinformatics* 2006, **7**:529.

Annex II :

Publication n° 4

A Comprehensive Benchmark Study of Multiple Sequence Alignment Methods: Current Challenges and Future Perspectives

Julie D. Thompson*, Benjamin Linard, Odile Lecompte, Olivier Poch

Département de Biologie Structurale et Génomique, IGBMC (Institut de Génétique et de Biologie Moléculaire et Cellulaire), CNRS/INSERM/Université de Strasbourg, Illkirch, France

Abstract

Multiple comparison or alignmentof protein sequences has become a fundamental tool in many different domains in modern molecular biology, from evolutionary studies to prediction of 2D/3D structure, molecular function and inter-molecular interactions etc. By placing the sequence in the framework of the overall family, multiple alignments can be used to identify conserved features and to highlight differences or specificities. In this paper, we describe a comprehensive evaluation of many of the most popular methods for multiple sequence alignment (MSA), based on a new benchmark test set. The benchmark is designed to represent typical problems encountered when aligning the large protein sequence sets that result from today's high throughput biotechnologies. We show that alignmentmethods have significantly progressed and can now identify most of the shared sequence features that determine the broad molecular function(s) of a protein family, even for divergent sequences. However,we have identified a number of important challenges. First, the locally conserved regions, that reflect functional specificities or that modulate a protein's function in a given cellular context,are less well aligned. Second, motifs in natively disordered regions are often misaligned. Third, the badly predicted or fragmentary protein sequences, which make up a large proportion of today's databases, lead to a significant number of alignment errors. Based on this study, we demonstrate that the existing MSA methods can be exploited in combination to improve alignment accuracy, although novel approaches will still be needed to fully explore the most difficult regions. We then propose knowledge-enabled, dynamic solutions that will hopefully pave the way to enhanced alignment construction and exploitation in future evolutionary systems biology studies.

Introduction

Evolutionary theory provides a unifying framework for analysing genomics data and for studying various phenomena in molecular, cell, or developmental biology [1]. Thus, evolutionary-based inference systems are playing an increasingly important role in diverse areas, such as elucidation of the tree of life [2], studies of epidemiology and virulence [3], drug design [4], human genetics [5], cancer [6] or biodiversity [7]. Essential prerequisites for such evolutionary-based studies are the multiple sequence alignment (MSA) and its subsequent analysis [8,9,10]. By placing the sequence in the framework of the overall family, MSAs can be used to characterise important features thatdetermine the broad molecular function(s) of the protein, such as the 3-dimensional structure or catalytic sites, and that have been conserved throughout evolution.However, most proteins act in complex, dynamic networks that are dependent on the biological context, for example subcellular localisation, temporal and spatial expression patterns, or environment. Here, MSAs will alsohave a crucial role to play in identifying the specific features, also known as "specificity determining positions" (SDPs), that modulate a

protein's function in a given context, for example, interaction domains, regions or sites, targeting signals in the different cell machineries, pathways orcompartments, or post-translational modification sites(phosphorylation, cleavage, etc.) [11,12,13].

MSA algorithms have been an active area of research since the 1980s. Traditionally the most popular approach has been the progressive alignment procedure [14], which exploits the fact that homologous sequences are evolutionarily related. A multiple sequence alignment is built up gradually using a series of pairwise alignments, following the branching order in a phylogenetic tree. A number of different alignment programs based on this method have been developed, includingboth global and local approaches. A global MSA algorithm is defined here as one that tries to align the full length sequences from one end to the other. Once the global alignment has been constructed, other methods are often used to identify the more conserved or reliable regions within the alignment. A local algorithm attempts to identify subsequences sharing high similarity. The unreliable or low similarity regions are then either excluded from the alignment, or are differentiated, for example, by the use of upper/lower case characters. Comparisons of many of these methods based on 'gold standard' benchmarks

[15,16] showed that none of the existing algorithms were capable of providing accurate alignments for all the test cases. As a consequence, iterative algorithms were developed to construct more reliable multiple alignments, using for example iterative refinement strategies [17], Hidden Markov Models [18] or Genetic Algorithms [19]. These methods were shown to be more successful at aligning the most conserved regions for a wide variety of test cases, although some accuracy was lost for distantly related sequences, in the 'twilight zone' of evolutionary relatedness [20,21].

In the post-genomic era, the growing complexity of the multiple alignment problem has lead to the development of novel methods that use a combination of different alignment algorithms [22,23,24,25] or that incorporate biological information other than the sequence itself [26,27]. A number of specific MSA problems have also been addressed by programs such as POA [28] for the alignment of non-linear sequences or PRANK [10] for the detailed evolutionary analysis of more closely related sequences. These new MSA construction methods are generally evaluated using one or more alignment benchmarks, for example, BAliBASE [15], OxBench [29] or PREFAB [24], and it is clear that this benchmarking has had a positive effect on their development [30]. Most of the widely used MSA benchmarks were compared in [21] and are also discussed in [31]. The use of objective benchmarks leads to a better understanding of the problems underlying poor performance, by highlighting specific weak points or bottlenecks. Thus, benchmarking can help the developer improve the performance of his software. In turn, the software improvements imply that the benchmarks must continually evolve, if they are to represent the current problems and challenges in the domain [31].

Today, new high throughout biotechnologies are providing us with enough data to build complete evolutionary histories of large sets of genes [32]. For the first time, it will be possible to compare sequences from hundreds of diverse organisms, both present and extinct, to perform detailed studies of the evolutionary patterns and forces that shaped extant genes and to reconstruct the genetic changes that are responsible for the phenotypic differences between organisms. Although the current flood of data clearly provides unique opportunities for systems-level studies, it also poses many new challenges, in addition to the obvious scalability issues. First, although the range of organisms studied has increased recently, a relatively small number of model organisms still dominate the public databases. Second, the protein families represented in today's sequence databases are often more complex, with multidomain architectures, large unstructured (natively disordered) regions, numerous splicing variants, etc. Third, the new sequences are mostly predicted by automatic methods and thus, contain a significant number of sequence errors [33,34]. For example, the EGASP assessment of gene prediction algorithms showed that the best gene prediction systems are able to predict entirely correct sequences for protein transcripts in the human genome only 50% of the time [35]. The problem has been further exacerbated by the next generation (massively parallel) DNA sequencing instruments that can sequence up to one billion bases in a single day at low cost [36]. These new technologies produce read lengths as short as 35–40 nucleotides, resulting in fragmentary protein sequences that pose problems for bioinformatics analyses [37]. If MSA methodology is to keep pace with the new challenges presented by this complex and often 'noisy' sequence data, the alignment benchmarks used for evaluation must now evolve to reflect this changing biological sequence space.

Here, we describe a new protein sequence alignment benchmark designed to reproduce today's sequence exploration requirements and a comprehensive assessment of the performance of some of the most popular MSA programs. Our study was motivated by two major observations. First, most of the existing MSA benchmarks - and as a consequence, most MSA construction algorithms - have focused on the patterns conserved in the majority of the sequences and not enough attention has been paid to the less frequent patterns, or SDPs, that might indicate subfamily-specific or context-specific functions. Second, current MSA programs for protein sequences generally model globular domain structure and evolution. Nevertheless, many proteins, particularly in eukaryotes, are unstructured (natively disordered) or contain large unstructured regions. These regions frequently contain motifs, such as signalling sequences or sites of posttranslational modifications, that are involved in the regulatory functions of a cell [38,39]. While this complexity alone represents a significant challenge for today's MSA algorithms, another major goal of our study was to investigate the effect of the 'noisy' data, including fragmentary or otherwise erroneous sequences, on MSA program performance.

Our benchmark, representing 218 large, complex protein families, has been incorporated in the BAliBASE benchmark suite and provides a complementary test to the existing reference sets. While the previous sets included mainly alignments of shared, structured domains, the reference set described here focuses on (i) subfamily specific features, (ii) motifs in disordered regions, (iii) the effect of fragmentary or otherwise erroneous sequences on MSA quality. The new benchmark tests were then used to evaluate the quality of the alignments produced by some of the most widely used programs for MSA construction. This comparative study allowed us to evaluate the recent progress achieved and to highlight a number of specific strengths and weaknesses of the different approaches. Finally, we propose new directions for the future development of multiple alignment construction and analysis methods.

Results

Benchmark alignments

The BAliBASE benchmark suite contains multiple sequence alignments, organised into 9 Reference Sets representing specific MSA problems, including small numbers of sequences, unequal phylogenetic distributions, large N/C-terminal extensions or internal insertions, repeats, inverted domains and transmembrane regions. Here, we have constructed a new BAliBASE test set, Reference 10, composed of 218 reference alignments and containing a total of 17892 protein sequences, which were obtained using a query-based database search protocol. Details of the benchmark alignments are provided in the Methods section. For each reference alignment, we then identified the locally conserved regions, or 'blocks', using an automatic method. This led to the definition of 9131 blocks, covering on average 46% of the total multiple alignment. The remaining regions of the reference alignments, corresponding to the unalignable or unstable segments, were excluded from the analyses performed in this work. The resulting benchmark alignments reflect some of the problems specific to aligning large sets of complex protein sequences. For example, many of the protein families (>64% of the alignments) have multidomain architectures and their members often share only a single domain. Another important feature of the alignments is linked to the distribution of the conserved blocks. The alignment of the highly studied P53/P63/P73 family (Figure 1A), illustrates this concept with only 18% of the blocks present in most (>90%) of the aligned sequences, while 30% are found in less than 10%. These 'rare' segments or patterns are often characteristic of context-specific functions, e.g. substrate binding sites, protein-

protein interactions or post-translational modification sites. Finally, the alignments have a high proportion of sequences with 'discrepancies', i.e. unexpected or discordant extensions, insertions or deletions, as shown in Figure 2. These discrepancies may correspond to naturally occurring variants or may be the result of artifacts, including PDB sequences (typically covering a single structural domain), proteins translated from partially sequenced genomes or ESTs, or badly predicted protein sequences. In the alignment in Figure 1A, 45% of the aligned sequences (61 out of 134) contain one or more of these discrepancies.

MSA program evaluation: overall alignment quality

For each of the 218 reference alignments in the benchmark, we applied eight alignment programs, resulting in a total of 1744 automatically constructed MSAs. The overall quality of these automatic alignments was measured using the Column Score (CS) described in Methods. This initial experiment generally confirmed previous findings, in terms of program ranking (Figure 3). Probcons, TCoffee and the most recent version of Mafft (linsi) (version 6.815) achieved the highest average scores (79.4% and 81.6% respectively). Nevertheless, Probcons and TCoffee took over 2.7 days to compute all the alignments, while Mafft (linsi) took 1.2 hours. The fastest program, Kalign, required only 3.0 minutes computation time, although some loss of accuracy was observed (74.3%). As expected, the more recent methods incorporating both local and global algorithms were generally more accurate than older methods, based on global (ClustalW: 64.4%) or local (Dialign-tx: 73.8%) algorithms alone. Individual alignment accuracy was highly variable even for the best programs (with a standard deviation of 19.6, 19.1 and 18.9 for Probcons, TCoffee and Mafft (linsi) respectively). This is in agreement with previous observations showing that some alignments are more difficult than others [10,20].

To investigate in more detail the factors affecting the performance of each program, we characterized each alignment using a number of 'global' attributes describing the overall full-length alignment, including the number of sequences to be aligned, their length, an MSA objective function (norMD) [40] and the percentage of the alignment covered by the blocks. Figure 4(A–D) shows the distributions of the overall alignment quality scores obtained by each MSA program for each global attribute. These distributions, together with a correlation analysis (Figure 4E), showed that more closely related sequences were generally aligned better (positive correlation for all programs with the norMD and percent coverage by blocks), as might be expected. For the more difficult alignment tests, e.g. with norMD<0.2, the mean CS scores were less than 0.5 for all the aligners included in this study. The length of the sequences had less effect on alignment quality, although longer sequences tended to be less well aligned. In contrast to some previous studies [20,21], we observed a negative correlation with the number of sequences in these alignments, i.e. the alignments with a larger number of sequences were less well aligned. For alignments with more than 80 sequences, only Mafft (linsi) achieved CS scores higher than 0.7.

Effect of sequence discrepancies on alignment quality

To study the effect of the new sequences resulting from high throughput biotechnologies, we identified sequence discrepancies that might be due to fragmentary or erroneous sequences using an empirical rule-based approach (described in Methods). The method exploits information from the reference alignments to classify the sequences in each alignment into a number of subfamilies and to construct a representative model for each protein subfamily, including characteristic conserved blocks and

typical start/stop sites. Each subfamily sequence was then compared to the model in turn, in order to identify 'outlier' sequences, with one or more discrepancies. The discrepancies we considered included: (i) divergence of the sequence from conserved core blocks that might indicate badly predicted exons, (ii) insertions that may be due to introns predicted to be coding, (iii) deletions that may be due to missing exons and (iv) potential start and stop site mispredictions. Although the method used here to detect sequence discrepancies may also identify a number of naturally occurring proteins, such as splicing variants, our main goal was to construct a set of reliable sequences for use in the following experiments.

In the first experiment, all the sequences (the reliable sequences and those with discrepancies) were used as input for each MSA program. The alignment quality scores were then calculated based only on the reliable sequences (ignoring the sequences with discrepancies) and compared to the scores obtained in the previous test for all sequences (Figure 5). Significant differences (one-tailed student t-test) were observed for all the MSA programs tested, implying that sequences with discrepancies are aligned less well than reliable sequences.

In the second experiment, the sequences with discrepancies were excluded from the benchmark test sets and each sequence set was realigned using the eight MSA programs. The quality of the resulting alignments was again measured using the CS score (Figure 5). No significant differences were observed for the alignment scores based on the reliable sequences, when sequences with discrepancies were included or excluded from the MSA.

Based on these two experiments, we conclude that the MSA programs tested are capable of accurately aligning the reliable sequences, even in the presence of a large proportion of sequences with discrepancies. Nevertheless, it is important to note that, in the presence of sequences with discrepancies, the subsequent exploitation of the MSA and in particular the identification of family-wide or context-specific motifs, is more complicated. In order to exploit the full potential of the new sequence resources, it is clearly necessary to characterise precisely the conserved segments within these sequences.

MSA program evaluation: alignment of locally conserved motifs

To investigate the ability of the MSA programs to identify context-specific or locally conserved motifs, we typified each individual block in the reference alignments using a number of different features: block length, sequence similarity in the block, the frequency with which the block is observed in the alignment, and the percentage of the block found in a natively disordered region. The alignment quality for each individual block was then measured using the Block Column Score (BCS) described in Methods. Figure 6(A–D) shows the distributions of the block scores obtained by each MSA program for each block attribute. BCS generally increased with increasing block length and increasing sequence similarity, as might be expected. Nevertheless, a correlation analysis (Figure 6A) showed that the programs did not respond in the same way to the different block features. For example, the scores obtained with the program Probcons were highly correlated with the frequency of the blocks, which implies that the blocks found in a small proportion of the sequences were aligned less well than those found in the majority of the sequences. In fact, for blocks found in less than 20% of the sequences, the mean BCS score for Probcons is 0.33, compared to 0.80 for blocks occurring in more than 80% of the sequences. This may be due to the probabilistic consistency-based objective function used in Probcons, which incorporates multiple sequence

Figure 1. An example benchmark alignment. (A) Reference alignment of representative sequences of the p53/p63/p73 family, with the domain organization shown above the alignment (AD: activation domain, Oligo: oligomerization, SAM: sterile alpha motif). Colored blocks indicate conserved regions. The grey regions correspond to sequence segments that could not be reliably aligned and white regions indicate gaps in the alignment. (B) Different MSA programs produce different alignments, especially in the N-terminal region (boxed in red in A) containing rare motifs and a disordered proline-rich domain.

doi:10.1371/journal.pone.0018093.g001

conservation information during the alignment of pairs of sequences. The defaultversion of Muscle and TCoffee were also affected by the frequency of the blocks. In the case of Muscle, this may be related to the iterative refinement stage, since the fast version with only 2 iterations was less sensitive. In contrast, Probcons and Muscle (default) were less sensitive than the other programs to the similarity of the blocks. The localization of the

block in a natively disordered region had an adverse effect on the scores obtained by all the programs tested. Thus, blocks with more than 20% of the residues in natively disordered segments were aligned with BCS scores less than 0.5 by all aligners. This is in agreement with our original observation that most MSA programs available today are designed to align the globular, folded domains in proteins.
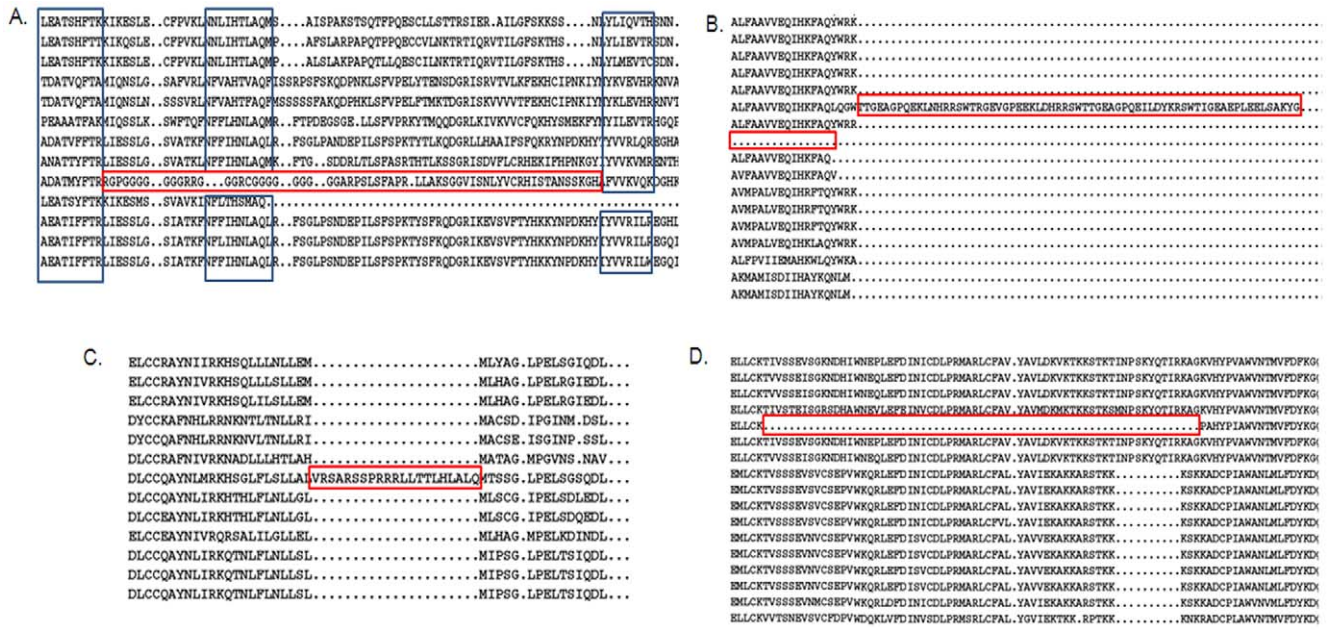
Figure 2. Examples of sequence discrepancies detected. Four types of sequence discrepancies are identified and highlighted by red boxes in the subfamily alignments. A. Potential mispredicted exons are predicted based on the scores of the conserved core blocks (blue boxes) in the subfamily alignment. Here, the ninth sequence contains a segment 'outlier' that scores below the defined threshold for the central core block. The region of the sequence identified as a discrepancy is extended to the nearest core blocks in which the sequence is correctly aligned. B. Potential start and stop site errors are predicted based on the distribution of the positions of the N/C-terminal residues. C. Identification of a potential inserted intron, based on the presence of a single sequence with the insertion in a given subfamily. D. Identification of a potential missing exon, based on the presence of a single sequence with a deletion in a given subfamily.
doi:10.1371/journal.pone.0018093.g002

Improving local alignment quality by combining methods

The experiments described above demonstrated some of the strengths and weaknesses of the different MSA construction methods. For a given set of sequences, different MSA programs often provide very different solutions, particularly outside the most conserved regions, as illustrated in Figure 1B. In order to test whether these differences could be exploited to improve local alignment accuracy, we determined a new score for each block, corresponding to the highest score obtained by any of the programs. This combined score was then compared to the block scores for each program individually (Figure 7). Of course, our combined score is a theoretical maximum, since it incorporates knowledge about the blocks from the reference alignment. Identifying *ab initio* conserved regions in different alignments and combining them in a single consensus alignment is more complicated. Nevertheless, the combined score represents a significant improvement over all the individual methods, with an increase in accuracy of almost 20%.

Based on these combined scores, we posed the following question: can we model or predict the 'alignability' of specific blocks based on the attributes we have defined here? In other words, can we use these attributes to distinguish the blocks that can be aligned from those that cannot? The good news is that, by exploiting the individual capabilities of the recent algorithmic developments, a new milestone is attained where the globular domains present in a majority of the sequences can be accurately aligned (Figure 8). Even short blocks (<10 residues)with low similarity (<0.5) can be aligned with 40–60% accuracy (Figure 8A). However, the frequency of occurrence in the alignment plays an important role. Blocks that occur in a majority of the sequences, even very divergent ones, are generally well aligned (Figure 8B).

Short blocks (<10 residues) that occur in a majority of the sequences are also well aligned (Figure 8C). Blocks in natively disordered regions are generally less well aligned than those in folded regions, and short, divergent blocks are misaligned by all programs (Figure 8D–F).

Discussion

We have used a new alignment benchmark to investigate whether MSA programs are capable of constructing high quality alignments for the sequences resulting from modern biotechnologies. The overall alignment quality scores obtained by the different programs generally confirmed the trends observed in previous benchmark studies. One notable exception was the fact that increasing the number of sequences in the alignment did not lead to more accurate alignments on average. We hypothesize that this is due to the greater complexity of the large alignments, generally representing divergent protein families with complex domain organisations and an increased number of fragmentary and erroneous sequences.

A more detailed study of local alignment quality then allowed us to highlight a number of differences in the MSA methods tested. For example, for very divergent blocks, Mafft (linsi), TCoffee and Probcons were more successful. The local alignment method, Dialign-tx, and Kalign performed better for blocks that were conserved in small subsets of the sequences, while Mafft (linsi) achieved the highest scores for short blocks less than 10 residues long. Based on these results, we demonstrated that better alignment accuracy could be achieved by combining the strengths of the different programs. Unfortunately, the alignment accuracy still decreases when the domains are found less frequently in the alignment.In the future, new approaches will be needed to
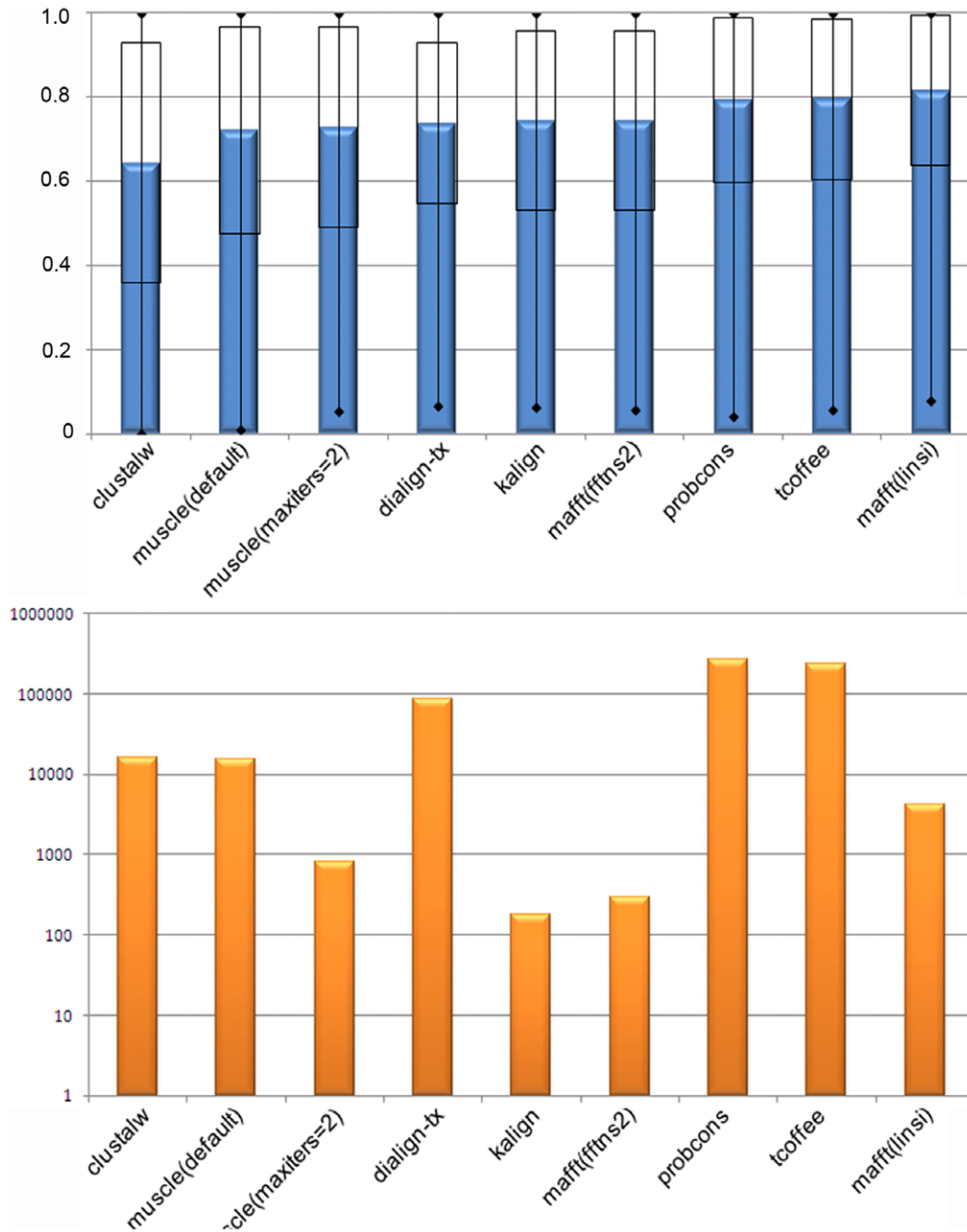
Figure 3. Overall alignment performance for each of the MSA programs tested. (A) Overall alignment quality measured using CS. Programs are shown ranked by increasing quality scores. Error bars correspond to one standard deviation.(B) Total run time for constructing all alignments (a log10 scale is used for display purposes).
doi:10.1371/journal.pone.0018093.g003

specifically address the problems of identifying the subfamily- or context-specific motifs and other blocks that occur less frequently in the alignment, and to handle the noise introduced by the numerousfragmentary and erroneous sequences.

There are a number of alternative solutions for coping with this additional complexity. First, assuming that the fragmentary and/ or erroneous sequences can be identified, they can be excluded from the alignment, although this would discard a significant amount of information. Second, the missing or erroneous portions of the sequences can be predicted [41]. This however is difficult without the information from the alignment itself. Third, new algorithms and programs can be developed to handle the specific

characteristics of the new sequences. Work in this direction has begun, with the development for example, of enhanced database searching algorithms such as CARMA [42], or MEGAN [43] that are more robust to the sequencing errors common in high throughput sequencing projects. In the MSA field, some aligners, such as Kalign, TCoffee or Probcons, provide estimators of local alignment accuracy that could be used to identify unreliable regions and eliminate them from subsequent analyses. The sensitivity/specificity of these accuracy scores has not been fully evaluated yet, although a comprehensive test could be performed using simulated sequences, where the true homology relationships between all sequence residues are known.
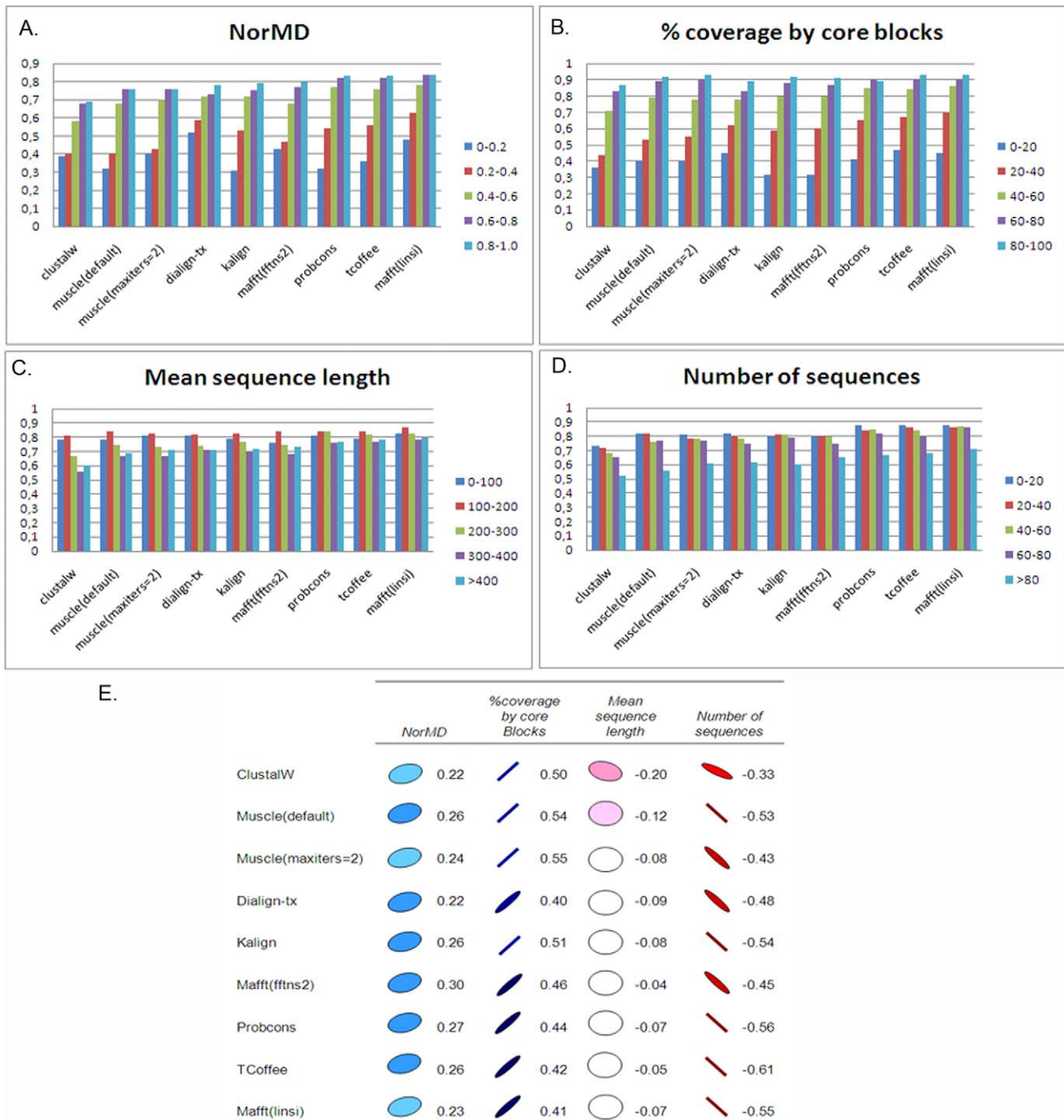
Figure 4. Factors affecting overall alignment quality. Average alignment quality scores (CS) for each MSA program tested and for each global alignment attribute:(A) CS versus NorMD, (B) CS versus the percentage of the alignment covered by the blocks, (C) CS versus mean sequence length, (D) CS versus the total number of sequences.(E) Pearson correlation coefficients of overall quality scores (CS) for each program with global alignment attributes (blue: positive correlation, red: negative correlation).

doi:10.1371/journal.pone.0018093.g004

The alignment of blocks in the natively disordered regions is even more problematic. This is probably because the default parameters used in most MSA programs have been optimized on alignments of globular, folded domains, and most of the benchmarks used to evaluate the programs are based on structural superpositions of these domains. Although the 3D fold gives important clues to function, it does not represent the whole protein [38,39]. The unstructured regions contain important regulatory signals, such as cellular localization or post-transcriptional modification sites, and many others waiting to be discovered. A number of groups have recently begun to develop new statistical models to represent many of these signals [44,45] and it will be crucial to incorporate these models in future MSA programs.
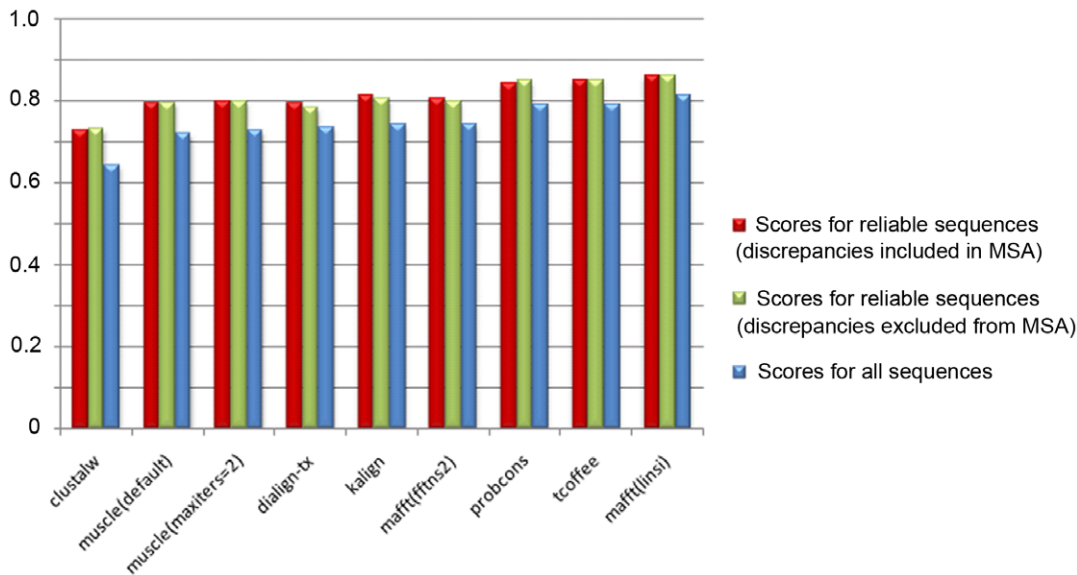
Figure 5. Comparison of alignment quality scores for sequence sets with and without potential error sequences. Quality scores (CS) for alignment of reliable sequences when discrepancies are included in the alignment set are shown in red. Quality scores for the same set of sequences when discrepancies are removed from the alignment set are shown in green. Scores for all sequences (from figure 2) are shown (in blue) for comparison purposes.
doi:10.1371/journal.pone.0018093.g005

So far, we have considered only the alignment of the conserved blocks that could be identified reliably, which cover less than 50% of the total alignment. The structural and/or functional roles of the remaining regions (shown in grey in Figure 1A) are still largely out of reach. We can draw parallels here with the evolving view of the human genome. When the genome was first sequenced, less than 5% of it was considered functional, the rest being 'junk DNA'. Now, it is known that this so-called 'dark matter' does in fact contain numerous functional elements [46].

It is clear that the sequence alignment field now needs to evolve to cope with the challenges posed by the overwhelming flood of data. We have shown that the partitioning of the alignment into well characterised blocks allows a judicious combination of complementary methods resulting in more accurate alignments, particularly in the less well conserved regions. These alignments will in turn allow to highlight both conserved family signatures and specific regions that might suggest neo- or sub-functionalization, or other important genetic events. The next generation of MSA methods will undoubtedly incorporate other novel approachesthat will allow us to reveal the detailed picture of a gene's function and evolution in the context of their complex interaction and regulatory networks. We propose two major directions for future developments. First, the definition of alignment and block attributes opens the way to the exploitation of the latest developments in the field of statistical pattern recognition and data mining, aimed at extracting interesting or informative correlations (rules, regularities, patterns or constraints) from large data sets. Some recent research in this area has focused on the identification of rare patterns e.g.[47] and the problems of how to differentiate valid rare patterns from noise. Second, MSA algorithms can benefit from the new structural and functional "omics" data. In the same way that 2D and 3D structure information has already been used in methods such as 3D-COFFEE [26] or Refiner [27], or information from database homology searches in programs such as PRALINE[48]orPRO-MALS [49], other important data resources could be exploited to shed light on the unstructured and other 'grey' regions. For

example, information about cellular localizationor specific molecular interactions could be used to guide the search for specific signals in these complex sequences.

Integration of these different algorithmic approaches and data types in knowledge-enabled, dynamic systems will ease and improve the complete MSA construction and analysis process; from the selection of a suitable set of sequences, via data cleaning and preprocessing, data mining and the evaluation of results, to the final knowledge presentation and visualization. Such systems could then be used to fully exploit the potential of MSAs as models of the underlying evolutionary processes that have created and fashioned extant genes and fine-tuned their structure, function and regulation.

Materials and Methods

Construction of reference alignments

The protein families used as benchmark test sets were selected to provide a variety of different multiple alignment problems (Figure 9). Thus, the number of sequences in each alignment ranges from 4 to 807. The mean sequence length for an alignment ranges from 56 to 3271 and mean residue percent identity ranges from 11 to 68. Detailed alignment statistics are available at ftp://ftp-igbmc.u-strasbg.fr/pub/msa_reference/stats.txt.

For each family, the reference alignment was constructed using a semi-automatic protocol similar to the one developed for the construction of the BAliBASE [50] alignment benchmark. Briefly, potential sequence homologs were detected by PSI-BLAST [51] searches in the Uniprot [52] and PDB [53] databases using a given query sequence. Of the 218 reference alignments, 122 (56%) have at least one sequence with known structure.Sequences with known 3D structure were then aligned using the SAP [54] 3D superposition program. Sequences with no known 3D structure were initially aligned by (i) identifying the most conserved segments in the PSI-BLAST HSP alignments with the Ballast [55] program and (ii) using these conserved segments as anchors for the progressive multiple alignment strategy implemented in DbClustal [56]. Unrelated sequences were removed from the
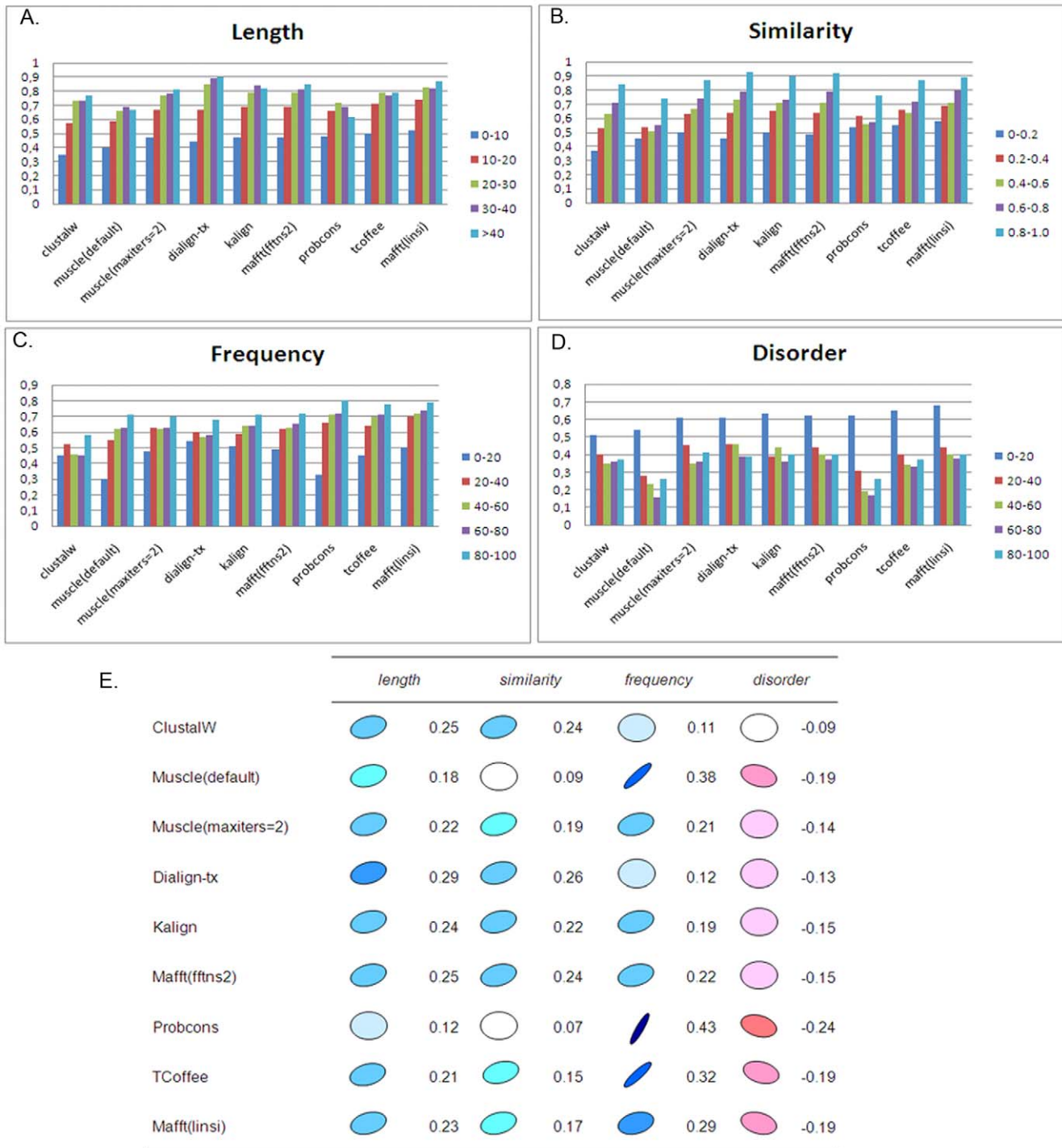
Figure 6. Factors affecting individual block alignment quality. Average block scores (BCS) for each MSA program and for each block attribute:(A) BCS versus similarity (= 1-MD) of the sequences in the block, (B) BCS versus block length: average residue length of the block, (C) BCS versus frequency of occurrence of the block in the alignment, (D) BCS versus disorder: percentage of residues in natively disordered regions compared to folded domains.(E) Correlation of individual block scores (BCS)for each program with the various block attributes.
doi:10.1371/journal.pone.0018093.g006

multiple alignment using the LEON [57] program and the quality of the alignment was evaluated using the NorMD objective function. Finally, structural and functional annotations (including known domains from the Interpro database: www.ebi.ac.uk/interpro/) were added using the multiple alignment information management system (MACSIMS) [58].

The automatic alignment was then manually verified and refined to correct any badly aligned sequences or locally misaligned regions. The manual refinementincluded the alignment of known secondary structure elements and functional residues.At this stage, a subset of the complete set of sequences detected in the database searches was selected to ensure that the benchmark
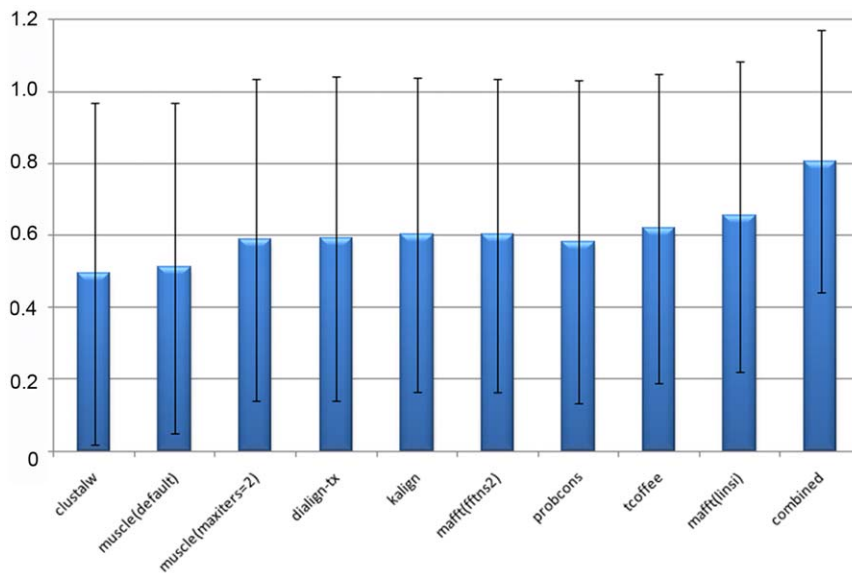
Figure 7. Comparison of block scores obtained by the different alignment programs. Mean block scores for the individual programs vary between 0.49 and 0.65. Combining the results from each program leads to an increased mean score of 0.81. Error bars correspond to one standard deviation. Asterisks indicate significant differences between the scores according to pairwise t-tests (significance level 0.05). doi:10.1371/journal.pone.0018093.g007

contains test sets of different sizes, thus representing a wide diversity of alignment problems. Alignments were edited with the JalView [59] editor which allows the user to visualize alignment conservation via various residue coloring schemes as well as conservation and consensus plots. The conserved regions were also explored according to the structural and functional information available for the sequence family.

Alignment block calculation

For each reference alignment, blocks are defined that correspond to the reliably aligned regions, using the RASCAL [60] program. Briefly, the alignment is first divided horizontally into sequence subfamilies using Secator [61]. For each subfamily, sequence conservation is measured using the NorMD objective function in a sliding window analysis (window length = 5) along the length of the alignment. A block is then defined as a region in the alignment consisting of at least 3 columns, in which the NorMD is above the threshold value of 0.2. For each block in each subfamily, a profile [62] is built from the alignment and pairwise profile-profile comparisons are made to identify blocks shared between a number of subfamilies. This protocol is similar to the method used to identify blocks in the previous BAliBASE alignment benchmark [50], although in this case only regions conserved in all the sequences were marked as blocks.

This protocol led to the identification of 7985 blocks, representing on average 46% of the total multiple alignment (coverage ranged from <20% to >80%). The remainder of the sequence segments could not be aligned reliably based only on the sequences and structures present in the alignment. Thus, the blocks exclude local segments that are either (i) unalignable by sequence alone or (ii) not biologically alignable.

Global alignment attributes

Four different attributes were calculated for each reference alignment, which reflect the overall difficulty of the alignment:

i.   the total number of sequences to be aligned,

ii.   the average length of the sequences to be aligned,

iii. the norMD score which is an objective function for MSA based on the Mean Distance (MD) scores introduced in ClustalX [63]. A score for each column in the alignment is calculated using the concept of continuous sequence space introduced by Vingron and Sibbald [64] and the column scores are then summed over the full length of the alignment. The norMD scores also take into account the size of the alignment by calculating the maximum score attainable given the lengths of each of the unaligned sequences and assuming that the sequences are all identical.

iv.   the percentage of the alignment covered by the blocks.

Block attributes

Four different attributes were calculated for each block in each reference alignment:

i.   the average similarity of the sequence segments in the block is estimated using: Similarity = 1-MD, where MD = mean distance [40] of the sequences in the block,

ii.   the length of the block, corresponding to the average number of residues for each sequence in the block,

iii. the frequency of occurrence of the block in the alignment, equal to the number of sequences in the block divided by the total number of sequences in the alignment,

iv.   the structural context of the block, measured by the percentage of the residues in the block found in a predicted natively disordered (unstructured) region. Natively disordered segments were predicted using the IUPred program [65].

Although the benchmark test sets are designed to represent many different alignment problems, the sampling of the four attributes described here is not always homogeneous. For example, the test sets contain few blocks in disordered regions, which are also long or which occur frequently in the alignments. This results
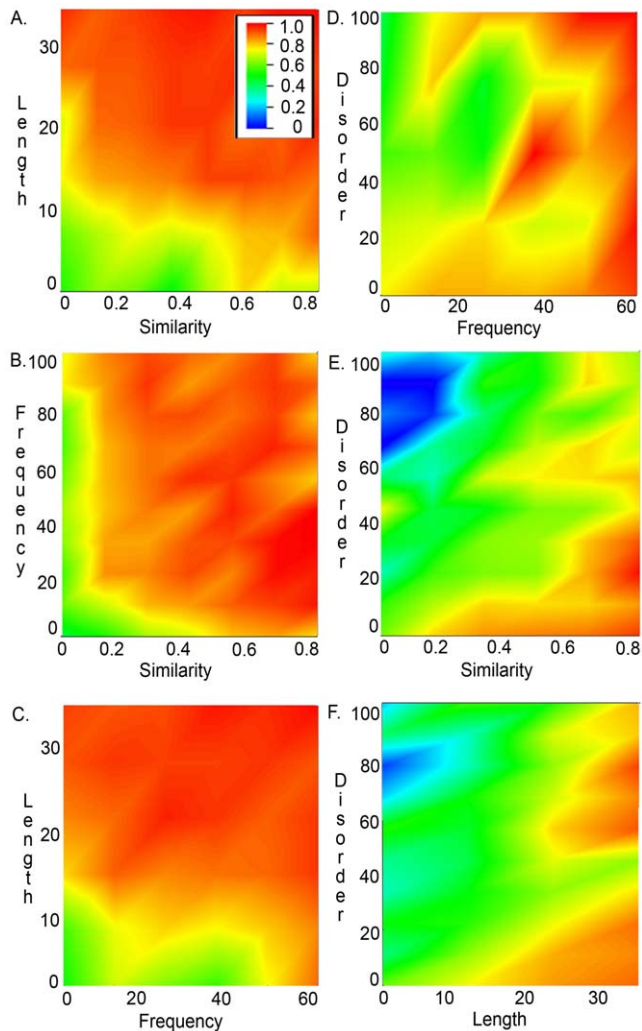
Figure 8. Alignability of blocks depends on various attributes.
By combining 8 different MSA programs, a majority of blocks can be
well aligned (red regions in the heat maps), but certain blocks remain
problematic (blue, green regions). (A) Short blocks (<10 residues) with
low similarity (<0.5) are aligned with 40–60% accuracy. (B) The
frequency of occurrence in the alignment plays an important role.
Blocks that occur in a majority of the sequences, even very divergent
ones, are generally well aligned. (C) Short blocks (<10 residues) that
occur in a majority of the sequences are also well aligned. (D to F)
Blocks in natively disordered regions are generally less well aligned than
those in folded regions, and short, divergent blocks are misaligned by
all programs (blue regions).
doi:10.1371/journal.pone.0018093.g008

in some heterogeneity in the subsequent analyses, such as the
results shown in figure 3 in the main text.

Detection of sequence discrepancies

The sequences in the benchmark test sets were extracted from
the public protein databases and may contain errors resulting from
inaccurate gene structure prediction. Different types of prediction
error were considered, such as excluding coding exons, including
introns as part of the coding sequence, or wrongly predicting start
and termination sites. We used the information in the reference
multiple alignment to build a model of the protein family and
sequences that deviated from this model were annotated as having
potential sequence errors.

The sequences in the complete alignment were first divided into
more related subfamilies using the Secator program [61]. Then,
for each subfamily, sequences with discrepancies that might
indicate errors in the corresponding gene structure, were identified
using an empirical rule-based approach:

1. Badly predicted exons are identified using the RASCAL
   algorithm [60] as 'outlier' sequence segments. The method is
   summarized here and in Figure 2A. First, conserved 'core
   blocks' are identified for the subfamily, representing the
   sequence segments that are reliably aligned in the majority of
   the sequences within the subfamily. Then, for each core block,
   a weighted profile is built from the alignment and each
   sequence within the subfamily is assigned a score against the
   profile. Finally, a threshold score for each core block is defined
   based on the upper and lower quartiles of the sequence scores.
   Sequence segment outliers that score below the threshold are
   annotated as 'discrepancies' or potential errors.

2. Badly predicted start or stop sites are identified by considering
   the positions of the N/C-terminal residues for each sequence in
   the subfamily alignment (Figure 2B). For each sequence, the
   position of the terminal residue in the alignment is noted. A
   window, W, of 'normal' values is then determined, as follows:
   $Q_1-10<W<Q_3+10$, where Q_1 and Q_3 are the lower and upper
   quartiles respectively of the distribution of terminal positions.
   Sequences with terminal positions outside this window are
   annotated as potential deletion/extension errors.

3. Inserted introns (Figure 2C) are detected using the following
   rule: a potential inserted intron is detected if two subfamily
   alignment columns (i,j) exist such that $((n_i = N_i)$ AND $(n_j = N_j)$
   AND $(N_k = 1$ for $i<k<j)$ AND $(j-i> = 10))$, where N_i is the total
   number of sequences in the subfamily (excluding fragments at
   column i), n_i is the number of residues in column i.

4. Missing exons (Figure 2D) are detected using the following rule:
   a potential missing exon is detected if two subfamily alignment
   columns (i,j) exist such that $((n_i = N_i)$ AND $(n_j = N_j)$ AND
   $(N_k = N-1$ for $i<k<j)$ AND $(j-i> = 10))$, where N_i is the total
   number of sequences in the subfamily (excluding fragments at
   column i), n_i is the number of residues in column i.

Multiple alignment programs evaluated

The latest versions of 8 different multiple alignment programs
(see below) were used to construct an alignment for each of the
benchmark test sets. The programs were run using the default
options for protein alignment, except for Mafft and Muscle. Mafft
is a suite of programs offering various multiple alignment
strategies, of which two complementary versions were tested: a
rapid, less accurate version (fftns2) and an iterative refinement
(linsi). For Muscle, two versions were tested: a fast, average
accuracy version that limits the refinement to a maximum of 2
iterations (iters = 2), and the default options, which limits the
refinement to a maximum of 16 iterations. The parallel version of
TCoffee was run on 8 processors. Thus, a total of eight different
versions of the alignment programs were tested (Table 1).

All programs were run on a Sun Enterprise V40z server (4
Opteron processors with 4×16 Gb memory) under RedHat
Enterprise Linux.

Evaluation procedure

Overall alignment quality scores. The alignments
obtained from each of the 8 programs were compared to the
corresponding reference alignments. Suppose we have a test
alignment of N sequences and M blocks. For each block, b in the
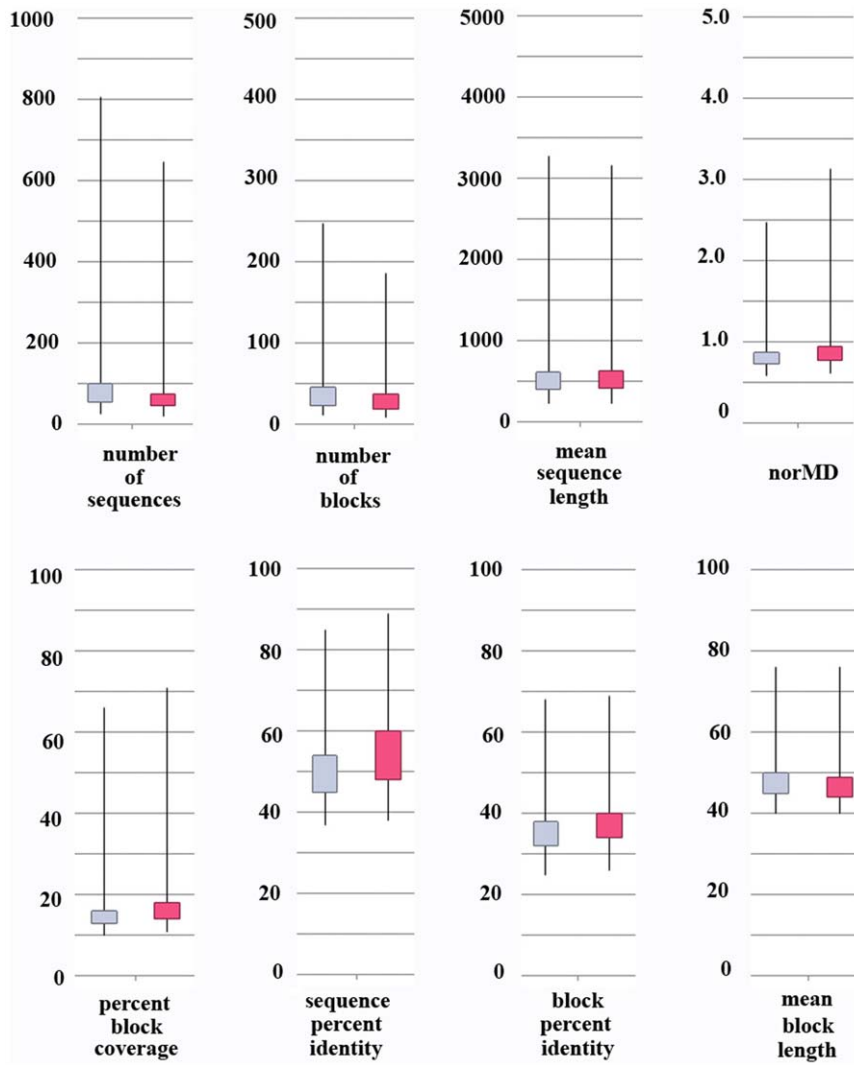
Figure 9. General statistics computed for the benchmark alignments. In the box-and-whisker plots, boxes indicate lower and upper quartiles, and whiskers represent minimum and maximum values. Blue boxes correspond to the alignment of all sequences. Red boxes correspond to the alignments containing only reliable sequences, with no identified sequence discrepancies.
doi:10.1371/journal.pone.0018093.g009

Table 1. Multiple sequence alignment programs used in this study.

Program	version	Availability
ClustalW[67]	2.0.12	www.clustal.org
Dialign-tx [68]	1.0.2	dialign-tx.gobics.de
Kalign [69]	2.03	msa.cgb.ki.se
Mafft (fftns2) [70]	6.815	align.bmr.kyushu-u.ac.jp/mafft/software
Mafft (linsi) [70]	6.815	align.bmr.kyushu-u.ac.jp/mafft/software
Muscle (iters = 2) [71]	3.8.31	www.drive5.com/muscle
Muscle (default) [71]	3.8.31	www.drive5.com/muscle
T-Coffee (parallel)[72]	8.99	www.tcoffee.org
Probcons [73]	1.12	probcons.stanford.edu

doi:10.1371/journal.pone.0018093.t001

alignment containing n_b sequences and m_b columns, the ith column of the block is assigned a score $C_{bi} = 1$ if all the residues in the column are aligned correctly, otherwise $C_{bi} = 0$. The score for each block ($= C_{bi}$ averaged over its columns) is then weighted by the number of sequences in the block. The overall alignment quality, or Column Score (CS), is then:

$$CS = \frac{\sum_{b=1}^{M} \dfrac{n_b \sum_{i=1}^{m_b} C_{bi}}{m_b}}{\sum_{b=1}^{M} n_b}$$

Block alignment quality scores. For each block, b in the alignment containing n_b sequences and m_b columns, the ith column of the block is again assigned a score $C_{bi} = 1$ if all the residues in the column are aligned correctly, otherwise $C_{bi} = 0$. The ability of the programs to align a specific block was estimated

by calculating the block column score, (BCS) = mean column score in the block:

$$BCS = \frac{\sum_{i=1}^{m_b} C_{bi}}{m_b}$$

In this case, the block column scores are not weighted by the number of sequences in the block. Instead, each block has a maximum score of 1, regardless of the frequency with which it is observed in the alignment.

Combining block alignment quality scores for different programs

For each reference alignment, a "combined score" was calculated corresponding to the maximal score possible if all correctly aligned blocks from each program were combined in a single alignment. For each block in the reference alignment, the maximum score obtained by any of the programs was selected and these maximal block scores were then averaged over the whole alignment.

Availability

Unaligned sequences for all the reference alignments are available in FASTA format from ftp://ftp-igbmc.u-strasbg.fr/pub/msa_reference/msa_reference.tar.gz. The annotated alignments, including the block definitions, are provided in an XML format based on the MAO Multiple Alignment Ontology [66] and used by the MACSIMS systems [56]. The source code for the scoring schemes used here is available from ftp://ftp-igbmc.u-strasbg.fr/pub/msa_reference/bali_score_src_v4.tar.gz.

Author Contributions

Conceived and designed the experiments: JDT OP. Performed the experiments: JDT BL. Analyzed the data: JDT BL OL OP. Wrote the paper: JDT BL OL OP.

References

1. Harvey PH, Pagel MD (1991) The Comparative Method in Evolutionary Biology. Oxford University Press Paris.
2. Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, et al. (2008) Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452: 745–749.
3. Bao Y, Bolotov P, Dernovoy D, Kiryutin B, Zaslavsky L, et al. (2008) The influenza virus resource at the National Center for Biotechnology Information. J Virol 82: 596–601.
4. Kuipers RK, Joosten HJ, van Berkel WJ, Leferink NG, Rooijen E, et al. (2010) 3DM: systematic analysis of heterogeneous superfamily data to discover protein functionalities. Proteins 78: 2101–2113.
5. Singh S, Tokhunts R, Baubet V, Goetz JA, Huang ZJ, et al. (2009) Sonic hedgehog mutations identified in holoprosencephaly patients can act in a dominant negative manner. Hum Genet 125: 95–103.
6. Zhang J, Chen X, Kent M, Rodriguez C, Chen X (2009) Establishment of a dog model for the p53 family pathway and identification of a novel isoform of p21 cyclin-dependent kinase inhibitor. Mol Cancer Res 7: 67–78.
7. Eaton MJ, Martin A, Thorbjarnarson J, Amato G (2009) Species-level diversification of African dwarf crocodiles (Genus Osteolaemus): a geographic and phylogenetic perspective. Mol Phylogenet Evol 50: 496–506.
8. Levasseur A, Pontarotti P, Poch O, Thompson JD (2008) Strategies for reliable exploitation of evolutionary concepts in high throughput biology. Evol Bioinform Online 4: 121–137.
9. Wong KM, Suchard MA, Huelsenbeck JP (2008) Alignment uncertainty and genomic analysis. Science 319: 473–476.
10. Löytynoja A, Goldman N (2008) Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. Science 320: 1632–1635.
11. Brown DP, Krishnamurthy N, Sjolander K (2007) Automated protein subfamily identification and classification. PLoS Comput Biol 3: e160.
12. Brandt BW, Feenstra KA, Heringa J (2010) Multi-Harmony: detecting functional specificity from sequence alignment. Nucleic Acids Res 38 Suppl: W35–40.
13. Rausell A, Juan D, Pazos F, Valencia A (2010) Protein interactions and ligand binding: from protein subfamilies to functional specificity. Proc Natl Acad Sci U S A 107: 1995–2000.
14. Feng DF, Doolittle, RF (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J Mol Evol 25: 351–360.
15. Thompson JD, Plewniak F, Poch O (1999) BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. Bioinformatics 15: 87–88.
16. Gardner PP, Wilm A, Washietl S (2005) A benchmark of multiple sequence alignment programs upon structural RNAs. Nucleic Acids Res 33: 2433–2439.
17. Gotoh O (1996) Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments. J Mol Biol 264: 823–838.
18. Eddy S (1998) Profile hidden Markov models. Bioinformatics 14: 755–763.
19. Notredame C, Higgins DG (1996) SAGA: sequence alignment by genetic algorithm. Nucleic Acids Res 24: 1515–1524.
20. Thompson JD, Plewniak F, Poch O (1999) A comprehensive comparison of multiple sequence alignment programs. Nucleic Acids Res 27: 2682–2690.
21. Blackshields G, Wallace IM, Larkin M, Higgins DG (2006) Analysis and comparison of benchmarks for multiple sequence alignment. In Silico Biol 6: 321–339.
22. Wallace IM, O'Sullivan O, Higgins DG, Notredame C (2006) M-Coffee: combining multiple sequence alignment methods with T-Coffee. Nucleic Acids Res 34: 1692–1699.
23. Katoh K, Toh H (2008) Recent developments in the MAFFT multiple sequence alignment program. Brief Bioinform 9: 286–298.
24. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792–1797.
25. Do CB, Mahabhashyam MS, Brudno M, Batzoglou S (2005) ProbCons: Probabilistic consistency-based multiple sequence alignment. Genome Res 15: 330–340.
26. O'Sullivan O, Suhre K, Abergel C, Higgins DG, Notredame C (2004) 3DCoffee: combining protein sequences and structures within multiple sequence alignments. J Mol Biol 340: 385–395.
27. Chakrabarti S, Lanczycki C, Panchenko A, Przytycka T, Thiessen P, et al. (2006) Refining multiple sequence alignments with conserved core regions. Nucleic Acids Res 34: 2598–2606.
28. Lee C, Grasso C, Sharlow MF (2002) Multiple sequence alignment using partial order graphs. Bioinformatics 18: 452–464.
29. Raghava GP, Searle SM, Audley PC, Barber JD, Barton GJ (2003) OXBench: a benchmark for evaluation of protein multiple sequence alignment accuracy. BMC Bioinformatics 4: 47.
30. Dessimoz C, Gil M (2010) Phylogenetic assessment of alignments reveals neglected tree signal in gaps. Genome Biol 11: R37.
31. Aniba MR, Poch O, Thompson JD (2010) Issues in bioinformatics benchmarking: the case study of multiple sequence alignment. Nucleic Acids Res 38: 7353–7363.
32. Koonin EV (2009) Darwinian evolution in the light of genomics. Nucleic Acids Res 37: 1011–1034.
33. Bakke P, Carney N, Deloache W, Gearing M, Ingvorsen K, et al. (2009) Evaluation of three automated genome annotations for Halorhabdus utahensis. PLoS One 4: e6291.
34. Keller O, Odronitz F, Stanke M, Kollmar M, Waack S (2008) Scipio: using protein sequences to determine the precise exon/intron structures of genes and their orthologs in closely related species. BMC Bioinformatics 9: 278.
35. Guigo R, Flicek P, Abril JF, Reymond A, Lagarde J, et al. (2006) EGASP: the human ENCODE Genome Annotation Assessment Project. Genome Biol 7 Suppl 1: S2 1-31.
36. Mardis ER (2008) The impact of next-generation sequencing technology on genetics. Trends Genet 24: 133–141.
37. Pop M, Salzberg SL (2008) Bioinformatics challenges of new sequencing technology. Trends Genet 24: 142–149.
38. Dunker AK, Oldfield CJ, Meng J, Romero P, Yang JY, et al. (2008) The unfoldomics decade: an update on intrinsically disordered proteins. BMC Genomics 9 S1.
39. Wong WC, Maurer-Stroh S, Eisenhaber F (2010) More than 1,001 problems with protein domain databases: transmembrane regions, signal peptides and the issue of sequence homology. PLoS Comput Biol 6: e1000867.
40. Thompson JD, Plewniak F, Ripp R, Thierry JC, Poch O (2001) Towards a reliable objective function for multiple sequence alignments. J Mol Biol 4: 937–951.

41. Bianchetti L, Thompson JD, Lecompte O, Plewniak F, Poch O (2005) vALId: validation of protein sequence quality based on multiple alignment data. J Bioinform Comput Biol 3: 929–947.

42. Krause L, Diaz NN, Bartels D, Edwards RA, Pühler A, et al. (2006) Finding novel genes in bacterial communities isolated from the environment. Bioinformatics 22: e281–289.

43. Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. Genome Res 17: 377–286.

44. Chica C, Labarga A, Gould CM, López R, Gibson TJ (2008) A tree-based conservation scoring method for short linear motifs in multiple alignments of protein sequences. BMC Bioinformatics 9: 229.

45. Sankararaman S, Sjölander K (2008) INTREPID—INformation-theoretic TREe traversal for Protein functional site Identification. Bioinformatics 24: 2445–2452.

46. Amaral PP, Dinger ME, Mercer TR, Mattick JS (2008) The eukaryotic genome as an RNA machine. Science 319: 1787–1789.

47. Koh YS, Rountree N (2009) Rare Association Rule Mining And Knowledge Discovery: Technologies For Infrequent And Critical Event Detection. (IGI Global, Hershey, PA).

48. Simossis V, Heringa J (2005) PRALINE: a multiple sequence alignment toolbox that integrates homology-extended and secondary structure information. Nucleic Acids Res 33: W289–294.

49. Pei J, Grishin N (2007) PROMALS: towards accurate multiple sequence alignments of distantly related proteins. Bioinformatics 23: 802–808.

50. Thompson JD, Koehl P, Ripp R, Poch O (2005) BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. Proteins 61: 127–136.

51. Schäffer A, Aravind L, Madden T, Shavirin S, Spouge J, et al. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res 29: 2994–3005.

52. The UniProt Consortium (2009) The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Res. In press.

53. Berman HM (2008) The Protein Data Bank: a historical perspective. Acta Cryst A64: 88–95.

54. Taylor WR (2000) Protein structure comparison using SAP. Methods Mol Biol 143: 19–32.

55. Plewniak F, Thompson JD, Poch O (2000) Ballast: blast post-processing based on locally conserved segments. Bioinformatics 16: 750–759.

56. Thompson J, Plewniak F, Thierry J, Poch, O (2000) DbClustal: rapid and reliable global multiple alignments of protein sequences detected by database searches. Nucleic Acids Res 28: 2919–2926.

57. Thompson JD, Prigent V, Poch O (2004) LEON: multiple aLignment Evaluation Of Neighbours. Nucleic Acids Res 32: 1298–1307.

58. Thompson JD, Muller A, Waterhouse A, Procter J, Barton GJ, et al. (2006) MACSIMS: multiple alignment of complete sequences information management system. BMC Bioinformatics 7: 318.

59. Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ (2009) Jalview Version 2—a multiple sequence alignment editor and analysis workbench. Bioinformatics 25: 1189–1191.

60. Thompson JD, Thierry JC, Poch O (2003) RASCAL: rapid scanning and correction of multiple sequence alignments. Bioinformatics 19: 1155–1161.

61. Wicker N, Perrin GR, Thierry JC, Poch O (2001) Secator: a program for inferring protein subfamilies from phylogenetic trees. Mol Biol Evol 18: 1435–1441.

62. Gribskov M, McLachlan AD, Eisenberg D (1987) Profile analysis: detection of distantly related proteins. Proc Natl Acad Sci USA 84: 4355–4358.

63. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucl Acids Res 25: 4876–4882.

64. Vingron M, Sibbald PR (1993) Weighting in sequence space: a comparison of methods in terms of generalized sequences. Proc Natl Acad Sci USA 90: 8777–8781.

65. Dosztányi Z, Csizmok V, Tompa P, Simon I (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics 21: 3433–3434.

66. Thompson JD, Holbrook SR, Katoh K, Koehl P, Moras D, et al. (2005) MAO: a Multiple Alignment Ontology for nucleic acid and protein sequences. Nucleic Acids Res 33: 4164–4171.

67. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, et al. (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23: 2947–2948.

68. Subramanian AR, Kaufmann M, Morgenstern B (2008) DIALIGN-TX: greedy and progressive approaches for segment-based multiple sequence alignment. Algorithms Mol Biol 3: 6.

69. Lassmann T, Frings OS, Sonnhammer, EL (2009) Kalign2: high-performance multiple alignment of protein and nucleotide sequences allowing external features. Nucleic Acids Res 37: 858–865.

70. Katoh K, Toh H (2008) Recent developments in the MAFFT multiple sequence alignment program. Brief Bioinform 9: 286–298.

71. Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5: 113.

72. Notredame C, Higgins DG, Heringa J (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol 302: 205–217.

73. Do CB, Mahabhashyam MS, Brudno M, Batzoglou S (2005) ProbCons: Probabilistic consistency-based multiple sequence alignment. Genome Res 15: 330–340.

Annex III :

Publication n° 5

Toward community standards in the quest for orthologs

Christophe Dessimoz[1,*], Toni Gabaldón[2], David S. Roos[3], Erik L. L. Sonnhammer[4], Javier Herrero[5]; and the Quest for Orthologs Consortium[†]

[1]ETH Zurich and Swiss Institute of Bioinformatics, Universitätstrasse 6, 8092 Zürich, Switzerland, [2]Bioinformatics and Genomics Programme, Centre for Genomic Regulation (CRG), and UPF. Dr. Aiguader, 88, 08003 Barcelona, Spain, [3]Department of Biology and Penn Genome Frontiers Institute, University of Pennsylvania, Philadelphia, PA 19104, USA, [4]Stockholm Bioinformatics Center, Swedish eScience Research Center, Department of Biochemistry and Biophysics, Stockholm University, Science for Life Laboratory, Box 1031, SE-17121 Solna, Sweden, [5]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Associate Editor: Alfonso Valencia

ABSTRACT

The identification of orthologs—genes pairs descended from a common ancestor through speciation, rather than duplication—has emerged as an essential component of many bioinformatics applications, ranging from the annotation of new genomes to experimental target prioritization. Yet, the development and application of orthology inference methods is hampered by the lack of consensus on source proteomes, file formats and benchmarks. The second 'Quest for Orthologs' meeting brought together stakeholders from various communities to address these challenges. We report on achievements and outcomes of this meeting, focusing on topics of particular relevance to the research community at large. The Quest for Orthologs consortium is an open community that welcomes contributions from all researchers interested in orthology research and applications.

Contact: dessimoz@ebi.ac.uk

1 INTRODUCTION

The concepts of orthology and paralogy are central to comparative genomics. These terms were coined more than four decades ago (Fitch, 1970) to distinguish between two classes of gene homology: those descended from a common ancestor by virtue of a speciation event (orthologs) versus those that diverged by gene duplication (paralogs). This distinction permits accurate description of the complex evolutionary relationships within gene families including members distributed across multiple species. Detection of orthology and paralogy has become an essential component of diverse applications, including the reconstruction of evolutionary relationships across species (reviewed in Delsuc *et al.*, 2005), inference of functional gene properties (e.g. Chen and Jeong, 2000; Hofmann, 1998; Tatusov *et al.*, 1997), and identification and testing of proposed mechanisms of genome evolution (e.g. Mushegian and Koonin, 1996; Tatusov *et al.*, 1997). In today's context, with the number of fully sequenced genomes growing by the day, accurate

and efficient inference of orthology has become an imperative. A plethora of computational methods have been developed for inferring orthologous relationships, many of which provide their predictions in form of web-accessible databases (reviewed in Alexeyenko *et al.*, 2006; Gabaldón, 2008; Koonin, 2005; Kristensen *et al.*, 2011).

In 2009, the first Quest for Orthologs meeting was organized to bring together scientists working in the fields of orthology inference, genome annotation and genome evolution to exchange ideas, tackle common challenges, aiming at removing barriers and redundancy (Gabaldón *et al.*, 2009). The main objectives identified were concerted effort toward standardized formats, datasets and benchmarks, and establishment of continuous communication channels including a mailing list, a website and a regular meeting.

Following the first Quest for Ortholog meeting in 2009, a second meeting was held in June 2011, bringing together 45 participants from 27 different institutions on 3 continents, representing >20 orthology databases (http://questfororthologs. org/orthology_databases). The meeting was structured to include plenary sessions devoted to topics of general interest (reference datasets, orthology detection methodology, practical applications of orthology), and additional discussions focusing on benchmarking, standardized formats, alternative transcripts, ncRNA orthology, etc. In this letter, we summarize the discussions and specific outcomes of the meeting, as well as some of the most important achievements of the Quest for Orthologs community in the past 2 years.

2 DEFINITIONS AND EVOLUTIONARY MODELS

Orthology finds application in multiple, diverse research areas. Depending on the context, the reasons for identifying orthologous genes can vary considerably, sometimes driving the use of subtly differing definitions of orthology and its extension to groups of genes. Brigitte Boeckmann (Swiss Inst Bioinformatics, Geneva, Switzerland) and Christophe Dessimoz (ETH Zürich, Switzerland) reviewed the definitions and objectives of orthologous groups within a unifying framework and discussed the implications of these differences for the interpretation and benchmarking of ortholog databases (Boeckmann *et al.*, 2011). The need for clear evolutionary definitions is particularly acute for multidomain proteins, as their underlying coding sequences often have distinct, and even conflicting, evolutionary histories. In an attempt to salvage

the gene as the fundamental evolutionary unit, Dannie Durand (Carnegie Mellon University, Pittsburgh, USA) proposed a model of gene homology based on the genomic locus, not the constitutive nucleotides of the gene (Song *et al.*, 2008).

3   DEBATING THE 'ORTHOLOG CONJECTURE'

The 'ortholog conjecture'—that at a similar degree of sequence divergence, orthologs are generally more conserved in function than paralogs—has been a prevailing paradigm, originally supported by theory rather than empirical studies. At the previous Quest for Orthologs meeting, Bill Pearson (University Virginia, Charlottesville, USA) questioned the ortholog conjecture and contended that the sequence similarity be the primary determinant of functional conservation (Gabaldón *et al.*, 2009). Several studies have now been undertaken to compare the properties of orthologs versus paralogs, and generally appear to support the importance of distinguishing orthologs from paralogs.

Erik Sonnhammer (Stockholm University, Sweden) reported significant support for the ortholog conjecture based on conserved domain architecture (Forslund *et al.*, 2011) and intron positions (Henricson *et al.*, 2010). David Roos (University of Pennsylvania, Philadelphia, USA) showed that protein structure is significantly more conserved for orthologs than for paralogs, particularly within protein active sites. Indeed, it is even possible to quantify the importance of orthology, in terms of sequence conservation or RMSD, for structural modeling (Peterson *et al.*, 2009). Toni Gabaldón (Center for Genomic Regulation, Barcelona, Spain) and colleagues found that human–mouse orthologs exhibit more conserved tissue expression than paralogs of a similar age (Huerta-Cepas *et al.*, 2011). Similarly, Klaas Vandepoele (Ghent University, Belgium) reported that for 77% of orthologs between *Arabidopsis* and rice, the expression patterns were more highly conserved than the background distribution, and that expression patterns can also be used to tease out functional similarity even among in-paralogs (Movahedi *et al.*, 2011).

In other tests, however, orthologs were not found to be functionally more conserved than paralogs. Just days before the meeting, Nehrt *et al.* (2011) reported that Gene Ontology (GO) functional annotations (du Plessis *et al.*, 2011) may be less similar among orthologs than among paralogs, and that human–mouse co-expression data across tissues argues against the ortholog conjecture. Discussion at the meeting noted an inherent bias favoring conservation between homologs in the same species, which may inflate the scores of paralogs. Furthermore, using correlation coefficients as a measure of gene expression conservation may also cause problems (Pereira *et al.*, 2009). Overall, this discussion suggests that the debate remains far from being settled.

4   INNOVATIONS IN ORTHOLOGY INFERENCE: INCREMENTAL METHODS AND META-METHODS

Much of the meeting focused on innovations in orthology inference. One trend involves the application of incremental methods, minimizing the need to recompute results as new datasets are added. Ikuo Uchiyama (National Institute for Basic Biology, Okazaki, Japan) described how the Microbial Genome Database (MBGD) uses such an approach to cope with new genomes, and also

to identify orthologs in metagenomic samples (Uchiyama *et al.*, 2010). Likewise, the most recent release of the OrthoMCL database permits new genes (and even entire genomes) to be assigned to putative ortholog groups (Chen *et al.*, 2006). Ingo Ebersberger (CIBIV, Vienna, Austria) showed how an incremental approach based on hidden Markov models can be used to identify orthologs in EST libraries, which typically only cover a fraction of all genes (Ebersberger *et al.*, 2009), and Radek Szklarczyk (2012) introduced a new profile-based iterative procedures that pushes the boundaries of reliable homology detection and helps identify disease genes in human.

Another trend involves the application of meta-methods to integrate predictions from multiple datasets, combining their strengths so as to outperform any single underlying method. Michiel Van Bel (Ghent University, Belgium) presented an ensemble method intended to detect orthologs in plant species combining different orthology inference methods—a notorious challenge due to extensive whole genome duplication and paleopolyploidy. This concept lies at the heart of the PLAZA database (Proost *et al.*, 2009). Michael S. Livstone (Princeton University, USA) described how the P-POD database (Heinicke *et al.*, 2007) enables users to compare orthology and paralogy predictions from multiple homology inference methods on 12 reference genomes from the Gene Ontology Consortium (Reference Genome Group of the Gene Ontology Consortium, 2009). With MetaPhOrs, Gabaldón showed that combining the orthologs inferred from several large-scale phylogenetic resources is not only meaningful to increase the total number of predictions, but also to assess the accuracy based on the consistency across different sources (Pryszcz *et al.*, 2011).

5   STANDARDS AND BENCHMARKING

A primary motivation for this meeting has been to establish standards for efficient data exchange in the orthology community. Until now, virtually every ortholog database has used a different format, posing a major impediment for consumers of orthology data, including annotators and for comparative genomicists. Likewise, the source data for orthology analysis (proteomes) has used a variety of formats (mostly *ad hoc* variations of the Fasta format). To resolve these issues, a working group has developed XML-based formats for both sequence and orthology data (OrthoXML and SeqXML, respectively) (Schmitt *et al.*, 2011). These formats were endorsed by meeting participants, representing many orthology databases, and by the reference proteome project. Documentation and tools are available at http://OrthoXML.org and http://SeqXML.org.

Following on from suggestions at the previous meeting, the Quest for Orthologs 'Reference Proteomes' serves as a common dataset to compare orthology inference methods. Eleanor Stanley (EBI, Hinxton, UK) gave an overview of UniProt's commitment to curate this dataset. Meeting participants suggested that an annual release schedule would be appropriate, and should ensure that most methods are applied to a common and reasonably current dataset. Although driven by the need to benchmark ortholog detection algorithms against a common dataset, we anticipate that the reference proteome project will be useful beyond the orthology prediction community. For example, UniProt curators are eager to test how different ortholog predictions against a consistent dataset can be used to facilitate protein annotation. Complementing the reference proteome project, Raja Mazumder (Georgetown University, Washington,

USA) presented an automated approach to identify representative proteomes—relatively small subsets of all proteomes that capture most of the information available (Chen *et al.*, 2011).

The availability of standardized datasets should significantly ease the challenge of sourcing genomes faced by all providers of ortholog detection, and holds great promise for orthology inference benchmarking. Indeed, previous benchmarking studies have been forced to evaluate orthology predictions based on inconsistent datasets (Altenhoff and Dessimoz, 2009; Boeckmann *et al.*, 2011; Hulsen *et al.*, 2006; Trachana *et al.*, 2011), or have been limited to comparatively small datasets analyzed only by methods available as stand-alone programs (Chen *et al.*, 2007; Salichos and Rokas, 2011). Leveraging the Reference Proteomes, Adrian Altenhoff (ETH, Zürich, Switzerland) presented a web server prototype for orthology benchmarking. The service gathers predictions submitted by ortholog providers and runs a battery of tests, such as an assessment of how well the predictions satisfy a standard definition of orthology (Fitch, 1970), and a test assessing accuracy in predicting GO function annotations (du Plessis *et al.*, 2011).

6   FUNCTIONAL PREDICTIONS

One of the chief benefits of ortholog group assignment is the potential for inferring putative function—particularly as new sequencing methodologies make it increasingly possible to assemble genomes and define genes from species where experimental data is lacking. Such computational inference can be risky, however, as the accuracy of existing annotations is often unknown, particularly for electronically assigned annotations, leading to rampant *in silico* propagation of errors (Gilks *et al.*, 2002). Paul Thomas (USC, Los Angeles, USA) outlined activities of the Gene Ontology (GO) Reference Genomes Project (Reference Genome Group of the Gene Ontology Consortium, 2009), and described a pilot project assigning GO terms to internal nodes of a reference tree (Gaudet *et al.*, 2011). Incorporating a concept of evolutionary breadth (and confidence) into the annotation process would greatly enhance the specificity of orthology-based inference. Nives Škunca (ETH, Zürich, Switzerland) reported an innovative effort to estimate the quality of electronic GO annotations, by tracking changes in stability, coverage and specificity over time. This study suggests a strategy for identifying high confidence electronic annotations that can be relied upon for transitive inference. The availability of a web-based platform for comparing the performance of orthology detection methods (see above) should greatly facilitate the assessment of functional prediction performance. In addition, the development of a curated catalog of ortholog genes with similar function, using experimental data, such as RNAi, expression data or mutant phenotype, would be a useful resource and could improve functional prediction.

7   ADDITIONAL TOPICS

Homology prediction based on similarity is a prerequisite for many orthology prediction methods, and a workshop was held to discuss current approaches and upcoming challenges in assessing sequence similarity. Much discussion was devoted to the need for more realistic models of sequence evolution, which would enable the proper assessment of what level of similarity is expected for two evolutionary related sequences. Tina Koestler (CIBIV,

Vienna, Austria) and Jean-Baka Domelevo (LIRMM, Montpellier, France) presented profile-based models of evolution, taking into account particularities of functional or structural regions of protein sequences. Further discussions stressed the necessity of elucidating the mode of evolution of multidomain proteins, particularly in the context of domain rearrangements. In a different take on homology inference, Vincent Miele (LBBE, Lyon, France) reported new methodology to identify robust homologous groups from the structure of similarity networks.

Orthology inference has been traditionally focused on the study of protein coding genes, but there is increasing interest in applying similar analyses to non-coding RNAs (ncRNAs). For example, both Ensembl (Flicek *et al.*, 2011) and miROrtho (Gerlach *et al.*, 2009) have started to provide orthology predictions for a subset of ncRNAs, largely based on synteny. Most of the discussion centered on the difficulties in use of phylogenetic methods for the analysis of ncRNAs: phylogenetic models used for protein coding genes usually assume that sites evolve independently, but ncRNAs often violate this assumption, owing to the importance of secondary structure conservation. Several models specifically developed for RNA sequences have been implemented in phylogenetic packages [e.g. PHASE (Gowri-Shankar and Rattray, 2007) or RAxML (Stamatakis, 2006)], but these models are not widely known. Other limitations hindering phylogenetic study of ncRNAs, include the difficulty in reliably detecting these genes. The RFam database (Gardner *et al.*, 2011) contains a high-quality set of ncRNA families, but its scope is limited to families for which an expert multiple alignment is available. A central repository for RNA sequences has been recently proposed (Bateman *et al.*, 2011) and we see this as important for boosting interest and helping to drive evolutionary studies on RNA sequences.

8   ACHIEVEMENTS AND OUTLOOK

The disparate but interconnected communities represented at this meeting have taken an important step toward better understanding one another. Inferring orthology is a non-trivial task, for many reasons. There are certainly significant computational and algorithmic challenges, but at a more basic level, differing applications driving the quest for orthologs has led to differing definitions of orthology (particularly with respect to subcategories, such as in-paralogs or co-orthologs), the use of different source datasets and different metrics for evaluating performance. The most important achievement to emerge from the Quest for Orthologs effort thus far is a series of consensus agreements, on:

- reference proteome datasets, including a minimal set suggested for benchmarking ortholog detection algorithms, and a larger set, greatly facilitating data sourcing;

- data exchange formats, including OrthoXML and SeqXML; and

- an analysis platform providing for comparison of developer-supplied ortholog calls using diverse metrics (include metrics supplied by users and developers).

The many different uses of orthology detection ensure that there will continue to be a multitude of useful algorithms. Some will be optimized for computational efficiency and/or scalability. Some

will focus on specific phylogenetic groups, which may be highly homogenous or relatively diverse, may or may not exhibit synteny and may include introns or operons, etc. Still other methods will be tailored to handle multidomain proteins, alternative transcription units, metagenomics data, etc. (Dessimoz, 2011).

The availability of reference datasets permits all groups to use the same proteomes, while also minimizing the effort to source the raw data. The OrthoXML format allows predictions to be exchanged efficiently, and the benchmarking platform permits consistent assessment of the results. One of the highlights of the June 2011 meeting was the discussion of orthology prediction methods—a discussion that could only take place because different algorithms were applied to the same source data. Proposed benchmarks are publicly accessible from the Quest for Orthologs portal (http://questfororthologs.org), in order to encourage other researchers to use this platform.

It will be exciting to see the progress of Quest for Orthologs initiatives over the coming years—the next meeting is tentatively scheduled for 2013. In the meantime, the reference proteomes will be updated and enlarged to sample taxonomic space, and the benchmarking service will be made publicly available. We invite all interested parties to join the orthology community, using the contacts available at the aforementioned Quest for Orthologs portal.

REFERENCES

Alexeyenko,A. *et al.* (2006) Overview and comparison of ortholog databases. *Drug Discov Today*, **3**, 137–143.

Altenhoff,A.M. and Dessimoz,C. (2009) Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Comput. Biol.*, **5**, e1000262.

Bateman,A. *et al.* (2011) RNAcentral: a vision for an international database of RNA sequences. *RNA*, **17**, 1941–1946.

Boeckmann,B. *et al.* (2011) Conceptual framework and pilot study to benchmark phylogenomic databases based on reference gene trees. *Brief. Bioinform.*, **12**, 423–435.

Chen,R. and Jeong,S. (2000) Functional prediction: identification of protein orthologs and paralogs. *Protein Sci.*, **9**, 2344–2353.

Chen,F. (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.*, **34**, D363–D368.

Chen,F. *et al.* (2007) Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS One*, **2**, e383.

Chen,C. *et al.* (2011) Representative proteomes: a stable, scalable and unbiased proteome set for sequence analysis and functional annotation. *PLoS One*, **6**, e18910.

Delsuc,F. *et al.* (2005) Phylogenomics and the reconstruction of the tree of life. *Nat. Rev. Genet.*, **6**, 361–375.

Dessimoz,C. (2011) Editorial: orthology and applications. *Brief. Bioinform.*, **12**, 375–376.

du Plessis,L. *et al.* (2011) The what, where, how and why of gene ontology–a primer for bioinformaticians. *Brief. Bioinform.*, **12**, 723–735.

Ebersberger,I. *et al.* (2009) HaMStR: Profile hidden markov model based search for orthologs in ESTs. *BMC Evol. Biol.*, **9**, 157.

Fitch,W. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool.*, **19**, 99–113.

Flicek,P. *et al.* (2011) Ensembl 2011. *Nucleic Acids Res.*, **39**, D800–D806.

Forslund,K. *et al.* (2011) Domain architecture conservation in orthologs. *BMC Bioinformatics*, **12**, 326.

Gabaldón,T. (2008) Large-scale assignment of orthology: back to phylogenetics? *Genome Biol.*, **9**, 235.

Gabaldón,T. *et al.* (2009) Joining forces in the quest for orthologs. *Genome Biol.*, **10**, 403.

Gardner,P.P. *et al.* (2011) Rfam: Wikipedia, clans and the "decimal" release. *Nucleic Acids Res.*, **39**, D141–D145.

Gaudet,P. *et al.* (2011) Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. *Brief. Bioinform.*, **12**, 449–462.

Gerlach,D. *et al.* (2009) miROrtho: computational survey of microRNA genes. *Nucleic Acids Res.*, **37**, D111–117.

Gilks,W.R. *et al.* (2002) Modeling the percolation of annotation errors in a database of protein sequences. *Bioinformatics*, **18**, 1641–1649.

Gowri-Shankar,V. and Rattray,M. (2007) A reversible jump method for bayesian phylogenetic inference with a nonhomogeneous substitution model. *Mol. Biol. Evol.*, **24**, 1286–1299.

Heinicke,S. *et al.* (2007) The Princeton Protein Orthology Database (P-POD): a comparative genomics analysis tool for biologists. *PLoS One*, **2**, e766.

Henricson,A. *et al.* (2010) Orthology confers intron position conservation. *BMC Genomics*, **11**, 412.

Hofmann,K. (1998) Protein classification and functional assignment. *Trends Guide Bioinformatics*, 18–21.

Huerta-Cepas,J. *et al.* (2011) Evidence for short-time divergence and long-time conservation of tissue-specific expression after gene duplication. *Brief. Bioinform.*, **12**, 442–448.

Hulsen,T. *et al.* (2006) Benchmarking ortholog identification methods using functional genomics data. *Genome Biol.*, **7**, R31.

Koonin,E.V. (2005) Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.*, **39**, 309–338.

Kristensen,D.M. *et al.* (2011) Computational methods for Gene Orthology inference. *Brief. Bioinform.*, **12**, 379–391.

Movahedi,S. *et al.* (2011) Comparative network analysis reveals that tissue specificity and gene function are important factors influencing the mode of expression evolution in arabidopsis and rice. *Plant Physiol.*, **156**, 1316–1330.

Mushegian,A.R. and Koonin,E.V. (1996) A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc. Natl Acad. Sci. USA*, **93**, 10268–10273.

Nehrt,N.L. *et al.* (2011) Testing the ortholog conjecture with comparative functional genomic data from mammals. *PLoS Comput. Biol.*, **7**, e1002073.

Pereira,V. *et al.* (2009) A problem with the correlation coefficient as a measure of gene expression divergence. *Genetics*, **183**, 1597–1600.

Peterson,M.E. *et al.* (2009) Evolutionary constraints on structural similarity in orthologs and paralogs. *Protein Sci.*, **18**, 1306–1315.

Proost,S. *et al.* (2009) PLAZA: a comparative genomics resource to study gene and genome evolution in plants. *Plant Cell*, **21**, 3718–3731.

Pryszcz,L.P. *et al.* (2011) MetaPhOrs: orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score. *Nucleic Acids Res.*, **39**, e32.

Reference Genome Group of the Gene Ontology Consortium (2009) The Gene Ontology's Reference Genome Project: a unified framework for functional annotation across species. *PLoS Comput. Biol.*, **5**, e1000431.

Salichos,L. and Rokas,A. (2011) Evaluating ortholog prediction algorithms in a yeast model clade. *PLoS One*, **6**, e18755.

Schmitt,T. *et al.* (2011) SeqXML and OrthoXML: standards for sequence and orthology information. *Brief. Bioinform*.

Song,N. *et al.* (2008) Sequence similarity network reveals common ancestry of multidomain proteins. *PLoS Comput. Biol.*, **4**, e1000063.

Stamatakis,A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, **22**, 2688–2690.

Szklarczyk,R. *et al.* (2012) Iterative orthology prediction uncovers new mitochondrial proteins and identifies C12orf62 as the human ortholog of COX14, a protein involved in the assembly of cytochrome c oxidase. Genome Biol. (in press).

Tatusov,R.L. *et al.* (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.

Trachana,K. *et al.* (2011) Orthology prediction methods: A quality assessment using curated protein families. *BioEssays*, **33**, 769–780.

Uchiyama,I. *et al.* (2010) MBGD update 2010: toward a comprehensive resource for exploring microbial genome diversity. *Nucleic Acids Res.*, **38**, D361–D365.

Annex IV :

Publication n° 6

KD4v: comprehensible knowledge discovery system for missense variant

Tien-Dao Luu, Alin Rusu, Vincent Walter, Benjamin Linard, Laetitia Poidevin, Raymond Ripp, Luc Moulinier, Jean Muller, Wolfgang Raffelsberger, Nicolas Wicker, Odile Lecompte, Julie D. Thompson, Olivier Poch and Hoan Nguyen*

Laboratoire de Bioinformatique et Génomique Intégratives, Institut de Génétique et de Biologie Moléculaire et Cellulaire, 67404 Illkirch, France

ABSTRACT

A major challenge in the post-genomic era is a better understanding of how human genetic alterations involved in disease affect the gene products. The KD4v (Comprehensible Knowledge Discovery System for Missense Variant) server allows to characterize and predict the phenotypic effects (deleterious/neutral) of missense variants. The server provides a set of rules learned by Induction Logic Programming (ILP) on a set of missense variants described by conservation, physico-chemical, functional and 3D structure predicates. These rules are interpretable by non-expert humans and are used to accurately predict the deleterious/neutral status of an unknown mutation. The web server is available at http://decrypthon.igbmc.fr/kd4v.

INTRODUCTION

A wide variety of human diseases have been linked to non-synonymous SNPs (nsSNPs), also called Missense Variants, which result in an alteration of the amino acid sequence of the encoded protein and can affect the function, solubility or structure of the mutated protein. Today, with the huge amount of protein information available in various biomedical databases, it is now possible to better understand the correlation between a nsSNP and the associated phenotypes.

Several methods (1) have been developed to predict the effects of nsSNPs on the 3D structure of a protein and its function, based on the hypothesis that variants that modify the structure/function at the molecular level are more likely to be deleterious. The methods can be divided into two main categories: (i) sequence-based methods using multiple sequence alignments and incorporating different approaches to quantify residue

conservation: SIFT (2), PANTHER (3), SNAP(4) and SNP/GO (5) and (ii) methods combining sequence and 3D structure features such as the widely used Polyphen-2 (6), nsSNPAnalyzer (7) and SNPs3D (8). Most of these methods can classify a nsSNP as either deleterious (strong functional effect) or neutral (weak functional effect) with high accuracy. However, they only provide a final score and in general, no information is provided that could be used to evaluate the classification and to estimate the relationships between genotypic and phenotypic variation.

To overcome these limitations, the KD4v (Comprehensible Knowledge Discovery System for Missense Variant) server aims to discover, exploit and provide the user with links between the computed impact of a mutation and the human disease phenotype. We applied the ILP method (9) to a set of nsSNPs involved in human diseases that are mapped to 3D structure and annotated by the MSV3d (MisSense Variant mapped to 3D structure) pipeline (10). KD4v provides two complementary services: (i) a knowledgebase consisting of ILP rules based on 16 sequence/structure/evolution predicates that characterize deleterious mutations in any human gene and that can be interpreted by biologists and (ii) a tool for mutation prediction based on the ILP rules with performances similar to the most widely used methods: PolyPhen-2 and SIFT. In addition, the KD4v server links the human genes to a rich set of up-to-date information encompassing tissue expression, protein-protein interactions or phenotypic descriptions hosted by SM2PH (11).

MATERIALS AND METHODS

Missense variant annotation

The nsSNPs observed in all human proteins were annotated by the MSV3d pipeline, which automatically performs a sequence/structure/evolution analysis and has

*To whom correspondence should be addressed. Tel: +33 3 88 65 32 65; Fax: +33 3 88 65 32 01; Email: nguyen@igbmc.fr

been shown to be robust and efficient (6,12). This includes various parameters which describe, among others, the physico-chemical changes induced by the amino acid substitution, the conservation pattern of the mutated residue, the status of mutated residues with respect to functional features. In KD4v, this multi-level sequence-based characterization of nsSNPs is complemented by parameters related to 3D models or the 3D Fold classification in SCOP (13). This results in pre-computed annotations for over 63 000 known nsSNPs in the 10 713 proteins with known or modelled 3D structures currently available. In addition, the user can also request a prediction for any new or unknown missense variant, if the protein can be mapped to a 3D structure.

The characterization of the background conservation and exploitation of the different types of evolutionary data has been described in detail previously (10). Briefly, we used MACSIMS (14) to annotate a multiple alignment, containing both Uniprot and PDB sequences, with information such as: (i) taxonomic data, (ii) functional descriptions, (iii) known domains or domains similar to a known 3D structure, (iv) potential disordered regions, (v) blocks that do not correspond to disordered regions or known domains but that are conserved at the family or subfamily level and thus may constitute uncharacterized domains and (vi) conservation pattern of domains and residues. If the variant position is mapped to an identified 3D structure, the structural context of each individual mutation is modelled based on several descriptors combining sequence/structure-related data using several software tools such as MODELLER (15), CSU (16), I-Mutant (17). Details of the predicates used in the KD4v server and computational methods/software are provided on the KD4v help page.

Dataset compilation and computer resource

We used the variant set from the Polyphen-2 training set (6) extracted from SwissVar (18) to train and test the KD4v server. Only nsSNPs that are mapped to 3D structures were retained and randomly split into a training set (6000 disease-causing mutations associated with distinct 881 OMIM phenotypes and 2000 neutral polymorphisms) and a first validation set (658 disease-causing variants associated with 311 distinct OMIM phenotypes and 298 neutral polymorphisms). We also created a second validation set (173 disease-causing mutations associated with distinct 39 OMIM phenotypes and 179 neutral polymorphisms), in which not only variants, but also protein sequences, were different in the training and validation sets. Our goal is to predict the deleterious nature of human variants, i.e. those variants associated with disease phenotypes, and it should be noted that these datasets do not specifically identify mutations that have a weaker effect on the function of the protein. The datasets are available for download from our website.

To guarantee a permanent powerful CPU resource for the KD4v server, we deployed the software on the Décrypthon grid (19) including a total of 58 machines and 475 processors under the AIX operating system distributed on six nodes.

Induction logic programming implementation

Induction Logic Programming (ILP) combines Machine Learning and Logic Programming (9). Briefly, given a formal encoding of the background knowledge and a set of examples, an ILP system will derive hypotheses explaining all positive examples and none, or almost none, negative examples. In this approach, logic is used as a language to induce hypotheses from the examples and background knowledge. The result of the learning step is a set of rules represented as logical formula, typically a Prolog program, that can be reused as a prediction service. The creation of the KD4v is based on distinct predicates deduced from the multi-level characterization provided by MSV3d (Supplementary Table S1) and involves various steps detailed in Supplementary Figure S1. We have limited our study to the task of discriminating the mutations linked to human diseases (deleterious) from those associated with the 'polymorphism' term (neutral). Thus, a positive example in Prolog syntax is defined as: 'is_deleterious(m_Q92947.p.Gly390Ala)' which indicates that, in protein Q92947, the replacement of the glycine at position 390 by an alanine is deleterious.

The implementation of the server also includes the optimization of the predicates using a 5-fold cross-validation on the training set with standard performance indicators including sensitivity, specificity, precision, recall, accuracy and F-measure (see legend of Supplementary Table S2 for a complete description). Thus, the final ILP model consists of 16 predicates (Supplementary Table S1) which can be separated into two major types: predicates describing the mutated residue or protein (functional and structural features) and predicates describing the physical, chemical or structural changes introduced by the substitution.

KD4v RULE SERVICE

Currently, the server hosts 111 rules that are comprehensible by humans. These ILP rules can be used, for example, to uncover the relationships between the deleterious effect of a mutation and the multi-class conservation pattern or the type of the physico-chemical alterations (e.g. size, charge and hydrophobicity) introduced by the substitution. Figure 1 shows some induced rules on the web page. To illustrate how to interpret ILP rules, we can consider the humvar398_44 rule:

```
deleterious(A) :-
    modif_charge(A, charge_increase) and
    modif_hydrophobicity(A, hydrophobicity_decrease) and
    secondary_struc(A, helix) and wt_accessibility(A, buried) and
    mut_accessibility(A, buried).
```

This rule states that a mutation A is deleterious if: (i) the charge of the residue is increased by the mutation; (ii) its hydrophobicity is decreased; (iii) the residue is found in a helix; (iv) the wild-type residue is buried; and (v) the mutant residue is also buried. This rule correctly identified 191 (3.18% of the 6000 studied) deleterious mutations, while misclassifying five neutral mutations as

There are total 111 rules.

How to interpret the rules

	Id	If Statement	Then	Coverage Positive	Coverage Negative	Rank
		Enter a key word: [ ] Submit				
⊞	humvar398_8	conservation_class(A, global_conservation_rank_1) and freq_at_pos(A, B) and B>=2.	deleterious(A)	475 (7.92%)	2 (0.1%)	1
⊞	humvar398_42	freq_at_pos(A, B) and B>=2 and secondary_struc(A, other) and wt_accessibility(A, buried).	deleterious(A)	397 (6.62%)	2 (0.1%)	2
⊞	humvar398_35	freq_at_pos(A, B) and B>=3 and secondary_struc(A, other).	deleterious(A)	249 (4.15%)	5 (0.25%)	3
⊞	humvar398_12	g_or_p(A, g_or_p_unchanged) and conservation_class(A, global_conservation_rank_1) and secondary_struc(A, other) and wt_accessibility(A, buried).	deleterious(A)	214 (3.57%)	3 (0.15%)	4
⊞	humvar398_37	modif_charge(A, charge_unchanged) and modif_polarity(A, polarity_increase) and conservation_class(A, global_conservation_rank_1).	deleterious(A)	211 (3.52%)	3 (0.15%)	5
⊞	humvar398_78	modif_hydrophobicity(A, hydrophobicity_decrease) and is_in_site(A, yes) and freq_at_pos(A, B) and B>=2 and secondary_struc(A, other).	deleterious(A)	211 (3.52%)	4 (0.2%)	6
⊞	humvar398_50	g_or_p(A, g_or_p_disparition) and freq_at_pos(A, B) and B>=2 and secondary_struc(A, other).	deleterious(A)	208 (3.47%)	5 (0.25%)	7
⊞	humvar398_11	modif_polarity(A, polarity_increase) and conservation_class(A, global_conservation_rank_2) and mut_accessibility(A, buried) and stability(A, decrease).	deleterious(A)	200 (3.33%)	5 (0.25%)	8
⊞	humvar398_55	modif_charge(A, charge_increase) and modif_polarity(A, polarity_increase) and conservation_class(A, global_conservation_rank_1).	deleterious(A)	196 (3.27%)	5 (0.25%)	9
⊞	humvar398_9	conservation_class(A, global_conservation_rank_1) and gain_contact(A, dc) and wt_accessibility(A, buried).	deleterious(A)	194 (3.23%)	4 (0.2%)	10
⊞	humvar398_44	modif_charge(A, charge_increase) and modif_hydrophobicity(A, hydrophobicity_decrease) and secondary_struc(A, helix) and wt_accessibility(A, buried) and mut_accessibility(A, buried).	deleterious(A)	191 (3.18%)	5 (0.25%)	11

Figure 1. ILP rules. The first column provides a link to the positive (deleterious mutations) and negative (neutral mutations) examples covered by a given rule and that can be seen by clicking on the + icon. The second column provides the rule identifier (Id). The next two columns provide the 'if' and 'then' clauses of the induced rules. The two right most columns indicate the number of positive and negative examples covered by the rule in each row.

deleterious (0.25% of the 2000 neutral mutations in the training set).
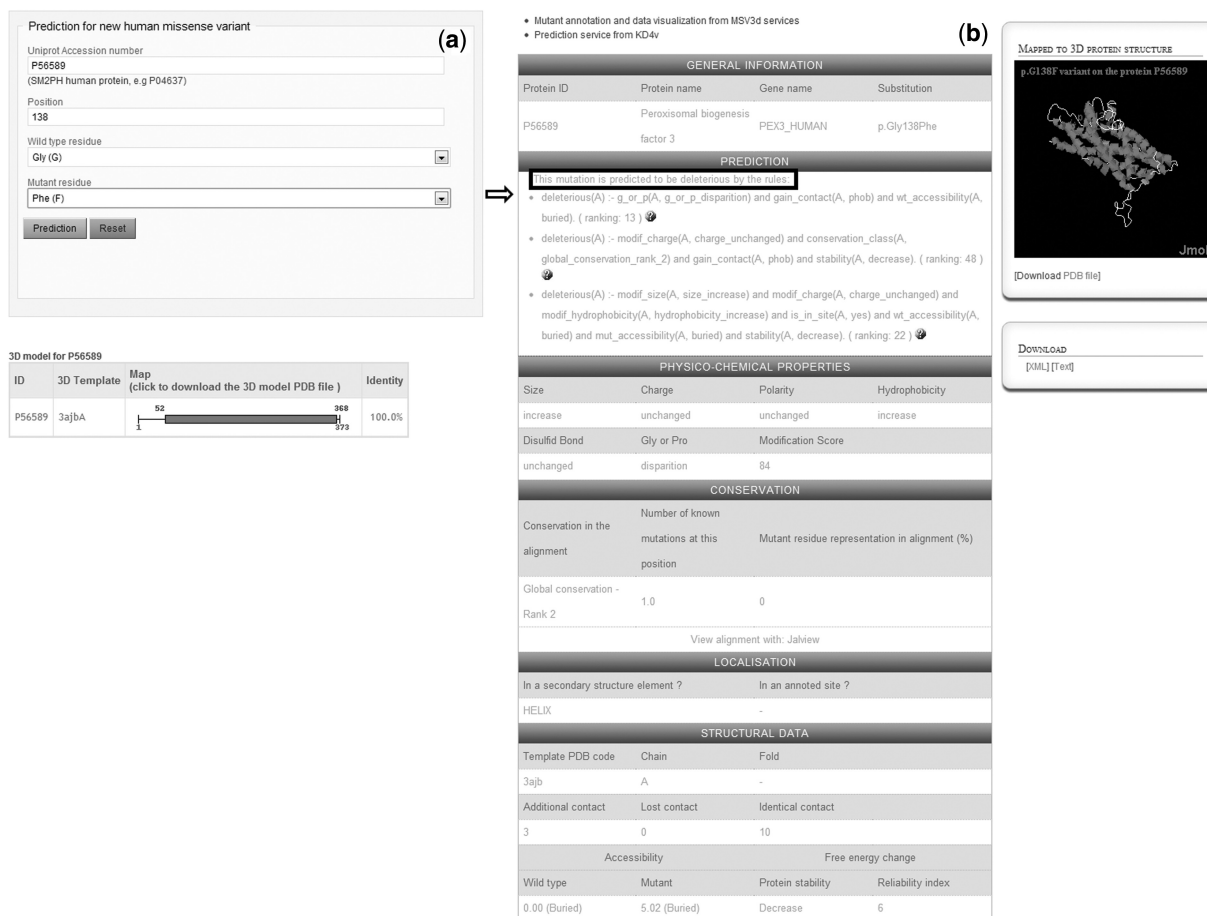
KD4v PREDICTION SERVICE

Input and output

KD4v provides a service aimed at estimating nsSNP effects based on the ILP rules. It can be accessed via the web interface or via the SOAP Web Service, which can be downloaded from the website. The input form of the web interface (Figure 2a) is supported by Ajax to facilitate the identification of the protein accession number and the location of a mutation on the protein sequence or on the schematic 3D map provided. Given the input data, the MSV3d pipeline generates a multi-level characterization of the variant to be predicted. If a 3D model is available, these values are translated into prolog facts, which then become the input for the prediction service. Thanks to the Prolog engine, the deductive reasoning process immediately derives a conclusion (deleterious or neutral nsSNP) with identified rules. Figure 2b shows the KD4v output for the substitution Gly138Phe in the human peroxisomal biogenesis factor 3, predicted to be deleterious. In the 3D model of this protein, which is involved in the Zellweger syndrome, this residue is

buried and located in one of the central helices shaping the protein fold. Analyzing the rule associated with this deleterious prediction, it can be seen that, although this residue is not highly conserved (67% identity which corresponds to the rank2 in our conservation pattern classification), the gain in hydrophobic contact and the decrease in the overall stability might be responsible for the deleterious effect.

Prediction evaluation

We compared the performance of our ILP-based prediction service with two widely used methods: SIFT and PolyPhen-2. The different measures of predictive performance are reported for two independent nsSNP validation sets (Tables 1 and 2). The accuracy (72.28% in Table 1, 75.57% in Table 2) and F-measure (78.61% in Table 1, 71.52% in Table 2) indicate that the KD4v prediction service based on ILP is comparable to SIFT and PolyPhen-2 (although PolyPhen-2 is more accurate on one of the validation sets) and thus represents a competitive alternative solution. Moreover, the KD4v provides ILP rules associated with deleterious predictions that are more interpretable than the previous prediction methods. These rules should help to improve the understanding of

Figure 2. (**a**) Screenshot of the input form of the prediction service. (**b**) Screenshot of the output page providing the prediction results as well as the multi-level characterizations of the mutation. The rules are described if the variant is 'deleterious'. The annotated information related to the mutated position can be visualized in the MSV3d interface on the right.

Table 1. Comparison of prediction methods based on the PolyPhen-2 validation set [658 disease-causing (OMIM phenotype) mutations and 298 neutral polymorphisms]

	TP	FP	FN	TN	Sensitivity	Specificity	Precision	Recall	Accuracy	F-measure
SIFT	398	38	260	260	0.6049	0.8725	0.9128	0.6049	0.6883	0.7276
PolyPhen-2	576	111	77	184	0.8821	0.6237	0.8384	0.8821	0.8017	0.8597
KD4v	487	94	171	204	0.7401	0.6846	0.8382	0.7401	0.7228	0.7861

Table 2. Comparison of prediction methods based on the validation set that excludes proteins present in the training set (173 disease-causing mutations (OMIM phenotype) and 179 neutral polymorphisms)

	TP	FP	FN	TN	Sensitivity	Specificity	Precision	Recall	Accuracy	*F*-measure
SIFT	106	23	67	156	0.6127	0.8715	0.8217	0.6127	0.7443	0.702
PolyPhen-2	139	70	34	109	0.8035	0.6089	0.6651	0.8035	0.7045	0.7278
KD4v	108	21	65	158	0.6243	0.8827	0.8372	0.6243	0.7557	0.7152

the relationships between physico-chemical and structural features and deleterious mutations.

CONCLUSION

The KD4v server uses the available or modelled 3D structures and information provided by the MSV3d pipeline to characterize and predict the phenotypic effect of a mutation. The main advantages of KD4v are (i) valuable predicates and ILP rules associated with the predictions, allowing biologists to identify deleterious mutations and interpret the results, (ii) an ergonomic web interface, incorporating the comprehensive annotation of missense variants, complemented with a SOAP-based remote API

for multiple predictions. Furthermore, the effects of any unknown missense variant (1 of approximately 32 000 000 variants corresponding to all positions of mapped 3D structures and all possible amino acid replacements) can be predicted upon request by the user. In the future, we will extend the background knowledge, first by adding structural surface topology descriptions (20) of the proteins, allowing the precise mapping of different functional regions such as the protein core and the non-interacting or interacting surfaces, and second, by integrating useful knowledge about the functional impact of missense variants from the SNPdbe database (21). Finally, we intend to enhance the prediction performance by combining ILP with other machine learning methods.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1 and 2, Supplementary Figure 1.

ACKNOWLEDGEMENTS

FUNDING

Conflict of interest statement. None declared.

REFERENCES

1. Thusberg,J., Olatubosun,A. and Vihinen,M. (2011) Performance of mutation pathogenicity prediction methods on missense variants. *Hum. Mutat.*, **32**, 358–368.
2. Ng,P.C. and Henikoff,S. (2003) SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.*, **31**, 3812–3814.
3. Thomas,P.D., Campbell,M.J., Kejariwal,A., Mi,H., Karlak,B., Daverman,R., Diemer,K., Muruganujan,A. and Narechania,A. (2003) PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res.*, **13**, 2129–2141.
4. Bromberg,Y., Yachdav,G. and Rost,B. (2008) SNAP predicts effect of mutations on protein function. *Bioinformatics*, **24**, 2397–2398.
5. Calabrese,R., Capriotti,E., Fariselli,P., Martelli,P.L. and Casadio,R. (2009) Functional annotations improve the predictive score of human disease-related mutations in proteins. *Hum. Mutat.*, **30**, 1237–1244.
6. Adzhubei,I.A., Schmidt,S., Peshkin,L., Ramensky,V.E., Gerasimova,A., Bork,P., Kondrashov,A.S. and Sunyaev,S.R. (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.
7. Bao,L., Zhou,M. and Cui,Y. (2005) nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms. *Nucleic Acids Res.*, **33**, W480–W482.
8. Yue,P., Melamud,E. and Moult,J. (2006) SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics*, **7**, 166.
9. Muggleton,S. (1991) Inductive logic programming. *N. Gen Comput*, **8**, 295–318.
10. Luu,T.D., Rusu,A.M., Walter,V., Ripp,R., Moulinier,L., Muller,J., Toursel,T., Thompson,J.D., Poch,O. and Nguyen,H. (2012) MSV3d: database of human MisSense variants mapped to 3D protein structure. *Database J. Biol. Databases Curation*, **2012**, bas018.
11. Friedrich,A., Garnier,N., Gagniere,N., Nguyen,H., Albou,L.P., Biancalana,V., Bettler,E., Deleage,G., Lecompte,O., Muller,J. *et al.* (2010) SM2PH-db: an interactive system for the integrated analysis of phenotypic consequences of missense mutations in proteins involved in human genetic diseases. *Hum. Mutat.*, **31**, 127–135.
12. Plewniak,F., Bianchetti,L., Brelivet,Y., Carles,A., Chalmel,F., Lecompte,O., Mochel,T., Moulinier,L., Muller,A., Muller,J. *et al.* (2003) PipeAlign: A new toolkit for protein family analysis. *Nucleic Acids Res.*, **31**, 3829–3832.
13. Andreeva,A., Howorth,D., Chandonia,J.M., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2008) Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.*, **36**, D419–D425.
14. Thompson,J.D., Muller,A., Waterhouse,A., Procter,J., Barton,G.J., Plewniak,F. and Poch,O. (2006) MACSIMS: multiple alignment of complete sequences information management system. *BMC Bioinformatics*, **7**, 318.
15. Eswar,N., Eramian,D., Webb,B., Shen,M.Y. and Sali,A. (2008) Protein structure modeling with MODELLER. *Methods Mol. Biol.*, **426**, 145–159.
16. Sobolev,V., Sorokine,A., Prilusky,J., Abola,E.E. and Edelman,M. (1999) Automated analysis of interatomic contacts in proteins. *Bioinformatics*, **15**, 327–332.
17. Capriotti,E., Fariselli,P. and Casadio,R. (2005) I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res.*, **33**, W306–W310.
18. Mottaz,A., David,F.P., Veuthey,A.L. and Yip,Y.L. (2010) Easy retrieval of single amino-acid polymorphisms and phenotype information using SwissVar. *Bioinformatics*, **26**, 851–852.
19. Bard,N., Bolze,R., Caron,E., Desprez,F., Heymann,M., Friedrich,A., Moulinier,L., Nguyen,N.H., Poch,O. and Toursel,T. (2010) Decrypthon grid - grid resources dedicated to neuromuscular disorders. *Stud. Health Technol. Informat.*, **159**, 124–133.
20. Albou,L.P., Poch,O. and Moras,D. (2011) M-ORBIS: mapping of molecular binding sites and surfaces. *Nucleic Acids Res.*, **39**, 30–43.
21. Schaefer,C., Meier,A., Rost,B. and Bromberg,Y. (2012) SNPdbe: constructing an nsSNP functional impacts database. *Bioinformatics*, **28**, 601–602.

Annex V :

Awards

PRIX DE LA MEILLEURE PRÉSENTATION ORALE

13èmes Journées Ouvertes en Biologie, Informatique et Mathématiques

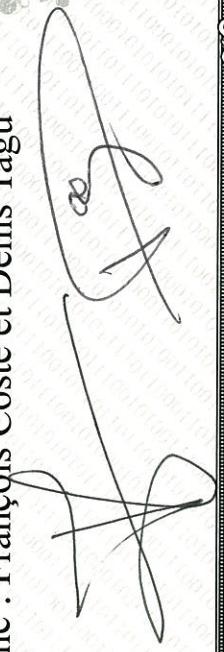Rennes, 3-6 juillet 2012

Prix de la meilleure présentation orale

attribué à :

Benjamin Linard

pour sa présentation intitulée :

EvoluCode : an original view of Human Systems Evolution

Rennes, le 6 juillet 2012

Présidents du comité de programme : François Coste et Denis Tagu

INRIA
MICROSOFT RESEARCH

CENTRE DE RECHERCHE COMMUN

JOBIM 2012

Benjamin LINARD

**Développement de méthodes
évolutionnaires d'extraction de
connaissance et application à des
systèmes biologiques complexes**

Résumé

La biologie des systèmes s'est beaucoup développée ces dix dernières années, confrontant plusieurs niveaux biologiques (molécule, réseau, tissu, organisme, écosystème…). Du point de vue de l'étude de l'évolution, elle offre de nombreuses possibilités. Cette thèse porte sur le développement de nouvelles méthodologies et de nouveaux outils pour étudier l'évolution des systèmes biologiques tout en considérant l'aspect multidimensionnel des données biologiques. Ce travail tente de palier un manque méthodologique évidant pour réaliser des études haut-débit dans le récent domaine de la biologie évolutionnaire des systèmes. De nouveaux messages évolutifs liés aux contraintes intra et inter processus ont été décrites. En particulier, mon travail a permis (i) la création d'un algorithme et un outil bioinformatique dédié à l'étude des relations évolutives d'orthologie existant entre les gènes de centaines d'espèces, (ii) le développement d'un formalisme original pour l'intégration de variables biologiques multidimensionnelles permettant la représentation synthétique de l' histoire évolutive d'un gène donné, (iii) le couplage de cet outil intégratif avec des approches mathématiques d'extraction de connaissances pour étudier les perturbations évolutives existant au sein des processus biologiques humains actuellement documentés (voies métaboliques, voies de signalisations…).

Keywords : orthologie, extraction de connaissance, évolution, réseaux biologiques

Summary

Systems biology has developed enormously over the 10 last years, with studies covering diverse biological levels (molecule, network, tissue, organism, ecology…). From an evolutionary point of view, systems biology provides unequalled opportunities. This thesis describes new methodologies and tools to study the evolution of biological systems, taking into account the multidimensional properties of biological parameters associated with multiple levels. Thus it addresses the clear need for novel methodologies specifically adapted to high-throughput evolutionary systems biology studies. By taking account the multi-level aspects of biological systems, this work highlight new evolutionary trends associated with both intra and inter-process constraints. In particular, this thesis includes (i) the development of an algorithm and a bioinformatics tool dedicated to comprehensive orthology inference and analysis for hundreds of species, (ii) the development of an original formalism for the integration of multi-scale variables allowing the synthetic representation of the evolutionary history of a given gene, (iii) the combination of this integrative tool with mathematical knowledge discovery approaches in order to highlight evolutionary perturbations in documented human biological systems (metabolic and signalling pathways...).

Keywords : orthology, knowledge extraction, evolution, biological networks