UNIVERSITÉ LOUIS PASTEUR STRASBOURG

Thèse présentée pour obtenir le titre de

Docteur de l'Université Louis Pasteur Strasbourg 1

Discipline : Sciences du Vivant

Spécialité : Bioinformatique

par

Julie THOMPSON-MAALOUM

De l'évolution de l'alignement multiple : vers une exploitation efficace des données et une extraction des connaissances à l'ère post-génomique

(Evolution of multiple alignments : towards efficient data exploitation and knowledge extraction in the post-genomique era)

Soutenue publiquement le 22 novembre 2006 devant le jury :

Directeur de thèse	Olivier POCH, Directeur de recherche, IGBMC, Strasbourg
Co-Directeur de thèse	Patrice KOEHL, Professeur, UC Davis, California
Rapporteur interne	Jean-Luc SOUCIET, Professeur, ULP, Strasbourg
Rapporteur externe	Desmond HIGGINS, Professeur, University College, Dublin
Rapporteur externe	Miguel ANDRADE, Health Research Institute, Ottawa
Examinateur	Eric WESTHOF, IBMC, Strasbourg

Remerciements

Je tiens à exprimer ma profonde reconnaissance à messieurs Miguel Andrade, Desmond Higgins, Jean-Luc Souciet et Eric Westhof pour l'honneur qu'ils me font de juger cette thèse.

Je voudrais également exprimer ma sincère gratitude à Dino Moras qui m'a offert la possibilité de rejoindre le Laboratoire de Biologie et de Génomique Structurales et m'a toujours soutenue dans mon travail.

Je tiens à remercier Jean-Claude Thierry pour ses conseils judicieux et pour l'intérêt qu'il a porté à mon travail.

Un grand merci à Patrice Koehl qui a partagé ses connaissances structurales avec moi et qui m'a fait voir le monde en trois dimensions.

Et bien sûr, un grand merci à Olivier Poch, sans qui je ne serais pas là où je suis aujourd'hui ! Merci pour son enthousiasme et son énergie débordante. Merci pour la confiance qu'il m'a accordée et qui m'a fait avancer dans la science, mais aussi dans la vie.

Merci à tous les membres du Laboratoire Bioinformatique et Génomique Structurales qui m'ont apporté aide et soutien pendant cette thèse. Grâce à leurs compétences et leurs qualités humaines, mon travail s'est déroulé dans une ambiance chaleureuse (et parfois bruyante !). Un merci tout particulier :

à Odile Lecompte, mon écrivain préféré, pour son amitié et son soutien constant dans les moments difficiles,
à Luc Moulinier avec qui les discussions sont toujours « stimulantes »,
à Raymond Ripp pour ses photos, aussi bien que sa gentillesse et sa disponibilité.

Merci aussi aux membres de la Plate-forme Bioinformatique de Strasbourg, et en particulier à Frédéric Plewniak pour son aide tout au long de ces années.

Merci à Serge Uge pour son assiduité et sa persévérance contre les « trials and tribulations » du système Linux.

Next, I would like to thank Toby Gibson for introducing me to the world of science and to computational biology (and also to German beer!). Thanks also to everybody in the Bioinformatics group at the European Molecular Biology Laboratory, and in particular to Sian Etherington and Simon Hubbard, for their friendship and support during my time there.

Thanks to Stephen, Linda, David and Wendy for always being there when I needed them and a big thanks to my parents for having given me the possibility to pursue my dreams…

Finally, a big thanks to Sammy and Amina for their understanding and encouragement during the past few years. Without them, this thesis would probably not exist.

Résumé

Les bases de la bioinformatique

Depuis la mise en évidence de l'ADN comme source première de l'information génétique et la détermination, en 1953, de la structure de la double hélice d'ADN, la bioinformatique est devenue une discipline à part entière dans la recherche et les développements des sciences du vivant. Initialement conçue autour de méthodes informatiques dédiées à l'organisation et à l'analyse des données déposées dans les premières bases de données biologiques, la bioinformatique s'est structurée, dans le courant des années 80, autour de différents champs d'application pour aboutir à une discipline de recherche indépendante. Schématiquement, trois branches majeures sont souvent distinguées correspondant aux aspects de stockage et de récupération des données, aux aspects de traitements et analyses statistiques et informatiques des données et enfin ceux couvrant le développement de nouveaux algorithmes de prédiction à même de fournir de nouvelles informations. Classiquement, les analyses bioinformatiques étaient réalisées par des experts qui validaient visuellement ou expérimentalement les résultats obtenus *in silico*. Cependant, à l'ère post-génomique, la bioinformatique est traversée par une véritable révolution liée à la disponibilité de nombreuses séquences de génomes complets coïncidant avec la production d'une vaste quantité de données liées à l'émergence des technologies à haut débit et recouvrant des domaines aussi variés que la transcriptomique, la protéomique ou l'interactomique. Dès lors, les bases de données biologiques sont littéralement inondées par un mélange hétéroclite d'informations validées expérimentalement ou prédites *in silico* avec leur corollaires d'approximation. Dans ce contexte, de nouveaux systèmes intégrés sont développés pour la gestion des données incluant les aspects de stockage et d'extraction efficaces de données hétérogènes jusqu'aux aspects de fouille de l'information et de mise en évidence des connaissances. Ces développements permettent d'envisager des études à haut débit de systèmes biologiques complexes et offrent comme perspective ultime la compréhension fine des processus et relations à l'œuvre dans le passage de l'information génétique vers les niveaux supérieurs de complexité tels ceux liés à la fonction moléculaire, aux grandes voies et réseaux biologiques, voire à la physiologie d'organisme entier ou aux systèmes écologiques.

Des séquences et structures tertiaires à la fonction

S'il est admis que l'information génétique présente dans le génome contient le schéma directeur pour le développement et la vie d'un organisme, il est clair que l'exploitation de cette information s'organise autour de différents niveaux de complexité tous fortement liés aux fonctions des produits des gènes (acides nucléiques ou protéines). Dès lors, une des applications les plus importantes de la bioinformatique a été l'étude des relations existant entre séquences d'acides nucléiques ou de protéines, structures tertiaires et fonctions biologiques. Si ces travaux ont révélé une relation directe entre similarité de séquences protéiques et conservation d'un même repliement structural, la relation entre repliement et fonction est apparu pour l'instant plus complexe. Ce résultat est sans doute à rattacher à la notion même de fonction d'un gène qui peut être décrite à différents niveaux allant de l'activité biochimique *stricto sensu* jusqu'à son rôle dans l'organe ou l'organisme en passant par les processus ou voies biologiques dans lesquels le gène est impliqué. Cependant, par delà ces différents plans de complexité, la comparaison des séquences d'acides nucléiques ou de protéines a été largement utilisée aussi bien pour révéler des motifs fonctionnels conservés que pour identifier des éléments distincts résultant d'événements ou de perturbations spécifiques.

En particulier, les comparaisons ou alignements multiples de séquences jouent un rôle fondamental dans la majorité des approches bioinformatiques mises en œuvre dans l'analyse de génome ou de protéome, et ce, depuis l'identification d'un gène jusqu'à la caractérisation des fonctions moléculaires et cellulaires du produit d'un gène. Initialement utilisés surtout dans des analyses liés à l'évolution et à l'exploration des relations phylogénétiques entre organismes, les approches de l'alignement multiple ont été mises à profit par les nouveaux algorithmes de recherche dans les banques afin d'améliorer le traitement de séquences de plus en plus distantes. Enfin, les alignements multiples ont grandement contribué à l'amélioration des prédictions de fonctions ou de structures tertiaires en s'appuyant aussi bien sur la mise en évidence d'homologie entre séquences qu'en réalisant des prédictions *ab initio* basées sur un consensus.

Vers une exploitation des données efficace et la découverte de connaissances.

Durant ma thèse, différentes approches complémentaires ont été développées dans la continuité de travaux antérieurs concernant la création et l'analyse des alignements multiples de séquences complètes (MACS). Trois nouveaux axes de recherche ont été particulièrement explorés qui ont abouti à la réalisation : (i) d'un nouveau banc d'essai pour l'évaluation objective des algorithmes d'alignement multiple, (ii) d'une ontologie spécifique aux alignements multiples de séquences (ADN/ARN/protéines) et de structures, (iii) d'un système de gestion d'information dédié à l'intégration et à l'analyse de l'ensemble des données attachées à la notion de famille de protéines.

(i) Evaluation objective des algorithmes d'alignement multiple

Actuellement, les méthodes d'alignement multiple évoluent rapidement pour répondre aux nombreux défis soulevés par les données du haut débit. Dans ce contexte, par delà les aspects purement informatiques qui deviennent prépondérants au grée de la croissance exponentielle des banques, l'estimation objective de la fiabilité d'un alignement est probablement le critère le plus important dans ces développements. En informatique, la qualité d'un algorithme est souvent estimée en comparant les résultats obtenus à un ceux d'un jeu de référence pré-calculé utilisé comme étalon. Dans le cadre des alignements de séquences, une référence objective peut être construite en combinant les informations des structures tridimensionnelles à celles des motifs fonctionnels. Cette approche a été utilisée pour la construction de BAliBASE, l'un des jeux d'essais les plus utilisés dans le domaine des méthodes d'alignement multiple. Dans ce cadre, les premiers travaux ont porté sur le développement d'un nouveau protocole semi-automatique et sur l'obtention d'une nouvelle version de la banque BAliBASE réunissant de larges séries d'alignements multiple de référence basés sur la superposition des structures 3D tout en maintenant un haut niveau de qualité et une validation humaine des cas trop complexes. Les alignements sont répartis dans différentes classes de référence correspondant aux problèmes les plus fréquemment rencontrés dans le domaine de l'analyse automatique des données du haut débit. Cela recouvre des problèmes liés à l'identification et à l'alignement de domaines isolés, étape essentielle à la création automatisée de banques de domaines jusqu'à l'alignement de séquences multi-domaines complètes fréquemment rencontrées dans les recherches dans les banques de séquences.

(ii) Ontologie dédié à l'alignement multiple

La seconde partie du travail a concerné le développement de MAO, acronyme de « Multiple Alignment Ontology », une ontologie 'orientée tâche' dédiée aux alignements d'acides nucléiques, de protéines ou de structures. Récemment, de nombreuses ontologies ont été développées afin d'obtenir une organisation plus efficiente des connaissances biologiques. Classiquement, une ontologie fournit une représentation structurée des connaissances courantes d'un domaine particulier sous la forme d'un 'vocabulaire de termes' et de 'spécification de leur sens' comprenant des définitions formelles et connectées. Un tel formalisme fournit une trame propice aux traitements informatiques et algorithmiques aboutissant ainsi à la détection de motifs cachés au sein des données et à l'extraction aisée des connaissances. MAO a été développée en collaboration avec des experts provenant des deux communautés (acides nucléiques et protéines), et impliqués dans les domaines de la comparaison des séquences et des structures secondaires et tertiaires. Un des éléments les plus puissants de MAO est lié au fait qu'elle fournit un lien naturel et intuitif entre de nombreuses ontologies distinctes déjà développées dans les domaines de la génomique et de la protéomique de telle sorte que les données expérimentales et les informations prédites puissent être intégrées et estimées dans le contexte de leur conservation au sein d'une famille de séquences alignées.

(iii) système de gestion d'information d'alignement multiple

MAO a été mis à profit dans un nouveau système de gestion d'information, appelé MACSIMS (acronyme de « Multiple Alignment of Complete Sequence Information Management System »), utilisé pour l'intégration et l'organisation automatiques de différents types de données dans le cadre de l'alignement multiple. Une combinaison de méthodes exploitant l'analyse des bases de connaissances et la prédiction de séquences *ab initio* est utilisée pour réaliser des validations croisées s'appuyant sur les informations structurales et fonctionnelles issues des banques publiques de séquences. L'information validée des séquences connues est alors propagée aux séquences inconnues, les caractérisant ainsi par des annotations fiables et détaillées. Les informations collectées ou générées par MACSIMS sont disponibles dans un format structuré permettant une exploitation automatique à haut débit par ordinateur et sont aussi accessibles au biologiste pour l'analyse visuelle à travers une interface web simple et conviviale. MACSIMS facilite ainsi la collecte automatique d'informations et d'extraction de connaissances et fournit un outil interactif d'interrogation et de visualisation des résultats.

La puissance intégrative de MACSIMS a été exploitée dans une variété de projets distincts, incluant (i) les validations *in silico* de séquence de protéines (Bianchetti, 2005), (ii) l'annotation fonctionnelle de protéines basée sur 'Gene Ontology' (Chalmel *et al*., 2005), (iii) la caractérisation de cibles potentielles pour le projet SPINE (Structural Proteomics IN Europe) (Thompson *et al.*, 2006; http://www.spineurope.org/) et (iv) la prédiction des effets structuraux et fonctionnels de mutations génétiques humaines dans le contexte du projet MS2PH (de la Mutation Structurale aux Phénotypes des Pathologies Humaines) (Garnier *et al*., 2006).

Nous avons aussi démontré que MACSIMS, en combinaison avec la base de données BAliBASE, peut évoluer vers un véritable 'banc d'essai' capable de tester et de valider l'adéquation entre une information liée aux séquences et une question biologique spécifique. Cette approche a été validée dans le cadre d'une étude portant sur l'efficacité de prédiction des sites fonctionnels dans les protéines sur la base de différentes caractéristiques de séquence/structure/évolution. Les méthodes actuelles utilisent pour l'essentiel, la

conservation évolutive comme l'indicateur primaire de sites potentiels. Cependant, cette conservation ne reflète pas seulement la pression de sélection impliquée dans le maintien de la fonction de la protéine, mais aussi celle responsable de la stabilité du repliement tridimensionnel. Nous avons ainsi démontré qu'en combinant les résidus conservés dans les alignements multiples de séquences, avec les renseignements d'hydrophobicité, d'accessibilité à la surface et de contacts entres résidus, nous pouvons améliorer l'exactitude des prédictions de sites fonctionnels.

Conclusions et perspectives

Les travaux décrits constituent les premières étapes d'une évolution de l'alignement multiple traditionnelle permettant de passer d'un simple empilement de lettres à l'obtention d'un dispositif interactif intégrant non seulement les séquences, mais également les informations structurales et fonctionnelles ainsi que des données prédites. Dans le futur, MAO sera amélioré par l'incorporation d'autres informations, telles que celles ayant trait à la structure des gènes, aux mutations et leurs phénotypes associés ou aux résidus impliqués dans des interactions. Ces informations couplées à des stratégies d'analyses appropriées seront intégrées dans les futures versions de MACSIMS et fourniront les bases pour le développement de nouveaux algorithmes de création d'alignements multiples incorporant les connaissances disponibles ainsi qu'au développement d'une nouvelle fonction objective d'évaluation de la qualité des MACS.

Les applications potentielles de MACSIMS sont très nombreuses et touchent aussi bien aux aspects d'annotation automatique de protéines hypothétiques, dont le nombre ne cesse de grandir suite aux multiples projets de séquençage de génomes complets, qu'à des aspects plus structuraux tel que l'étude de motifs ou résidus spécifiques d'un repliement. A l'avenir, on peut penser que ces développements auront des implications dans les domaines aussi divers que le génie des protéines, la modélisation de voies biologiques, les études génétiques de la susceptibilité aux maladies humaines ou les stratégies de développements de médicaments.

Un autre domaine de recherche en plein croissance concerne l'utilisation des méthodes d'alignement multiple pour des données autres que des acides nucléiques ou aminés, et notamment, pour des 'alphabets structuraux' constitués de lettres correspondant à des fragments de structures tertiaires ou pour des 'alphabets événementiels' développés dans le cadre des sciences sociales afin de caractériser des successions temporelles d'événements ou d'activités. Ces axes de recherche sont assez récents et envisagent de tirer profit des stratégies et méthodologies d'alignement multiple développées dans le passé dans le contexte de la comparaison de séquences moléculaires. Cependant, il est clair que, dans le futur, ces nouveaux champs d'investigation auront des retombées particulièrement bénéfiques en contribuant à l'émergence de nouveaux concepts et à de nouvelles formulations de la problématique de l'alignement multiple en général.

Contents

List of Figures

List of Tables

"To those who would know the biochemical structure, function and origin of man and would strive to improve his lot."
MO Dayhoff, 1965

1 General introduction

The work described in this thesis concerns the study of biological sequences and the role of the encoded gene products at the molecular, cellular and organism levels. We will use multiple sequence alignments of complete sequences (MACS) to place the gene sequence in the context of its overall family, where patterns of conservation and divergence can be used to identify evolutionarily conserved features and important genetic events. The MACS will also be used as a tool for the integration of all the information related to a gene family, providing an ideal workbench for the study of the relationships between gene sequence, structure, function and evolution.

The foundations of bioinformatics

Since the discovery of DNA as the source of genetic information and the elucidation of the double-helical nature of the DNA molecule (Watson and Crick, 1953), bioinformatics has become an integral part of research and development in the biological sciences. Originally introduced for the analysis of biological sequences (e.g. Fitch 1966; Needleman and Wunsch 1970; Sankoff 1975; Smith and Waterman 1981), new computational methods were soon needed to organise and analyse the data stored in the first biological databases (e.g. Dayhoff, 1965; Bernstein *et al*., 1977; Bairoch and Boeckmann, 1991) and, in the 80's, the field of bioinformatics took shape as an independent research discipline. For the first time, efficient algorithms were developed to cope with an increasing volume of information, and their computer implementations were made available for the wider scientific community. Three major problems were addressed at this time: the storage and retrieval of the data, computational and statistical data analyses and algorithms for prediction. In general, bioinformatics studies were performed by experts who manually verified the results obtained.

Bioinformatics in the post-genomic era

Bioinformatics has now been transformed by the availability of numerous complete genome sequences, as well as the new information resources that are being created from the raw data produced by different high throughput technologies in fields such as transcriptomics, proteomics, or interactomics. As a consequence, the biological databases are now being inundated with a mixture of experimentally validated data and computational predictions with their inherent unreliability. In this context, information management systems are now being introduced to collect, store and curate all this heterogeneous information in ways that will allow its efficient retrieval and exploitation. These developments are opening up the possibility of new large scale studies, whose goal is to understand how genetic information is translated to molecular function, networks and pathways, all the way to physiology and even ecological systems. The success of these studies will depend on our ability to organise and validate the raw data, to extract previously unknown information, to infer new hypotheses and to present the results in a user-friendly way to the biologist.

From sequences and 3D structures to function and evolution

The genetic information encoded in the genome sequence contains the blueprint for the potential development and activity of an organism, but the implementation of this information depends on the functions of the gene products (nucleic acids and proteins). For example, RNA plays a key role in all steps of gene expression, as an intermediate carrier of genetic information and as a functional intermediate of the expression cascade which amplifies single genes into many copies of the encoded proteins. Many non-coding RNA (tRNA, snRNA, miRNA, etc.) are also involved in direct regulation of transcription and translation. Proteins perform a wide variety of biological functions in organisms, from catalysis of biochemical reactions, transport of nutrients or recognition and transmission of signals to structural and mechanical roles within the cell. As a consequence, one of the most important applications of bioinformatics has been the study of the relationships between the sequence of a gene and its 3D structure, biological function and evolution. The function of an RNA molecule depends mostly on its tertiary structure and this structure is generally more conserved than the primary sequence (Woese and Pace, 1993). In the case of proteins, a direct relationship between sequence similarity and conservation of 3D structure has been clearly established (Koehl and Levitt, 2002a). However, the relation between fold and function is much more complex (Watson *et al.,* 2005). Gene function can be described at many levels, ranging from biochemical function, via macromolecular complexes to cellular processes and pathways, up to the organ or organism level. Furthermore, as proteins and RNA evolve, they can acquire new roles and sequences with a common evolutionary origin do not necessarily share the same precise function. This complexity calls for a more rigorous description of molecular and cellular functions and several projects have been initiated recently to formally characterise functional information, such as the RNA Ontology (Leontis *et al.*, 2006), the Gene Ontology (Ashburner *et al.*, 2000) or the EC (Bairoch 2000) and Kegg (Kanehisa, 2002) databases.

The central role of multiple alignments

The comparison of protein or nucleic acid sequences has had a major impact on our understanding of sequence/structure/function/evolution relationships (Lecompte *et al.*, 2001). Multiple sequence comparisons or alignments were originally used in evolutionary analyses to explore the phylogenetic relationships between organisms (reviewed in Phylips *et al.*, 2000). More recently, new sequence database search methods have exploited multiple alignments to detect more and more distant homologues (e.g. Altschul *et al.*, 1997). Multiple sequence alignments of protein or nucleic acid sequences are also used to highlight conserved functional features and to identify major evolutionary events, such as duplications, recombinations or mutations. They have led to a significant improvement in predictions of both 3D fold (Moult 2005) and function (Watson *et al.*, 2005). Of course, in the current era of complete genome sequences, it is now possible to perform comparative multiple sequence analysis at the genome level (Margulies *et al.*, 2006).

Such studies have important implications in numerous fields in biology. Nucleic acid divergence is used as a molecular clock to study organism divergence under the evolutionary forces of natural selection, genetic drift, mutation and migration, with applications from the scientific classification or taxonomy of species to genetic fingerprinting. Conserved sequence features or markers are used to characterise groups of individuals in population genetics (Kidd *et al.*, 2004). Genotype/phenotype correlations can reveal candidate genes

associated with a particular trait (e.g. plant height) or inherited disease, such as schizophrenia (Owen *et al.*, 2005). In drug discovery, a protein family perspective can identify specific structural or functional features that facilitate protein-ligand interaction studies for high-throughput virtual compound screening methods (Lenz *et al.*, 2000). Thus, multiple alignments now play a fundamental role in most of the computational methods used in genomic or proteomic projects, ranging from gene identification and the functional characterisation of the gene products to genetics, human health and therapeutics. However, new bioinformatics approaches are now needed in order to manage and extract the important information from the mass of data generated by the new high-throughput technologies.

Objectives: towards efficient data exploitation and knowledge discovery

In this work I have built on past experience in the group concerning the construction and evaluation of high-quality, reliable Multiple Alignments of Complete Sequences (MACS) (Thompson *et al.*, 1997; Plewniak *et al.*, 2000; Thompson *et al.*, 2001; Thompson *et al.*, 2003; Plewniak *et al.*, 2003; Thompson *et al.*, 2004). The MACS provides an ideal environment for the reliable integration of information from a complete genome to a gene and its related products. In order to fully understand the functions and molecular interactions of a particular protein, such diverse information as cellular location, degradation and modification, 2D/3D structures, mutations and their associated illnesses, the evolutionary context and literature references must be assembled, classified and made available to the biologist. By placing the sequence in the framework of the overall family, multiple alignments can identify important structural or functional motifs that have been conserved through evolution, but can also highlight particular non-conserved features resulting from specific events or perturbations. Multiple alignments thus allow reliable data validation, consensus predictions and rational propagation of information from known to unknown sequences, and provide a valuable workbench for integrated systems analysis, hypothesis generation and experiment-planning advice.

The three major new developments described here are (i) a new benchmark for the objective evaluation of multiple alignment algorithms, (ii) an ontology for multiple alignments of DNA/RNA/protein sequences and structures, (iii) an information management system that exploits the MACS and the organisation provided by the ontology.

(i) Objective evaluation of multiple alignment algorithms

Multiple alignment methods are now evolving in response to the challenges posed by the new large-scale applications (reviewed in Thompson and Poch, 2006a). Alignment reliability is probably the most important criteria in these developments, though computational aspects will become more and more important as the databases increase in size. In computer science, the quality of an algorithm is often estimated by comparing the results obtained by a new method with a pre-defined benchmark or 'gold standard'. In the case of multiple alignments, an objective reference can be constructed by incorporating 3D structure information and functional motifs. This approach was used in the construction of BAliBASE (Thompson *et al.*, 1999a; Bahr *et al.*, 2001), one of the most widely used benchmarks for multiple sequence alignment methods. A comparison of different multiple alignment methods based on BAliBASE (Thompson *et al.*, 1999b) highlighted the fact that no single algorithm was capable of constructing high quality alignments for all test cases. Subsequently, the first alignment methods were introduced that combined both global and local information in a

single alignment program, resulting in more reliable alignments for a wide range of alignment problems.

Here, a new semi-automatic protocol has been developed in order to construct a new version of the BAliBASE database (Thompson *et al*., 2005a), which provides high quality, manually refined, reference alignments based on 3D structural superpositions. The alignments are organised into different reference sets, containing test cases that cover most of the current multiple alignment problems, from alignment of single domains e.g. in the construction of protein domain databases to the alignment of full-length, complex sequences, such as those detected by the database searches routinely performed in automatic, high throughput genome analysis projects. A comparison of the most recent alignment programs using BAliBASE version 3 has shown that significant improvements have been achieved since the last comprehensive evaluation was performed (Thompson *et al.*, 1999b). This is most probably due to the recent development of co-operative, knowledge-based methods that exploit the new structural and functional information available.

(ii) Multiple alignment ontology

The second part of this work concerned the development of the Multiple Alignment Ontology (MAO), a task-oriented ontology for nucleic acid and protein sequence and structure alignments (Thompson *et al.*, 2005b). In recent years, ontologies have been introduced in a number of areas for the management of biological knowledge. In computer science, an ontology is defined as a formal, structured representation of the knowledge in a particular domain (Gruber, 1993). It includes a "vocabulary of terms" and a "specification of their meaning" including definitions and inter-relations, which impose a structure on the domain and constrain the possible interpretations of terms. Such a formalism provides a framework for computational methods dedicated to the detection of hidden motifs and to knowledge discovery. MAO was designed to improve interoperation and data sharing between different alignment protocols for the construction of a high quality, reliable multiple alignment in order to facilitate knowledge extraction and the presentation of the most pertinent information to the biologist. The ontology has been developed in collaboration with domain experts from both the DNA/RNA and protein communities, including experts in the fields of both primary sequence and 2D/3D structure comparisons. One of the most powerful features of the MAO ontology is that it provides a natural, intuitive link between a number of different ontologies in the domains of genomics and proteomics, so that diverse experimental data and predicted information can be integrated in the context of the overall family alignment.

(iii) Information management system

MAO was then used in a new MACS-based Information Management System called MACSIMS (Thompson *et al.*, 2006b), for the integration of different types of data in the framework of the multiple alignment. MACSIMS was designed to combine knowledge-based methods with complementary *ab initio* sequence-based predictions for protein family analysis. A data collection system has been incorporated to automatically retrieve a wide range of information, from taxonomic data and functional descriptions to individual sequence features, such as structural domains and active site residues. A number of new algorithms have been developed for reliable data cross-validation, consensus predictions and rational propagation of information from the known to the unknown sequences. In this way, structural and functional data can be combined with information about the conservation of the family and the variability observed at different residue sites. All the information

collected or generated by MACSIMS is stored in XML format files that provide a structured format for automatic, high-throughput data parsing by computers. In addition, all the information is also easily accessible for manual analysis by biologists, via a simple-to-use, graphical interface implemented via the JalView alignment editor (Clamp *et al.*, 2004). MACSIMS thus facilitates automatic information retrieval and knowledge discovery, with functionalities for interactive queries and visual exploration of search results.

Example applications

The integrative power of MACSIMS has been used in a variety of different projects, including (i) the *in silico* validation of protein sequences (Bianchetti *et al.*, 2005), (ii) reliable protein function annotation based on the Gene Ontology (Chalmel et al., 2005), (iii) the characterisation of potential targets for the SPINE (Structural Proteomics in Europe) project (Albeck *et al.*, 2006; http://www.spineurope.org/) and (iv) the prediction of structural and functional effects of mutations in the context of the MS2PH (Structural Mutation to Human Pathologies Phenotype) project (Garnier *et al.*, 2006).

We have also shown that MACSIMS, combined with the BAliBASE benchmark database, represents an effective workbench for the testing of hypotheses related to the most pertinent information for a specific research question. The rationale is demonstrated by a study of the effectiveness of a number of sequence/structure characteristics for the prediction of functional sites in proteins. Most current methods use evolutionary conservation as the primary indicator of potential sites. However, sequence conservation reflects not only evolutionary selection to maintain protein function, but also to maintain the stability of the folded protein. We have demonstrated that by combining the conserved residues in multiple sequence alignments, with other information such as residue hydrophobicity, solvent accessibility and inter-residue contacts, we can improve the accuracy of such predictions.

Conclusion and perspectives

Genomics and proteomics technologies, together with the new systems biology strategies have led to a paradigm shift in bioinformatics. The traditional reductionist approach has been replaced by a more global, integrated view. Multiple sequence alignment is one example where a shift of thinking or focus is now leading to the development of new methods. The work presented here represents the first steps in the evolution of the traditional multiple sequence alignment from a simple stacking of letters to become an interactive tool, incorporating not only the sequence itself, but also structural/functional information in the context of the complete protein family. The MAO ontology has been developed in order to provide a data standard for the exchange of information and to facilitate collaborations between the different resources. The formalism and organisation established in MAO is exploited in a new information management system. In MACSIMS, knowledge-based techniques have been developed that include data mining components for finding subtle correlations and relationships and data management and analysis techniques have been applied to ensure that the pertinent information can be extracted and presented to the biologist in a clear, user-friendly format. In this context, version 3 of the BAliBASE benchmark represents a comprehensive reference for the evaluation and comparison of the new alignment methods that are now being introduced.

In the future, MAO will be extended to incorporate other information resources, such as gene structure, mutation and phenotype information and residue interaction data. This will

require more formal links between MAO and the other biological ontologies. The information retrieved from the numerous biological databases, together with complementary sequence analysis strategies will provide the basis for a new knowledge-based multiple alignment construction method, as well as a new objective function for the evaluation of the quality of the MACS.

The potential applications of MACSIMS are numerous, but will include such fields as the automatic annotation of the ever-increasing number of hypothetical proteins being produced by the high-throughput genome sequencing projects or the definition of characteristic motifs for specific protein folds. Hopefully, this will also have significant consequences for more wide-reaching areas, such as protein engineering, metabolic modelling, genetic studies of human disease susceptibility, and the development of new drug development strategies.

Another growing area of research is the application of multiple alignment methods for the comparison of other kinds of data, such as 3D structure fragment libraries or structural alphabets (e.g. Kolodny *et al.*, 2002; Camproux *et al.*, 2004), molecular networks (Sharan and Ideker, 2006), or even time use data and activity patterns in the social sciences (Thompson *et al.*, 1999c; Wilson, 2006). Here, data that is fundamentally a sequence of events is represented by an alphabet defined by experts in the field, who also define the similarity scores between the different events. These emerging fields are exploiting the power of the multiple alignment methodologies developed over the years for the comparison of molecular sequences, but will also contribute new concepts and formulations that will undoubtedly prove beneficial in the future.

"Is there a danger, in molecular biology, that the accumulation of data will get so far ahead of its assimilation into a conceptual framework that the data will eventually prove an encumbrance?"
John Maddox, Nature, 1988

2 Context: bioinformatics in the post-genomic era

Biology has been transformed by the availability of numerous complete genome sequences for a wide variety of organisms, ranging from bacteria and viruses to model plants and animals to humans. However, the completion of the genome sequence is just a milestone marking the beginning of efforts to decipher the meaning of the genetic "instruction book". The major challenge today is to understand how the genetic information encoded in the genome sequence is translated into the complex processes involved in the organism and the effects of environmental factors on these processes. Bioinformatics will play a crucial role in the systematic interpretation of genome information, associated with data from other high-throughput experimental techniques, such as structural genomics, proteomics or transcriptomics. The results of such large-scale analyses will have widespread implications for fundamental research, but will also have practical biotechnology applications in fields such as genetic fingerprinting and engineering, human health, diagnostics and therapeutics.

Section 2.1 describes the current data explosion in the field of molecular biology, brought about by the recent developments in high-throughput genomic, transcriptomic and proteomic technologies. The implications of this changing data landscape for systems-level studies are then discussed in section 2.2. Finally section 2.3 focuses in more detail on the study of the relationships between sequence and structure, function and evolution and the critical role such studies play in the new systems-level biology.

2.1 From a data-poor to a data-rich science

2.1.1 Genome sequencing

In the past ten years, high throughput genome sequencing and assembly techniques have lead to a rapid increase in the amount of sequence data available in the public databases. The first free-living organism to be sequenced was that of *Haemophilus influenzae* (1.8Mb) in 1995, and since then genomes have been sequenced at an ever-increasing pace. The human genome was completed by the Human Genome Project in 2004, and high-quality draft genome sequences are now available for many higher organisms, including the mouse, the domestic dog and the chimpanzee. At the time of writing, the Genomes OnLine Database (GOLD: http://www.genomesonline.org/) contained 364 complete genomes or chromosomes, and a further 607 eukaryotic genomes, 950 bacterial genomes and 58 archaeal genomes were being sequenced. Many of these model organisms occupy strategic positions in the tree of life and provide important information for evolutionary studies (Delsuc *et al.*, 2005). Other genomes have been sequenced because they have important industrial implications, such as the design of novel drugs (Regnstrom and Burgess, 2005), the development of therapies for the treatment of complex diseases (e.g. Cooke, 2006), or the design and production of new fuels, food ingredients or thermostable chemicals (e.g. Niehaus *et al.*, 1999).

The various genome sequencing projects have lead to a rapid increase in the number of sequences available in the nucleotide and protein databases. For example, figure 2.1 shows the exponential growth of the UniProt protein database (Wu *et al.*, 2006) since 1996 (http://www.expasy.org/). The UniProt database consists of two sections: (i) the TrEMBL section contains translations of all coding regions from the major nucleotide databases and includes over 2.5 million sequences (ii) Swiss-Prot is a smaller section with manually-annotated records with information extracted from literature and curator-evaluated computational analyses.

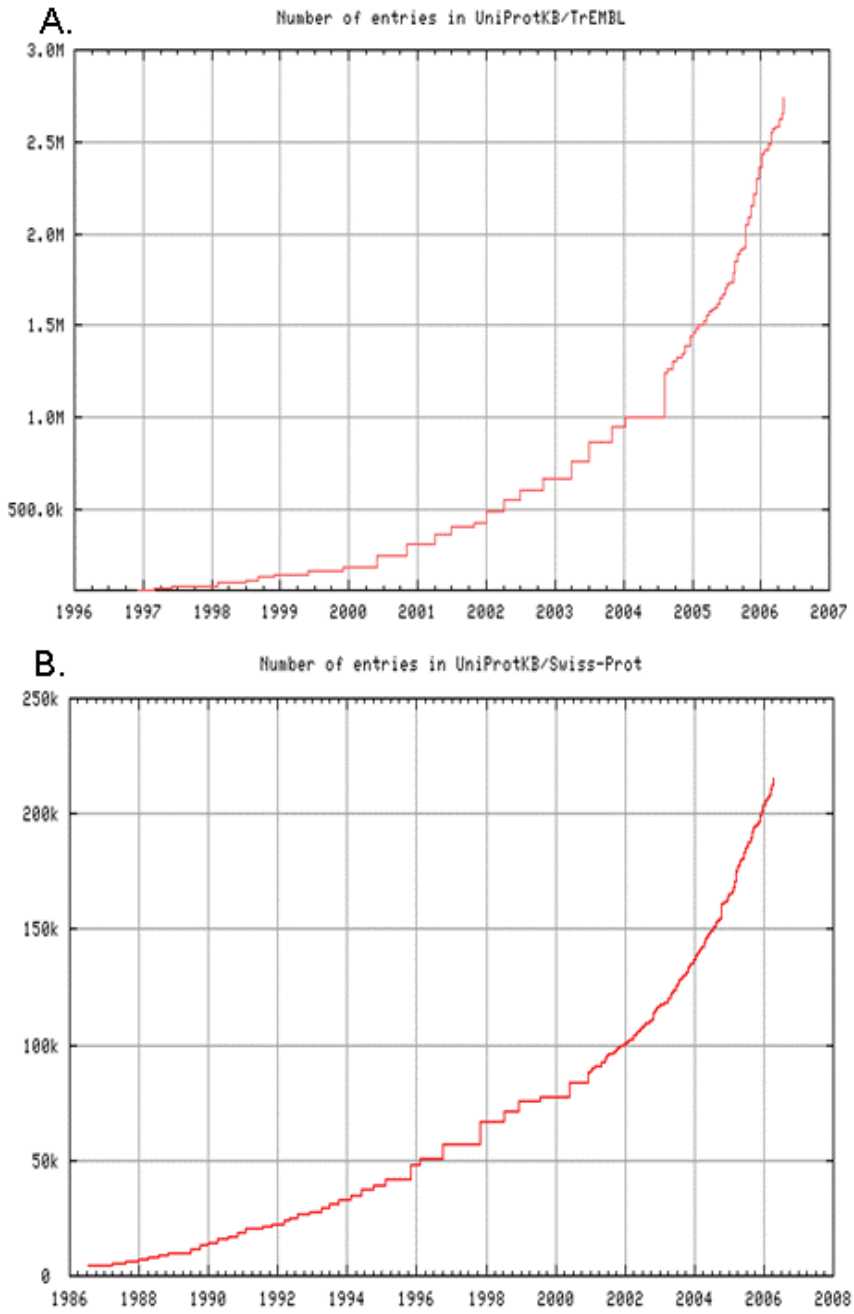


Figure 2.1 Exponential growth of TrEMBL and Swissprot sections of the Uniprot database

2.1.2 Structural genomics

This avalanche of protein sequences resulting from the determination of the complete genome sequences of diverse organisms, is now awaiting further structural and functional interpretation. Only a small fraction of the proteins encoded in the genomes has been experimentally studied, and uncharacterised proteins often represent more than half of the potential protein-coding regions of a genome (Roberts, 2004). To address this problem, large-scale experimental studies are now underway to gain a better understanding of the role and origins of the tens of thousands of these 'hypothetical' or uncharacterised proteins. For example, the goal of many of the structural genomics (SG) projects is to provide experimental 3D structures that cover the majority of the protein folding space. The structures of over 3000 proteins (see figure 2.2) have already been determined by such initiatives and deposited in the PDB database (Kouranov *et al.*, 2006)

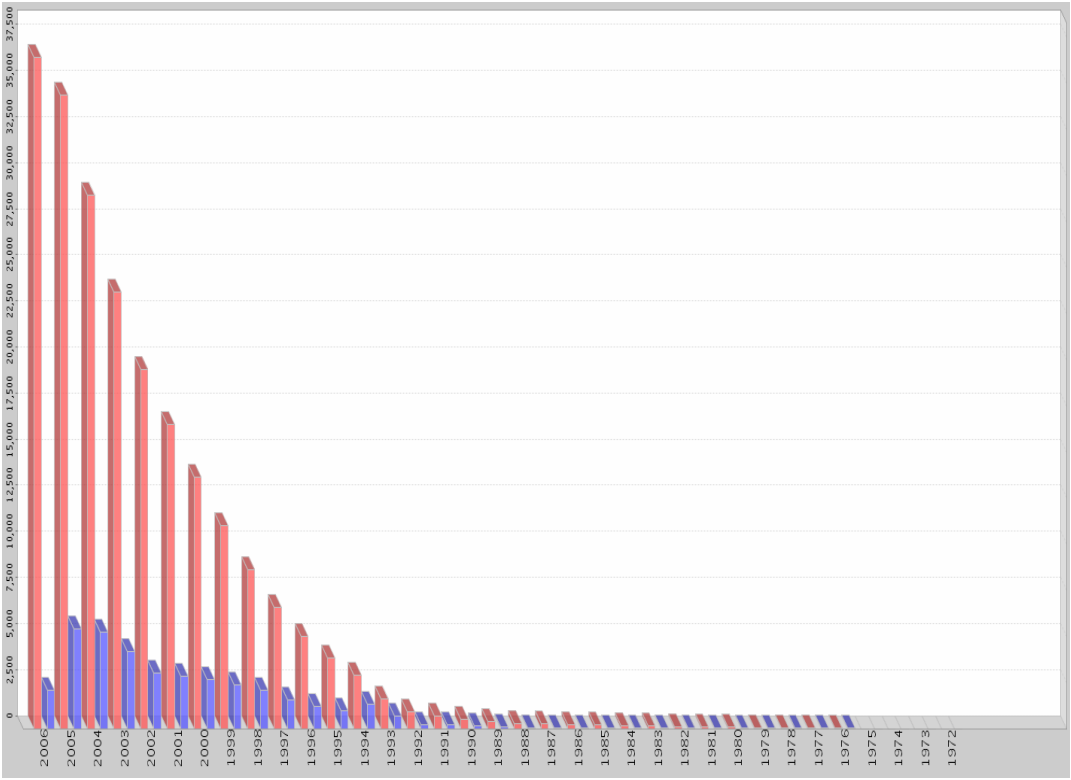


Figure 2.2 The number of solved structures in the PDB database
(http://www.rcsb.org/pdb/) The blue columns represent the yearly totals and the red columns represent the accumulated totals.

These experimental structures will have an even greater impact on our knowledge of the protein fold space because they can serve as representative templates for *in silico* structure homology modelling methods, providing valuable information for the large fraction of sequences whose structures have not been determined experimentally. Such comparative protein structure models are available in the MODBASE (Pieper *et al.*, 2006) database for all available protein sequences that can be matched to at least one known protein structure. Reliable structural models are available for domains in approximately 60% of the sequences in the Swiss-Prot and TrEMBL databases.

2.1.3 Other 'omics' resources

Other large-scale data resources are also emerging from high-throughput experimental technologies such as microarrays for systematically analysing gene expression profiles or yeast two-hybrid systems and mass spectroscopy for detecting protein-protein interactions. The impact of the genome projects of the past 10 years is thus not simply an increased amount of sequence data, but the diversification of molecular biology data. Table 2.1 lists some examples of these data resources, which have been denoted with the suffix '-ome' (from the Greek for 'all', 'every' or 'complete') to indicate studies undertaken on a large or genome-wide scale.

Transcriptome	the mRNA complement of an entire organism, tissue type, or cell
Proteome	the entire complement of proteins in a given biological organism or system at a given time
Metabolome	the population of metabolites in an organism
Secretome	the population of gene products that are secreted from the cell
Lipidome	the totality of lipids in an organism
Interactome	the complete list of interactions between all macromlecules in a cell
Spliceosome	the totality of the alternative splicing protein isoforms
Kinome	The totality of protein kinases in a cell
Neurome	The complete neural makeup of an organism
ORFeome	the totality of protein-encoding open reading frames (ORFs)
Unknome	The totality of genes of unknown function
Textome	The body of scientific literature which text mining can analyse
Resourceome	The full set of bioinformatics resources

Table 2.1 Some examples of the use of generic '-ome' terminology

The availability of comprehensive 'omics' datasets is changing the way we approach biological research. The emphasis in biology and bioinformatics is shifting from studying individual components, such as genes, RNA or proteins in isolation, to the study of the vast networks that biological molecules create, which regulate and control life.

2.2 Systems biology

System-level studies are aimed at the elucidation, design or modification of complex structures, such as macromolecular complexes, regulatory pathways, cells, tissues or even complete organisms. Systems biology aims to explain such complex biological systems by using a combination of experimental, theoretical and computational approaches. The goal is not simply to produce a catalogue of the individual components or even interactions, but to understand how the system components fit together, the effect of each individual part on its neighbours, and how various parameters such as concentrations, interactions, and mechanics change over time (Ideker *et al.*, 2001a; Kitano, 2002: Ge *et al.*, 2003). The new outlook is characterised by the basic idea of "emergent" properties, *i.e.* it considers global behaviour not explicable in terms of the individual, single components of the system (Chong and Ray, 2002).

Historically, there have been two main approaches to systems biology. One has its roots in biology, the other in mathematics. The former sees it as a way to integrate the different levels of information, from system-level experiments in developmental biology, cancer studies etc. to

lower level data pertaining to genes, mRNAs, proteins, and pathways (Hood *et al.*, 2001). For the latter, the main idea is that complex biological systems might be modelled successfully using the tools developed in systems theory and engineering.

2.2.1 Heterogeneous data integration

Initial studies in this field have begun to provide insights into how cellular networks may be organised (Uetz and Finley, 2005). For example, new biological pathways have been elucidated from analysis of protein-protein interaction networks derived from experimental methods such as the yeast two-hybrid approach, mass spectrometry or TAP-tag technology, in conjunction with other types of data, including protein function, subcellular location and gene expression profiles (Cho *et al.*, 2004). In addition to the definition of the system structure, system behaviour and response to external stimuli, such as mutations, environmental conditions, chemical injection or drug absorption are also being studied. For example, protein–protein interaction maps for the budding yeast *Saccharomyces cerevisiae* were combined with a genomic-scale data set describing the phenotypic role of all nonessential yeast proteins to study the recovery of the yeast from exposure to DNA-damaging agents (Said *et al.*, 2004). A systematic integration of technologies is also being applied in the pharmaceutical industry to identify molecular functions and pathways associated with a disease and to improve the drug discovery pipeline (e.g. Davidov *et al.*, 2003; Apic *et al.*, 2005).

2.2.2 Mathematical modelling

Most biological systems are sufficiently complex (with nonlinearities, emergent properties, loosely coupled modules, etc. that are the hallmarks of complexity) that 'systems biology' and 'complex systems' might be considered to be practically synonymous. Tools found useful in analysing the latter should prove of value to the study of the former. Mathematical modelling was traditionally used in biology to study population genetics. More recently, thanks to the recent development of mathematical tools such as chaos theory to help understand complex, nonlinear mechanisms in biology and an increase in computing power which enables calculations and simulations to be performed that were not previously possible, modelling of various pathways and networks in cellular and molecular biology has become an area of growing interest. Dynamic simulations of metabolic networks were performed in the 1970s (Garfinkel *et al.*, 1970; Wright, 1970; Loomis and Thomas, 1976). These pathway-centered kinetic models were followed by cell-scale models of metabolic networks (Heinrich and Rapoport, 1977), and by the 1990s, multi-level models were formulated of dynamic, large-scale systems, e.g. mitosis (Novak and Tyson, 1995) or control of MAPK signalling in oncogenesis (Schoeberl *et al.*, 2002) and even of complete organisms, such as *Haemophilus influenzae* (Edwards and Palsson, 1999). These models describe reconstructed networks and their possible functional states (phenotypes) and are now available at the genome-scale for a growing number of organisms.

Biological systems are inherently complex, *i.e.* they are systems consisting of interacting parts, in which the state of one part affects the state of one or more others. Common interactions include feedback loops, in which information from the output of a system transformation is sent back as input to the system. Negative feedback loops, in which the new data produce an output in the opposite direction to previous outputs, are typically responsible for regulation and are generally considered to provide stability. Positive feedback loops, in which the new input facilitates and accelerates the transformation in the same direction as the

preceding output, are often equated with undesired instability in a system. Nevertheless, they are an important factor in the dynamics of many complex systems.

In order to understand a system, a so-called "structural model" (this terminology is nothing to do with 'structural biology' in the sense of the 3D coordinates of atoms in a molecule) is needed, which shows all the components and, qualitatively, how they interact with each other. In order to understand in quantitative terms how these links behave, the model is based on equations and an associated set of parameters. Some of the mathematical formalisms that have been employed with some success in biology and bioinformatics to describe such complex systems, are directed graphs, Bayesian networks, Boolean networks and their generalizations, ordinary and partial differential equations, qualitative differential equations, stochastic equations, and rule-based formalisms (reviewed in de Jong, 2002). Using formal methods, the structure of systems can be described unambiguously, while predictions of their behaviour can be made in a systematic way. Modelling and simulation studies have predominantly used deterministic, coarse- to average-grained models, such as logical models and simple differential equation models. The few modelling and simulation studies using fine-grained, quantitative, and stochastic models have been restricted to regulatory networks of relatively small size and modest complexity that have been well-characterized already by experimental means.

2.2.3 Combined approaches

The two approaches described above are complementary and recently they have been used together in an iterative way (see figure 2.3).

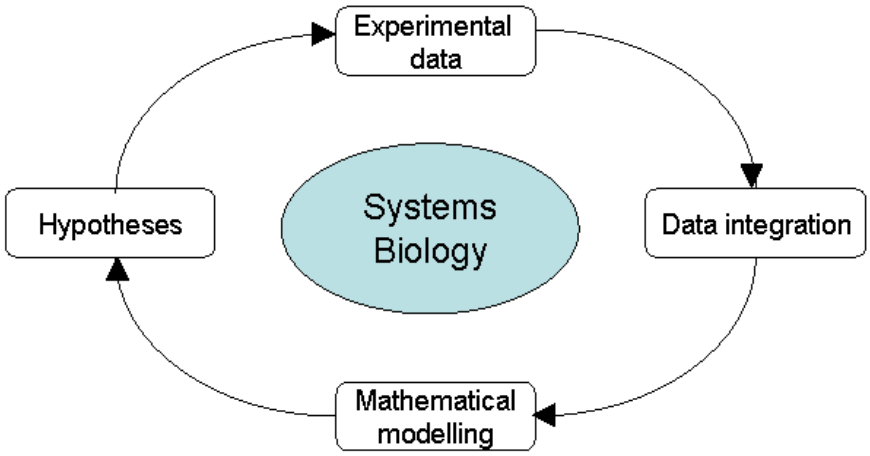


Figure 2.3 Overview of the new integrated approach to systems biology

An integrated systems approach to understanding biology can be described as an iterative process that includes (1) data collection and integration of all available information (ideally all components and their relationships in the organism), (2) mathematical modelling of the system, (3) generation of new hypotheses and (4) experimentation at a global level. In this new approach, global sets of biological data are integrated from as many hierarchical levels of information as possible. This is the initiation point for the formulation of detailed graphical or mathematical models, which are then refined by hypothesis-driven, iterative systems perturbations and new data integration. Cycles of iteration will result in a formal working model of how the systems function dynamically in the growth, development and maintenance of the organism in the context of its environment. Ultimately, these models will explain the

systems or emergent properties of the biological system of interest. Once the model is sufficiently accurate and detailed, it will allow biologists to accomplish two tasks not possible before: (i) to predict the behaviour of the system given any perturbation and (ii) redesign or perturb the gene regulatory networks to create completely new emergent systems properties.

Quantitative experimental data has been combined with mathematical modelling to predict the ability of certain bacteria to orient themselves to specific chemicals (Barkai and Leibler, 1997). In another data-intensive study (Ideker *et al.*, 2001b), the effect of perturbations of the sugar metabolism pathway in yeast on gene expression and protein activity was studied, using an integrated approach to build, test, and refine a model based on experimental data from DNA microarrays, quantitative proteomics, and databases of known physical interactions. The combined approach has also been used to model complete cells, for example, in the Virtual Cell or Vcell project (Schaff and Loew, 1999) or the e-cell project (Tomita, 2001). Both Vcell and e-cell use differential equations to model the basic chemical pathways in a cell and to predict what would happen if a pathway were to be interfered with or changed in some way. However, modelling a single cell gives little insight into multi-cellular processes and diseases, and single cell simulations are now being complemented by *in silico* multi-cellular systems.

Examples of multi-cellular models include genome-wide strategies to understand the control systems of specific components of development that are highly conserved, for example segment polarity genes of fruitfly (von Dassow *et al.*, 2000) and endomesoderm specification in the sea urchin embryo (Davidson *et al.*, 2002). More recently, an important international collaborative effort has been made in the context of the IUPS Physiome Project (Hunter, 2004), in order to provide a framework for linking models of biological structure and function in human and other eukaryotic physiology across multiple levels of spatial organization (from nano-scale molecular events to metre-scale intact organ systems) and multiple time scales (from microseconds to a human lifetime). The levels of biological organization, from genes to the whole organism are: gene regulatory networks, protein–protein and protein–ligand interactions, protein pathways, integrative cell function, tissue and whole organ structure/function relations, and finally the integrative function of the whole organism. Models of heart, lungs, musculo-skeletal system and the digestive system have already been constructed within the IUPS project, with applications ranging from educational software to virtual surgery and surgical training, medical diagnostics and drug discovery.

Although systems biology is still a young field, the pioneering studies described above have clearly established systems biology as a firm scientific discipline. Nevertheless, the field still faces significant experimental, technical and computational challenges that will need to be addressed in the future. Once these have been resolved, a detailed understanding of the interplay between different hierarchies of biological information within their environmental contexts will hopefully lead to new conceptual insights, as well as more practical innovations, such as predictive; preventive and personalised medicine or alternative sources of food and energy

2.3 Systems-level functional studies

The study of the relationships between sequence and structure, function and evolution will play a critical role in the new systems level studies. Although physiologists and ecologists have been using a systems approach for many years, it is only in the post-genomic era that it has become feasible to extend these studies to the level of molecular details. For the first time, it

will be possible to integrate knowledge across different levels of biological organization and to anchor this at the molecular level. For these applications to be realistic, though, apart from vastly increased computing power, it will be absolutely necessary to come to a fundamental understanding of the processes in cells at the smallest level. The basis for macro-level insights is still micro-level knowledge. Unfortunately, although genome sequence information has become available in unprecedented amounts, the absence of a direct functional correlation between genes and the corresponding nucleic acid or protein products represents a significant roadblock for improving the efficiency of biological discoveries (Bienkowska, 2005).

2.3.1 From DNA to RNA and proteins

The Central Dogma of Molecular Biology (Crick, 1958) states that the genetic information encoded in DNA is transcribed to generate RNA and that these RNA are then translated to form proteins which perform cellular functions (see figure 2.4). The RNA intermediary between DNA and protein is a messenger type of RNA, or mRNA. The Central Dogma stipulates that no genetic information is transferred from protein to protein, protein to RNA or protein to DNA.

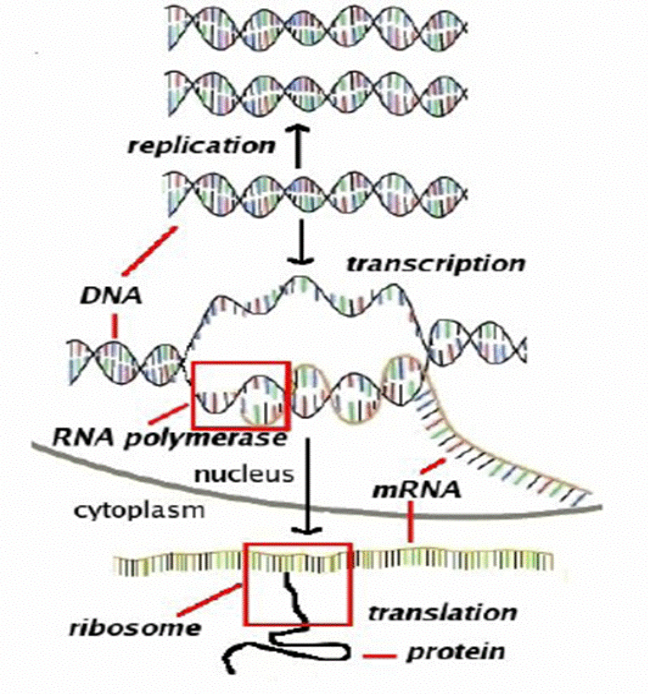


Figure 2.4 The Central Dogma of Molecular Biology

Over the last 50 years, many discoveries have challenged the Central Dogma and the fixed, deterministic view of DNA. They have revealed that, whilst the essential elements of the Central Dogma still hold, it is a rather over-simplistic model. Some examples of alternative information pathways that have been observed recently include:

- *Reverse transcription*. After the central dogma was expounded, retroviruses were discovered. These transcribe RNA into DNA through the use of a special enzyme called reverse transcriptase (reviewed in de Parseval and Heidmann, 2005). This confirmed that the flow from RNA to DNA does occur. Initially, this was thought to occur only in viruses, but, more recently, RNA to DNA flow has also been shown in higher animals, including

humans. For example, retrotransposons can copy themselves to RNA and then, via reverse transcriptase, paste multiple copies back to DNA (Mourier, 2005).

- *Viruses with RNA-only genomes*. Some virus species have their entire genome encoded in the form of RNA (Ahlquist, 2006). Thus, their information flow consists only of RNA to Protein.
- *Non-coding RNAs*. Many RNAs in an organism achieve a functional state capable of affecting the phenotype of the organism without ever being translated into a protein (Costa, 2005). Thus, their information flow consists only of DNA to RNA.
- *Prions*. Prions are proteins that propagate themselves by making conformational changes in other molecules of the same type (Bussard, 2005). This change affects the behaviour of the protein. In fungi, this change can be passed from one generation to the next, *i.e*. Protein to Protein.

Thus, the Central Dogma of Molecular Biology inspired by classical work in prokaryotic organisms accounts for only part of the genetic agenda of complex eukaryotes. In fact, gene expression is subject to a regulatory network of a complexity that it only just being realised (Lee *et al.*, 2002a). But, translation of the DNA genes to RNA or protein sequences is only the first step in the synthesis of functionally active molecules and further processing is required to obtain the final 3D structure and biochemical function of the gene product.

2.3.2 RNA sequence, structure and function

RNAs are single stranded polynucleotide molecules that often fold on themselves by base pairing to form structures called hairpin loops. Thus, most RNA molecules adopt specific tertiary structures. An example is shown in figure 2.5.

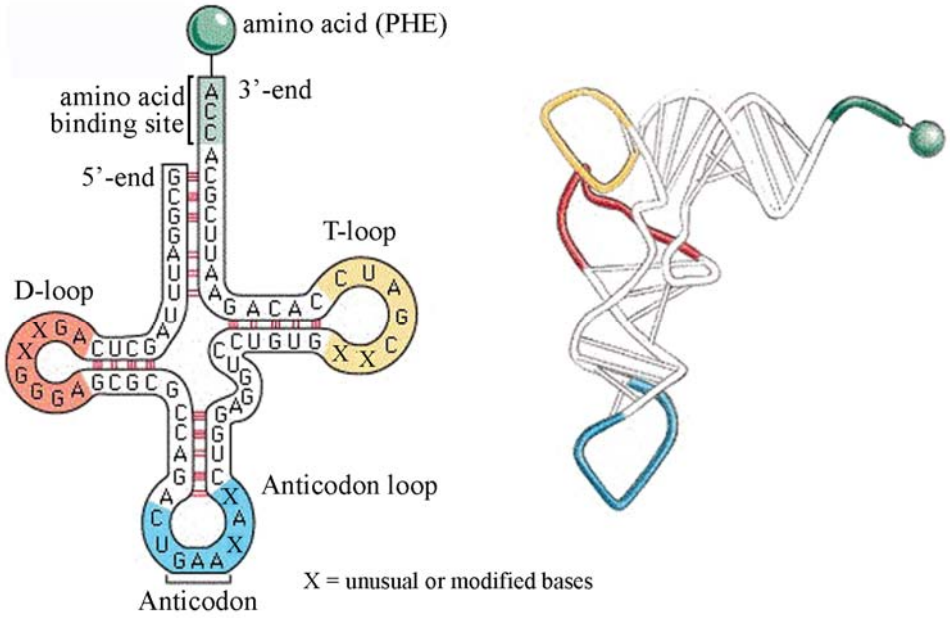


Figure 2.5 Different levels of RNA structure
The example shown is a tRNA molecule, which is a polynucleotide of about 60-95 nucleotides. tRNA exhibit a cloverleaf-like structure consisting of a stem and three main loops. The tertiary L-shaped structure interacts with ribosomes, aaRS, EFTU, etc.

.

Structural studies and comparative sequence analyses have suggested that biological RNAs are largely modular in nature, composed primarily of conserved structural building blocks or

motifs (Leontis and Westhof, 2003) of secondary (helices, and internal, external or junction loops) and tertiary (coaxial stacks, kissing hairpin loops, ribose zippers, etc.) structure. The secondary structure elements are significantly more stable and form faster than the tertiary interactions. Tertiary structure in RNA occurs via interactions involving two helices, two unpaired regions, or one unpaired region and a double-stranded helix. At the same time, detailed analysis of water, metal, ligand and protein binding to RNA has revealed the effect of these moieties on folding and structure formation (Holbrook, 2005). It is this 3D structure that largely determines the functional activity of the RNA.

RNA plays numerous key roles in biological processes, including protein synthesis, mRNA splicing, transcriptional regulation and retroviral replication. Messenger RNA (mRNA) is a single stranded molecule used as the template for protein translation. Other non-coding RNA (ncRNA) have been discovered more recently, that have functional or catalytic roles in many cell processes including the regulation of transcription, DNA replication and RNA processing and modification (reviewed in Costa, 2005). Currently, there are six main classes of ncRNA, namely transfer RNA (tRNA) and ribosomal RNA (rRNA), both of which are involved in the process of translation and gene expression, small nuclear RNA (snRNA), which are mostly components of the spliceosome, small nucleolar RNA (snoRNA), that are mostly involved in rRNA modifications, micro RNAs (miRNA), and small interfering RNAs (siRNA), which are both thought to regulate the expression of genes. One of the newest discoveries of ncRNA is RNA interference (RNAi). RNA interference (RNAi) is the process where the introduction of double stranded RNA into a cell inhibits gene expression in a sequence dependent fashion. RNAi is seen in a number of organisms such as Drosophila, nematodes, fungi and plants, and is believed to be involved in anti-viral defence, modulation of transposon activity, and regulation of gene expression (Henikoff, 2002).

The fact that RNA molecules can be both informational and diverse in structure has led to suggestions that RNA catalysis may have played a key role during the early evolution of life on this planet (Woese, 1967; Crick, 1968) and that the cell used RNA as both the genetic material and the structural and catalytic molecule, rather than dividing these functions between DNA and protein as they are today. This hypothesis became known as the "RNA world hypothesis" of the origin of life (Gilbert, 1986). In 2001, the RNA world hypothesis was given a major boost with the deciphering of the 3D structure of the ribosome (Yusupov *et al.*, 2001). Many long-standing questions were resolved by the crystal structure. A critical issue was whether the rRNA serves as a structural scaffold, or whether it is directly involved in ribosomal function. The structures showed that rRNA in fact does both of these things, creating the structural framework for the ribosome and at the same time, playing important roles in its functional sites (Noller, 2005).

2.3.3 Protein sequence, structure and function

Classified by biological function, proteins include the enzymes, which are responsible for catalyzing the thousands of chemical reactions of the living cell; structural proteins, such as tubulin, keratin or collagen; transport proteins, such as hemoglobin; regulatory proteins, such as transcription factors or cyclins that regulate the cell cycle; signalling molecules such as some hormones and their receptors; defensive proteins, such as antibodies which are part of the immune system; and proteins that perform mechanical work, such as actin and myosin, the contractile muscle proteins.

Every protein molecule has a characteristic three-dimensional shape or conformation, known as its native state. Fibrous proteins, such as collagen and keratin, consist of polypeptide chains arranged in roughly parallel fashion along a single linear axis, thus forming tough, usually water-insoluble, fibres or sheets. Globular proteins, e.g., many of the known enzymes, show a tightly folded structural geometry approximating the shape of an ellipsoid or sphere. The precise 3D structure of a protein molecule is generally required for proper biological function, since the specific conformation is needed that cell factors can recognise and interact with. If the tertiary structure is altered, e.g., by such physical factors as extremes of temperature, changes in *p*H, or variations in salt concentration, the molecule is said to be denatured; it usually exhibits reduction or loss of biological activity. The process by which a protein sequence assumes its functional shape or conformation is known as folding. Protein folding can be considered as a hierarchical process, in which sequence defines secondary structure, which in turn defines the tertiary structure (see figure 2.6). Other molecules, such as chaperones, may also direct the folding of large newly synthesized proteins into their native 3D structure.

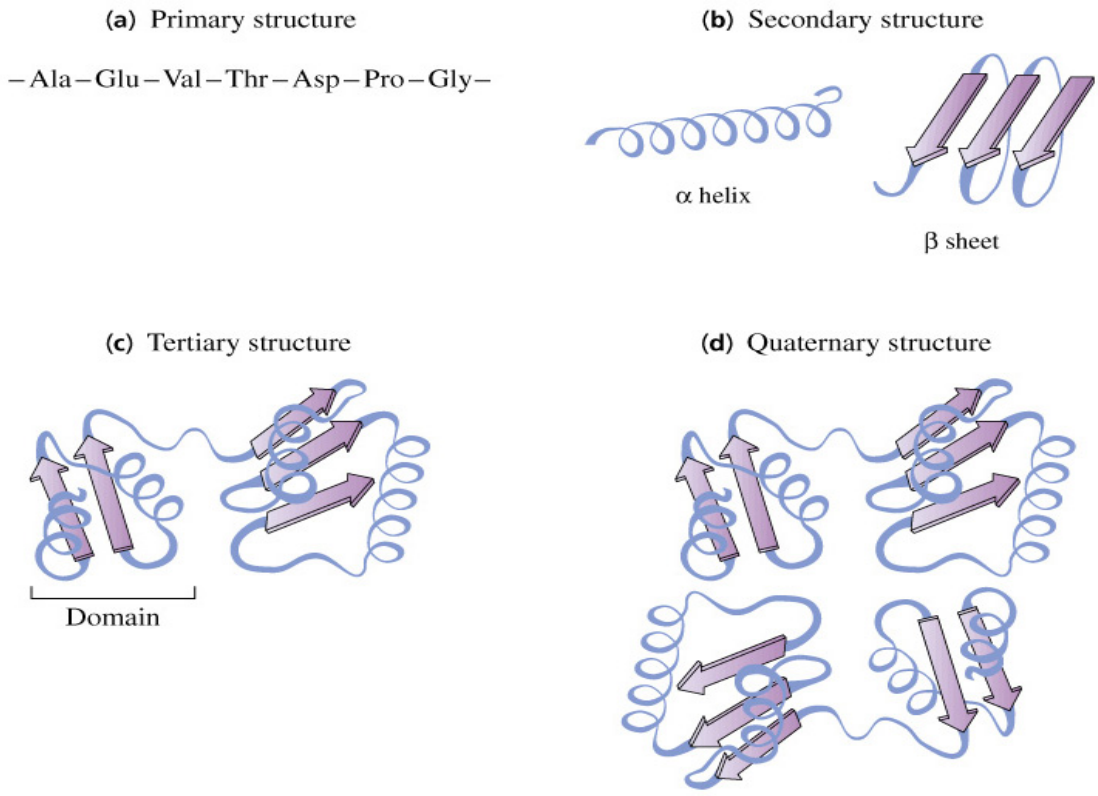


Figure 2.6 Different levels of protein structure
(from Principles of biochemistry, Horton, Moran, Ochs, Rawn, Scrimgeour). The ribbons represent examples of the four levels of protein structure. (a) The linear sequence of amino acid residues defines the primary structure. (b) Secondary structure consists of regions of regularly repeating conformations of the peptide chain, such as alpha helices and beta sheets. (c) Tertiary structure describes the shape of the fully folded polypeptide chain. The example shown has two domains. (d) Quaternary structure refers to the arrangement of two or more polypeptide chains into a multi-subunit molecule.

Although most protein sequences have a unique 3D confirmation, the inverse is not true. A 3D structure does not have a unique sequence, i.e. the size of the structure space is much smaller than the size of the sequence space. It is commonly assumed that there are around 1000 different protein folds, covering 10,000 different protein sequence families (Wang, 1998). A direct relationship has been clearly established between protein sequence similarity and

conservation of 3D structure (Chothia and Lesk, 1986; Yang and Honig, 2000; Koehl and Levitt, 2002a). Although exceptions exist, it is generally believed that when two proteins share 50% or higher sequence identity, they will generally share the same structural fold. However, in the so-called "twilight zone" of 20–30% sequence identity, it is no longer possible to reliably infer structural similarity (Chung and Subbiah, 1996). High sequence identity, but low structural similarity can occur due to conformational plasticity, solvent effects or ligand binding. Conversely, proteins in the 'twilight zone' of sequence similarity (<25% identity) can share surprisingly similar 3D folds (Gan *et al.*, 2002).

The relation between 3D fold and function is much more complex (Thornton *et al.*, 2000; Watson *et al.*, 2005) and the same fold is often seen to have different functions. After translation, the posttranslational modification (PTM) of amino acids can extend the range of functions of the protein by attaching to it other biochemical functional groups such as acetate, phosphate, various lipids and carbohydrates, by changing the chemical nature of an amino acid (e.g. citrullination) or by making structural changes, such as the formation of disulfide bridges (reviewed in Eichler and Adams, 2005). With respect to enzymes, local active-site mutations, variations in surface loops and recruitment of additional domains accommodate the diverse substrate specificities and catalytic activities observed within several superfamilies. Conversely, different folds can perform the same function, sometimes with the same catalytic cluster and mechanism (for example, trypsin and subtilisin proteinases). General rules seem to be that for pairs of domains that share the same fold, precise function appears to be conserved down to ~40 % sequence identity, whereas broad functional class is conserved to ~25 % (Wilson *et al.*, 2000). These results highlight the need to look beyond simple evolutionary relationships, at the details of a molecule's active site, to assign a specific function.

2.3.4 Towards a systemic definition of gene functions

Genes and gene products work together in complex, dynamic pathways to make a functional cell, tissue, and an organism as a whole. There is a lot of cross-talk between different proteins, DNA, and RNA to establish pathways, networks, and molecular systems. For example, a metabolic pathway is a representation of a set of frequently co-localized proteins, with various concentrations, interaction partners and 3D structures, that behave a certain way in the presence of particular metabolites. Other examples include RNA-mediated gene regulation, which is widespread in higher eukaryotes, and complex genetic phenomena like RNA interference, co-suppression, transgene silencing, imprinting, methylation, and possibly position-effect variegation and transvection, all involve intersecting pathways based on or connected to RNA signalling (Mattick, 2001). Clearly, understanding the biological role of genes requires an understanding of the spatial organization of the gene products into functional units such as complexes and organelles and the dynamic or temporal interactions between these units to control and carry out their various and complex biological functions (Aebersold, 2005). To add further complexity, during evolution, mutations in the gene can lead to different structure and function or can cause misfunctions, resulting in genetic disease phenotypes.

In the light of this growing complexity, the traditional definition of gene function, consisting of simple phrases of free text, is no longer sufficient. Traditionally, function has been defined in various terms, including molecular interaction, cellular process, or even phenotype of mutations, depending on the experimental perspective. These definitions now need to be standardised to allow automatic processing in high-throughout systems. Some progress has been made recently and the definition of function is moving toward a more

systematic representation. For example, ontologies are being developed that provide standardised function definitions for both RNA (ROC, Leontis *et al.*, 2006) and proteins (GO, Ashburner *et al.*, 2000). Other efforts have been directed towards organising functional data in databases, such as the ENZYME (Bairoch 2000) and KEGG (Kanehisa, 2002) databases.

To achieve a more global definition of gene function, ranging from its biochemical function to its role in the pathways, networks, cell and organism, such diverse information as 3D structures, cellular localisation, protein interactions and modifications, or mutations and their associated phenotypes must be assembled and classified. As a consequence, computational tools are needed for representing, integrating and modelling heterogeneous data as well as deciphering complex patterns and systems. The next two chapters will discuss the role of standardised vocabularies, known as ontologies, and information management systems in molecular biology.

"For speech is the means of association among men and in consequence, a wrong and inappropriate application of words obstructs the mind to a remarkable extent."
Francis Bacon, 1620

3 Ontologies

A more integrated approach to biology requires an interdisciplinary approach, with input from genetics and molecular biology, chemistry, pharmacology, computer science and mathematics amongst others. These diverse fields historically started rather independently and have grown quite distinguished terminologies. Standardised nomenclatures are now needed to facilitate the communication between experts from these different fields and to organise and merge the heterogeneous information produced by the different domains.

Data integration is currently hindered by syntactic differences in the file formats used by different applications and by semantic differences, such as naming conventions and terminology. The syntactic issue is now being addressed with the widespread adoption of standard file formats, such as the XML (eXtensible Markup Language) data exchange format. However, if data is to be truly understandable by multiple applications, semantic interoperability will also be necessary. Semantic ambiguities are ubiquitous, for example the same sequence may have different names in different sequence databases such as Genbank (Benson *et al.*, 2006) or EMBL (Cochrane *et al.*, 2006). Data integration is impeded by different meaning of identically named categories, overlapping meaning of different categories and conflicting meaning of different categories. Even the meaning of important high level concepts that are fundamental to molecular biology is ambiguous. For example, the term "gene" is shared by many disciplines, including genetics, molecular biology and population genetics. Historically, the term gene referred to an abstract concept to explain the hereditary basis of traits. A gene is now often defined in molecular terms as a complete chromosomal segment responsible for making a functional protein or RNA product, including both coding and regulatory regions (Snyder and Gerstein, 2003). The problem becomes more complex when natural language is used, for example for protein function definitions in the sequence databases, where the same function could be described as glycyl-tRNA synthetase, glycine-tRNA synthetase, glycine--tRNA ligase or simply glyQ. To resolve such semantic discrepencies, formal, structured vocabularies are now required, that constrain the use and interpretation of the terminology employed.

In the early nineties, ontologies were introduced in the context of 'knowledge modelling', rather than simple 'knowledge acquisition'. These structured depictions or models of known and accepted facts are being built today to make a number of applications more capable of handling complex and disparate information. Ontologies are used for example in artificial intelligence, the semantic web, and software engineering as a form of knowledge representation about the world or some part of it. They are used for communication between people and organisations by providing a common terminology over a domain. But perhaps the most important aspect of an ontology is that it creates a shared understanding of a domain in a format that can also be used by computers. They thus provide the basis for interoperability between different computational systems. They can be used for content explicitation for information resources and serve as an index to a repository of information. They can also be used as a basis for integration of information resources and as a query

model for information management systems that include automated inference and reasoning. Information management systems will be described in detail in chapter 4.

Section 3.1 introduces some basic concepts concerning the ontologies used in computer science. The formal representation of such ontologies is discussed in section 3.2. Some examples of successful biological ontologies are presented in section 3.3 and finally, section 3.4 introduces some of the main tools used to create and maintain bio-ontologies.

3.1 Ontologies in computer science

First, one should be aware of the distinction between ontology, the study of *being* as a branch of philosophy and individual (domain) ontologies, which are the result of the analysis of a particular domain of interest, possibly as broad as the universe. In philosophy, ontology is the most fundamental branch of metaphysics. It studies being or existence and their basic categories and relationships, to determine what entities and what types of entities exist. Ontology thus has strong implications for conceptions of reality. In computer science, an ontology is a data model that represents a domain and is used to reason about the objects in that domain and the relations between them. Ontologies can be of varying scope and content (Schulze-Kremer, 2002):

- *upper-level* ontologies are primarily concerned with general high level concepts that are the basis or our understanding of a particular domain;
- *domain* (also known as reference) ontologies are centred around a specific domain
- *task* (also known as application) ontologies are conceived for a specific problem solving task.

Gruber defines an ontology as "a formal, structured representation of the knowledge in a particular domain" (Gruber *et al.*, 1993). Ontologies generally consist of controlled concepts that represent classes or sets of instances in the world and the relationships that may exist between the concepts. For example, in the domain ontology shown in figure 3.1, some of the concepts associated with human development are represented, together with the fundamental relationships, *is_a* and *part_of.* Some ontologies also contain axioms that are used to constrain values for particular concepts or classes.

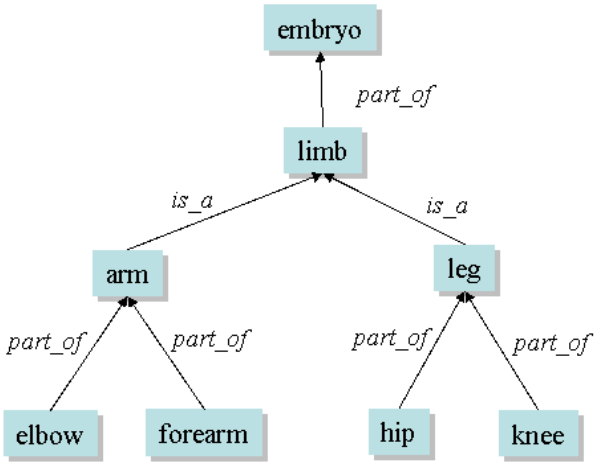


Figure 3.1 Example ontology
An extract from an ontology describing concepts associated with human development (Hunter et al., 2003). The concepts (blue boxes) are organised in a taxonomy with two different types of hierarchical relationships between concepts, is_a and part_of.

.

Instances are the `things' represented by a concept; for example, a human cytochrome C would be an instance of the concept Protein in a molecular biology ontology. Strictly speaking, an ontology should not contain any instances, because it is supposed to be a conceptualisation of the domain. The combination of an ontology with associated instances is what is known as a knowledge base.

3.1.1 Definition of concepts

A concept represents a set or class of entities or `things' within a domain. For example, protein is a concept within the domain of molecular biology. Concepts fall into two kinds:

- primitive concepts are those which only have necessary conditions (in terms of their properties) for membership of the class. For example, a globular protein is a kind of protein with a hydrophobic core, so all globular proteins must have a hydrophobic core, but there could be other things that have a hydrophobic core that are not globular proteins.
- defined concepts are those whose description is both necessary and sufficient for a thing to be a member of the class. For example, eukaryotic cells are kinds of cells that have a nucleus. Not only does every eukaryotic cell have a nucleus, every nucleus-containing cell is eukaryotic.

3.1.2 Definition of relations

Relations describe the interactions between concepts or the properties associated with concepts. Relations also fall into two broad kinds:

- Taxonomies that organise concepts into sub- or super-concept tree structures. The most common forms of these are :
    - The specialisation relationship commonly known as the *is_a* relationship. For example, an Enzyme *is_a* Protein, which in turn *is_a* Macromolecule.
    - The partitive relationship describes concepts that are part of other concepts, e.g. ModificationSite is *part_of* Protein.
- Associative relationships that relate concepts across tree structures. Commonly found examples include the following:
    - Nominative relationships describe the names of concepts, e.g. Protein *hasAccessionNumber* AccessionNumber (in the context of bioinformatics) and Gene *hasName* GeneName.
    - Locative relationships describe the location of one concept with respect to another, e.g. Chromosome *hasSubcellularLocation* Nucleus.
    - Associative relationships that represent, for example, the functions, processes a concept has or is involved in, and other properties of the concept, e.g. Protein *hasFunction* Receptor, Protein *isAssociatedWithProcess* Transcription and Protein *hasOrganismClassification* Species.

The relations, like concepts, can be organised into taxonomies. For example, *hasName* can be subdivided into *hasGeneName*, *hasProteinName* and *hasDiseaseName*. Relations also have properties that capture further knowledge about the relationships between concepts. These include, but are not restricted to:

- whether it is universally necessary that a relationship must hold on a concept. For example, when describing a protein database, we might want to say that Protein *hasAccessionNumber* AccessionNumber holds universally, i.e., for all proteins.
- whether a relationship can optionally hold on a concept, for example, we might want to describe that Enzyme *hasCofactor* Cofactor only describes the possibility that enzymes have a cofactor, as not all enzymes do have a cofactor.
- the cardinality of the relationship. For example, a particular AccessionNumber is the accession number of only one Protein, but one Chromosome may have many Genes.
- whether the relationship is transitive, for example if Protein *isAssociatedWithProcess* Transcription and Transcription *isAssociatedWithProcess* GeneExpression, then Protein *isAssociatedWithProcess* GeneExpression. The taxonomy relations *is_a* and *part_of* always have this property.

The OBO Relation Ontology provides consistent and unambiguous formal definitions of some widely used relational expressions used in bio-ontologies, in a way designed to assist developers and users in avoiding errors in coding and annotation. The aim of the Relation Ontology is to promote interoperability of ontologies and to support new types of automated reasoning about the spatial and temporal dimensions of biological and medical phenomena. Table 3.1 shows some examples of the different types of relationships defined in the Relation Ontology.

Foundational relations	Spatial relations	Temporal relations	Participation relations
is_a part_of instance_of	located_in contained_in adjacent_to	transformation_of derives_from preceded_by	has_participant has_agent

Table 3.1 Some of the relations described in the OBO Relation Ontology

3.2 Ontology representation

For ontologies to be used within an application, the ontology must be specified, that is, delivered using some concrete representation. The specification in the definition of ontologies by Gruber (see section 3.1) is the representation of this conceptualisation in a concrete form. The goal is to create a collaborative vocabulary and semantic structure for exchanging information about that domain. There are a variety of representations which can be used for ontologies, from lists of words, taxonomies, object-based knowledge representation languages such as Frames, and languages based on predicates expressed in logic such as Description Logics. These representations have varying characteristics in terms of their expressiveness, ease of use and computational complexity. Major considerations in the choice of representation are the expressivity and complexity of the encoding language, the rigour of the encoding and the semantics of a language. The three most widely used representations are:

- *Taxonomies* support the creation of simple tree-like inheritance structures. Although this provides great flexibility, the lack of structure in the representation can lead to difficulties with maintenance or preserving consistency, and there are usually no formally defined semantics. The single inheritance provided by a tree structure (each concept has only one parent in the *is_a* hierarchy) can also prove

limiting. Maintaining multiple inheritance hierarchies, however, is an arduous task.

- *Frame-based systems* provide more flexible structure. They are based around the notion of frames or classes which represent collections of instances (the concepts of the ontology). Each frame has associated attributes which can be filled by values or other frames. In particular, frames can have an *is_a* attribute which allows the assertion of a frame taxonomy. Frames are popular because frame-based modelling is similar to object-based modelling and is intuitive for many users. They have been used extensively for natural language processing, e.g. Ontolingua (Farquhar *et al.*, 1997).
- *Description Logics* (DLs) (Borgida, 1995) provides an alternative to frames. DLs describe knowledge in terms of concepts and relations that are used to automatically derive classification taxonomies. A major characteristic of a DL is that concepts are defined in terms of descriptions using other relations and concepts. In this way, the model is built up from small pieces in a descriptive way, rather than through the assertion of hierarchies. The DL supplies a number of reasoning services which allow the construction of classification hierarchies and the checking of consistency of these descriptions. These reasoning services can then be made available to applications that wish to make use of the knowledge represented in the ontology.

Description Logics and frames are similar, in that DLs are a logical reformulation of frames. An example is the OWL (Web Ontology Language), a markup language for publishing and sharing data using ontologies on the Internet. OWL is a vocabulary extension of the Resource Description Framework (RDF) and is derived from the DAML+OIL Web Ontology Language. Together with RDF and other components, these tools make up the Semantic Web project (Berners-Lee *et al.*, 2001).

3.3 Biological ontologies

In recent years, the utility of ontologies has been clearly demonstrated in several biological domains for the organisation and management of biological knowledge (Bard and Rhee, 2004). Ontologies are used for automatic annotation of data, for the sharing of information from different resources and for the presentation of domain knowledge to researchers, in particular to non-experts in the specific field. Clearly, the creation of an ontology demands a close collaboration between different disciplines, as shown in figure 3.2.

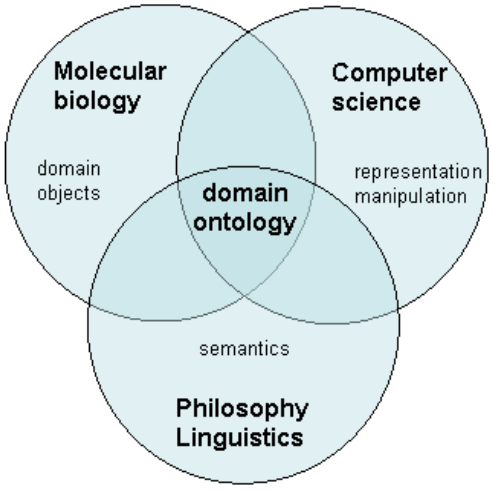


Figure 3.2 Interplay between ontologies, biology, computer science and linguistics
Molecular biologists discover facts that need to be organised and stored in databases. Computer scientists provide techniques for data representation and manipulation. Philosophers and linguists help organise the meaning behind database labels (adapted from Schulze-Kremer, 2002).

Since molecular biology and bioinformatics are in many aspects of enormous complexity, it is important to well understand beforehand the intended use for a newly to be developed ontology. Otherwise there is a great risk of losing focus and being overwhelmed by the multitude of facets leading to a failure of finishing a sufficiently complete, useful ontology. There is a tradeoff to be made between the flexibility and powerful reasoning capabilities of large, formal ontologies and the feasibility and pragmatism of more lightweight representations. To date, successful ontologies in computational biology have generally been of an intermediate level of technical complexity. By avoiding the more formal inference engines, and instead existing as large collections of hierarchical, taxonomic, categorisations of biological objects such as genes and proteins, they come closer to being specially structured databases. *i.e.* large relational objects.

3.3.1 Gene Ontology (GO)

The Gene Ontology (GO) (Ashburner *et al.*, 2000) is one of the most widely used biological ontologies. GO is an example of a *task* ontology, whose main use is as a controlled vocabulary for conceptual annotation of gene products in databases. It is essentially composed of three taxonomic hierarchies, representing the molecular function of a gene product; the process in which it takes place and the cellular location. Each set of categories is then equipped with one partial order representing *is_a* inheritance, and another representing *part_of* composition, so that more general categories are towards the top, and more specific categories are towards the bottom. It currently has over 10000 concepts within the ontology. The GO project began as a collaboration between three model organism databases: FlyBase (*Drosophila*), the *Saccharomyces* Genome Database (SGD) and the Mouse Genome Informatics (MGI) projects. Since then, the GO Consortium has grown to include many databases, including several of the world's major repositories for plant, animal and microbial genomes.

3.3.2 RiboWeb

RiboWeb (Altman *et al.*, 1999) is a frame-based *domain* ontology whose primary aim is to facilitate the construction of three-dimensional models of ribosomal components and compare the results to existing studies. The knowledge that RiboWeb uses to perform these tasks is captured in four ontologies: The physical-thing ontology; the data ontology; the publication ontology and the methods ontology. The physical-thing ontology describes ribosomal components and associated `physical things'. It has three principle conceptualisations: *Molecules*, *Molecule-Ensembles* and *Molecule-Parts*. The first describes covalently bonded molecules and includes the main biological macromolecules. *Molecule-Ensembles* captures non-covalently bonded collections of molecules, such as enzyme complexes. The *Molecule-Part* ontology holds knowledge about regions of molecules that do not exist independently, but need to be referred to by biologists. These would include amino acid side chains and the 3' and 5' ends of nucleic acid molecules. The data ontology captures knowledge about experimental detail as well as data on the structure of physical-things. The methods ontology contains information about techniques for analysing data. It holds knowledge of which techniques can be applied to which data, as well as the input and outputs of each method.

3.3.3 EcoCyc

Another example of a frame-based *domain* ontology is EcoCyc (Karp *et al.*, 1999a). The ontology covers *E. coli.* genes, metabolism, regulation and signal transduction and is used to specify a database schema. Scientists can visualise the layout of genes within the *E. coli.* chromosome, or of an individual biochemical reaction, or of a complete biochemical pathway with compound structures displayed. EcoCyc's use of an ontology to define a database schema has the advantages of its expressivity and ability to evolve quickly to account for the rapid schema changes needed for biological information. The user is not aware of this use of an ontology, except that the constraints expressed in the knowledge captured mean that the complexity of the data held is captured precisely. In EcoCyc, for example, the concept of *Gene* is represented by a class with various attributes, that link to other concepts: *Polypeptide product, Gene name, Synonyms* and *Identifiers* used in other databases etc. The representation system can be used to impose constraints on those concepts and instances which may appear in the places described within the system.

3.3.4 TAMBIS Ontology (TaO)

TAMBIS (Transparent Access to Multiple Bioinformatics Information Sources) (Baker *et al.*, 1998) uses an ontology represented by a DL to enable biologists to ask questions over multiple external databases using a common query interface. The TAMBIS ontology (TaO) (Baker *et al.*, 1999) describes a wide range of bioinformatics tasks and resources, and has a central role within the TAMBIS system. A user can form a complex, multi-source query, using the relationships defined in TaO. For example, starting with the concept *Protein*, the TaO is consulted as to which relationships can be used to join *Protein* to other concepts. Amongst many, the following two are offered: *isHomologous to Protein* and *hasAccessionNumber AccessionNumber*. Initially, the original *Protein* is extended to give a new concept *Protein isHomologous to Protein*. Then the second `protein' is extended with *hasAccessionNumber AccessionNumber*. The resulting concept (*Protein homologue of Protein with Accession Number*) describes proteins which are homologous to protein with a particular accession number. This concept can be used as a source independent query containing no information on how to answer such a query. The rest of the TAMBIS system takes this conceptual query and processes it to an executable program against the external sources. The TaO is available in two forms. The small TaO describes Proteins and Enzymes, as well as their motifs, secondary and tertiary structure, functions and processes. Important relationships include is*_component_of*, *has_name*, *has_function* and *is_homologous_to*.

3.3.5 Molecular Biology Ontology (MBO)

The Ontology for Molecular Biology (MBO) (Schulze-Kramer, 1997) is a general *upper-level* ontology, containing concepts and relationships that are required to describe biological objects, experimental procedures and computational aspects of molecular biology. In the conceptual part of the MBO, the primary relationship used is the *is_a* relationship. The root concept *Being* divides into *object* and *event*. *Object*, for example, is subdivided into *physical-* and *abstract- object*. This helps give a precise classification for lower level concepts - so, *physics* is an *abstract object* and *DNA* a *physical-object*. The actual biological content of the MBO is currently relatively small, ending at quite large grained concepts such as *Protein*, *Gene*, and *Chromosome*. The framework, however, exists for extending the MBO much further into the biological domain.

3.3.6 Open Biomedical Ontologies (OBO)

The feasibility and desirability of one comprehensive ontology for molecular biology versus several smaller task oriented ontologies has been extensively debated in the community (Schulze-Kremer, 2002). On the one hand, a comprehensive domain ontology would certainly be very helpful if it could be achieved and maintained. On the other hand, it seemed much more efficient and effective to have several smaller task or subdomain ontologies which take less time and expertise to grow and maintain and therefore are in the position to be put to use much sooner.

In principle, the approach of smaller subdomain ontologies is the more practical one. One of the major goals of the Open Biomedical Ontologies (OBO) consortium is to provide a set of compatible ontologies, which can be used in combination in order to integrate individual data resources into a coherent whole.

OBO (http://obo.sourceforge.net) is an ontology library for well-structured, controlled vocabularies for shared use across different biological domains. To date, over 50 ontologies have been registered at the site. Some examples are shown in figure 3.3. Acceptance on the OBO site implies that the ontology has been accepted as authoritative by the OBO group and that the ontology meets a number of specific criteria defined by the community. In particular, only a single ontology should be specified for each domain or task, and new ontologies should be orthogonal to the other ontologies already hosted within OBO.

Figure 3.3 The top level of the OBO hierarchy

The ontologies grouped together at the OBO web site cover a wide range of biomedical fields, such as specific organism anatomies, taxonomic classifications or transcriptomic and proteomic experimental protocols and data. Various ontologies have also been developed for particular aspects of single sequences, such as gene structure (SO) (Eilbeck *et al.*, 2005), protein function (GO) or protein–protein interactions (MI) (Hermjakob *et al.*, 2004). Some work has also begun to develop standard data formats to represent RNA sequences and structures and the RNA Ontology Consortium (ROC) (Leontis *et al.*, 2006) has been established to build a formal ontology.

The OBO ontologies can be accessed from a single location with a unified output format, using the Ontology Lookup Service (OLS) at http://www.ebi.ac.uk/ontology-lookup/. There are currently 43 ontologies available for querying and the new OBO ontologies will be automatically included.

3.4 Tools for ontology development

Tools are essential to aid the ontologist in constructing an ontology, and merging multiple ontologies. Such conceptual models are often complex, multi-dimensional graphs that are difficult to manage. For example, the DL GRAIL has associated tools to shield the ontologist from the logical formalism. An intermediate 'template' form is used to represent the conceptualisation, from which the encoding can be generated (Rogers *et al*., 1997). Ontology development tools also usually contain mechanisms for visualising and checking the resulting model. The MBO has an editor for creating and visualising the object based encoding used in that ontology (Schulze-Kremer, 1997). The frame based system used by EcoCyc also has the GKB editor (http://www.ai.sri.com/~gkb) for handling the conceptualisation and encoding in Frame Based Representations (Karp *et al.*, 1999). Such tools are essential for maintaining complex ontologies that are necessary for capturing knowledge within the biology domain. Many other tools for displaying and editing biological ontologies and some of the most widely used are listed here.

- OntoLingua from the Stanford Knowledge Systems Laboratory (http://www.ksl.stanford.edu/software/ontolingua/) provides a distributed collaborative environment to browse, create, edit, modify, and use ontologies. Later, the Chimaera tool was added (http://www.ksl.stanford.edu/software/chimaera/) in order to facilitate the merging of knowledge bases produced by different users for different purposes with different assumptions and different vocabularies. Later the goals of supporting testing and diagnosing ontologies arose as well. OntoLingua's and Chimaera's knowledge model is frame based.
- Protégé 2000 is an ontology editing software from Stanford Medical Informatics (http://smi.stanford.edu/projects/protege). Protégé is a free, open source ontology editor and knowledge-base framework. The Protégé platform supports two main ways of modelling ontologies via the Protégé-Frames and Protégé-OWL editors. Protégé ontologies can be exported into a variety of formats including RDF(S), OWL, and XML Schema.
- OilEd (http://www.ontoknowledge.org/oil/tool.shtml) is a graphical tool for creating and editing OIL (Ontology Inference Layer) ontologies developed at the University of Manchester. The knowledge model for OilEd is based on description logics.

- DAG-Edit (http://www.geneontology.org/doc/dagedituserguide/dagedit.html) was originally developed to maintain the Gene Ontology. The tool offers a graphical user interface to browse, search and edit GO files or other ontologies based on the directed acyclic graph (DAG) data structure. Ontologies can be output in the OBO file format.

These four editors were compared recently (Lambrix *et al.*, 2003) in terms of availability, functionality, visualisation and input and output formats, among other criteria. All the systems tested had particular strengths and weaknesses and no tool was superior in all aspects. The main strengths of Protégé 2000 compared to the other systems are its user interface, the extendibility and functionality of the plug-ins, as well as the different formats that can be imported and exported. Chimaera's main strengths concern its functionality, including ontology merging and diagnosis, the different formats that can be imported and exported, its help functionality, the shortcuts for expert users and the fact that multiple users can work with the same ontology. Howere, its user interface was its main weakness. The main advantage of OilEd is the fact that its model is description logic-based and that the underlying system can perform reasoning tasks such as classification and consistency checking. DAG-Edit was specifically built for GO ontologies and has an interface that is easy to use and learn.

3.5 Perspectives

The potential value of properly built ontologies for representing knowledge in the biological domain is immense. Ontologies will play a key role in the reconstruction of biological processes because they provide semantics of biological knowledge in a human- as well as in a computer-readable form. There has been some debate as to the most suitable representation for bio-ontologies. Tree-based ontologies are easier to build, and have been used in practical applications such as automatic database annotation. However, more rigorous formalisms provide a logical framework that, in the future, will allow automated reasoning and hypothesis generation.

Bio-ontologies for the obvious knowledge domains are now in place and are under active curation. A difficult problem is the interoperability between the different ontologies, because ontology development in biology is a relatively new field, and currently bio-ontologies contain many redundant or overlapping concepts. Therefore, the exchangeability and interoperability among ontologies and databases has to be addressed.

The majority of bio-ontologies were built to provide a common vocabulary and for a standard annotation; however, ontologies have far greater potential and could open up whole new possibilities for biological research. The formation of a set of integrated ontologies at different levels of representations will significantly increase interoperability between domain data and knowledge, and enable new intelligent bioinformatics applications. Unfortunately, most of the current search and analysis tools for mining these data do not exploit the full power of the ontologies and their associations with data objects.

"As a general rule the most successful man in life is the man who has the best information."
Benjamin Disraeli (1804 - 1881)

4 Information management systems

The new integrative, systems-level studies need to exploit the multitude of heterogeneous and autonomous data resources that include genomic sequences, 3D structures, cellular, phenotype and other types of biologically relevant information. As more and more biological data are generated, the problem of efficient retrieval and analysis of this data will become an important scientific bottleneck. The principal difficulties are due to:

- Volume: impact of terabyte scale experimentation
- Inaccessibility: navigation of diverse, distributed datasets
- Integrity: data that are of poor quality
- Intractability: good data, not useful for computational purposes (e.g. literature)

A major challenge for bioinformaticians is therefore the efficient processing of this mass of experimental and predicted data, in order to produce useful information and to render the information accessible to the biologist (Roos, 2001).

In the context of Knowledge Management (KM), the distinction between data, information and knowledge has been described explicitly (see figure 4.1) (Zeleny, 1987). In KM, data is defined as a list of simple facts or observations without any context or meaning. The context and the associations or relations between data are needed before the data can be transformed into useful information. Thus, information can be considered as being organised data that has been given meaning by way of the relationships between pieces of data. For example, single entries in a database are data, whereas reports created from intelligent database queries result in information.

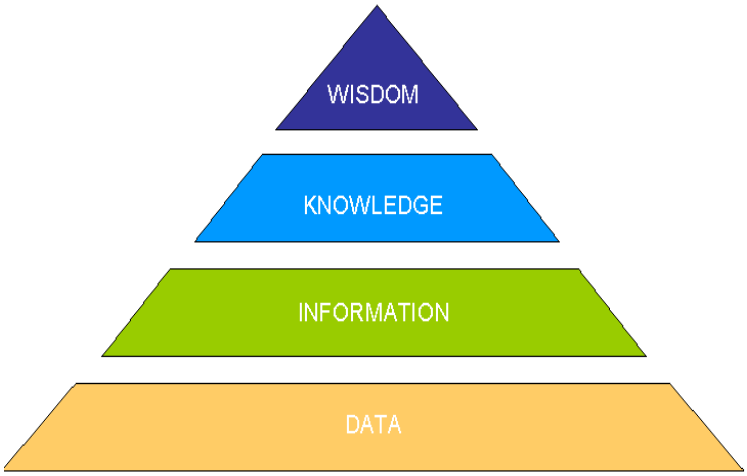


Figure 4.1 Transition of data into wisdom

While combining data and meaning to create information is extremely useful, the detection of patterns, trends and exceptions extends the value of the information. The next level of elevated understanding is knowledge. Knowledge is different from data or information in that it can be created by accumulating enough information or using logical inferences. Wisdom is the utilisation of accumulated knowledge to predict how the patterns or trends will change under certain conditions and to construct logical decision rules.

Traditionally, the information produced by bioinformatics studies has been interpreted by a human expert who had the experience necessary to understand the patterns revealed by the computational analyses. In the post-genomic era, the volume of data available requires automatic processing by 'intelligent' computer systems that are capable of understanding the relations and patterns hidden in the data. To achieve this, the basic knowledge in the domain of interest needs to be represented in a format that can be understood by the computer. As we saw in chapter 3, ontologies provide an ideal means of representing the fundamental concepts in a domain. Ontologies can thus provide the context and the explicit knowledge required for automatic information management and knowledge extraction systems.

The goal of ontology-based information management systems (IMS) is thus to combine information from different data resources into a unified system, such that the cumulative information provides greater biological insight than is possible if the individual information sources are considered separately. IMS are designed to help biologists systematically gather and exploit all the data crucial for their research, by automating many aspects, from data acquisition to knowledge discovery (see figure 4.2). The development of effective IMS requires a multidisciplinary domain drawing on research from such fields as databases and knowledge acquisition for expert systems, high performance computing, machine learning, reasoning with uncertainties and data visualization.

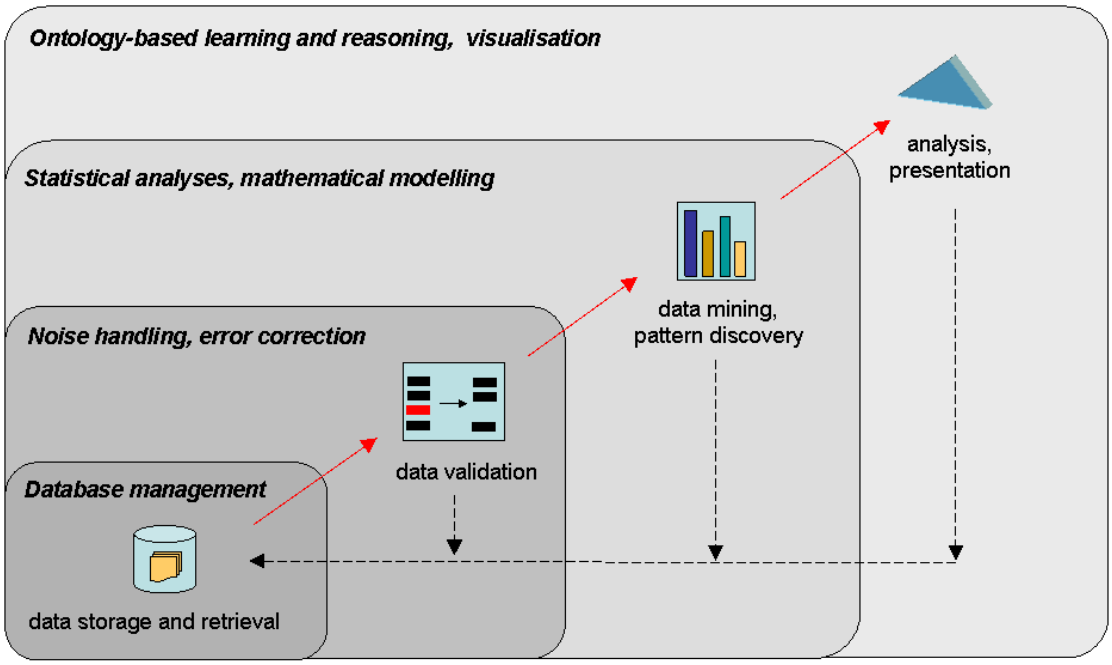


Figure 4.2 The knowledge discovery process

Four basic steps or processes in the knowledge discovery process can be identified (figure 4.2) : (i) database theories and tools provide the necessary infrastructure to store, access and manipulate data, discussed in section 4.1 (ii) data validation to filter or correct the errors and uncertainties from the large datasets, discussed in section 4.2 (iii) data mining to discover mainly hidden patterns, associations, structure and/or anomalies (section 4.3) and (iv) information analysis, interpretation and presentation (section 4.4).

4.1 Data storage and retrieval

In a dynamic heterogeneous environment such as bioinformatics, many different databases and software systems are used. Existing biological databases are typically highly focused, containing raw data of a specific type. However, entries stored in different databases can be strongly related and mutually dependent on each other. For example, the function of a gene depends on its biological context: its interactions with other genes, the pathways they are involved in, their expression under certain conditions, etc. Similarly, the function of an interaction depends on the function of the interacting partners. To retrieve the broader view of an entity, a biologist usually has to search multiple databases. This poses a number of problems. Most public databases have their own format and querying system. Links between databases are not always available and are not always coordinated between the different resources, giving rise to problems of consistency, redundancy, connectivity and synchronisation. Even on a small scale, for example for a single protein or complex, data integration becomes a daunting task.

To overcome these problems, new data integration systems are needed, that read data from multiple sources, perform simple transformations of data into a unified format and provide access to the data (Wong, 2002). Data sources might be simple text files, XML formatted documents or might be stored in relational database management systems (DBMS), such as Oracle or MySQL. To counter the increasing dispersion and heterogeneity of data, different approaches to integrating these data sources are appearing throughout the bioinformatics community. Two main approaches have been taken: the data warehousing approach and the distributed database approach (Davidson *et al.*, 1995).

4.1.1 Data warehousing: local storage and retrieval

The traditional approach to this problem has been data warehousing, where all the relevant databases are stored locally in a unified format and mined through a uniform interface. SRS (Etzold and Argos, 1993) and Entrez (Schuler *et al.*, 1996) are probably the most widely used database query and navigation systems for the life science community. They provide graphical user interfaces to access a broad range of scientific databases, including genome and protein sequences, metabolic pathways and literature abstracts. Entities are stored in a table and specific fields are extracted for indexing and cross-referencing. EnsMart (Kasprzyk *et al.*, 2004) is another warehouse-based system that is distinguished by its user-friendly query front end that allows users to compose complex queries interactively. EnsMart runs on Oracle and MySQL. Data types currently supported by EnsMart are genes, SNP data, and controlled vocabularies. More recently, the Atlas system (Shah *et al.*, 2005) provides a data warehouse based on relational data models, that locally stores and integrates biological sequences, molecular interactions, homology information and functional annotations of genes. First, Atlas stores data of similar types using common data models and second, integration is achieved through a combination of APIs, ontology and tools to perform

common sequence and feature retrieval tasks. Because the databases are installed locally, data retrieval is direct, efficient and relatively simple. Warehousing guarantees that the data needed are always available. Also users have control over the databases installed, which versions are used and when they are updated. The disadvantage of this approach is that the overhead costs can be very heavy in terms of the hardware required for database installation and maintenance.

4.1.2 Distributed databases and remote access

Distributed systems implement software to access heterogeneous databases that are dispersed over the internet and provide a query facility to access the data. Many examples have been developed. OPM (Object Protocol Model) (Topaloglou *et al.*, 1999) uses an entity model and generic servers that retrofit the data and unify data sources, and provides a query language OPM-MQL to query the distributed data. The integrated data can be output in XML format. IBM's DiscoveryLink (Hass *et al.*, 2001) uses a relational model and the SQL language for modelling and accessing distributed data. TAMBIS (Baker *et al.*, 1998) is a semantic-based system utilising ontologies and a services model to support user queries. BioMOBY (Wilkinson *et al.*, 2003) like TAMBIS, is also ontology-based and service-model driven. SEMEDA (Kohler and Schulze-Kremer, 2002) is an ontology based semantic metadatabase and is implemented as a 3 tiered architecture consisting of a relational database (backend) and jsp 1.1 (java server pages) as the middle tier, which dynamically generates the html frontend. Using this architecture has several advantages: data (ontologies and database metainformation) can be consistently stored independently from the application and can also be retrieved or imported by using the various built in interfaces and tools of the DBMS. These implementations do not house the data locally, but instead query the original data resource for available services before sending queries. These systems are powerful for interrogating disparate data sources. However, a disadvantage is that large queries may take a long time to return or may not be returned at all. Thus, remote access requires complex systems to manage communication between the server and the client, particularly when errors occur because remote systems are not available.

4.2 Data validation

Many of the large scale experimentally- or computationally-derived datasets used by systems biologists are error prone, including both false positives and false negatives. Some of the experimental errors are due to technical irreproducibility and some to biological features, such as the heterogeneity of gene expression levels in different cell populations, even under well-controlled conditions. Computational analyses based on faulty assumptions, improper statistical validation, or incomplete datasets, can also be misleading. The pre-existing biological literature is not perfect. For example, the level of error in genome functional annotations as been estimated to be about 5-8% for more general enzymatic functions to more than 30% for specific functions, such as substrate specificity (Devos and Valencia, 2001). At any given time, biological conclusions are drawn within a context from which key features might well be missing because they have not yet been discovered or conceptualized. A major technical challenge for integrative systems is developing procedures for handling error so that legitimate interpretations can be made from the available data. The reasons for wanting to minimize errors are straightforward. Errors can be propagated from one dataset to another and in hierarchical systems, cascading can occur where errors are allowed to propagate unchecked from layer to layer repeatedly, until a flood of incorrect information has been generated. Error detection and correction must therefore be integral parts of building and

maintaining databases.

4.2.1 Approaches to Noise Handling

Imperfections in a data set can be dealt with in three broad ways. We may leave the noise in, filter it out, or correct it (Teng, 2003). In the first approach, the data set is taken as is, with the noisy instances left in place. Algorithms that make use of the data are designed to be robust; that is, they can tolerate a certain amount of noise in the data. Robustness is typically accomplished by avoiding overfitting, so that the resulting classifier is not overly specialized to account for the noise. In the second approach, the data is filtered before being used. Instances that are suspected of being noisy according to certain evaluation criteria are discarded. A classifier is then built using only the retained instances in the smaller but cleaner data set. In the third approach, the noisy instances are identified, but instead of discarding these instances, they are repaired by replacing the corrupted values with more appropriate ones. These corrected instances are then reintroduced into the data set.

There are pros and cons to adopting each of these three approaches to noise handling. Robust algorithms do not require pre-processing of the data, but each algorithm has to institute its own noise handling routine, duplicating the effort required even if the same data set is used in each case. In addition, the noise in the data may interfere with the mechanism, affecting the performance of the resulting classifier. By filtering out the noisy instances from the data, there is a tradeoff between the amount of information available for building the classifier and the amount of noise retained in the data set. Filtering is not information-efficient; the more noisy instances we discard, the less data remains. Noise correction, when carried out correctly, preserves the maximal information available in the data set. A classifier built from this corrected data should have a higher predictive power and a more compact representation. Thus, the most efficient and practical solution will probably be to combine the advantages of the three different approaches.

4.3 Data mining

Sensitive data mining systems are now required to manage and extract the knowledge that is potentially buried in the hundreds of terabytes of data distributed over the various Internet-based resources. The goal of data mining, also known as Knowledge Discovery in Databases (KDD) is to detect patterns or relationships in the data that might lead to hidden information thereby enabling intelligent, knowledge-driven decision-making. New data mining techniques are being developed, in fields such as statistics, artificial intelligence and rule-based approaches, as well as in clustering and classification methods. For example, decision trees are being used to identify possible targets in high-throughput structural proteomics (Bertone *et al.*, 2001). Association rule discovery is used for finding and describing relationships between different items in a large data set (Oyama *et al.*, 2002; Creighton and Hanash, 2003). Correlation analysis and clustering is used to determine local structural information such as the catalytic triad, metal binding sites and the N-linked glycosylation site (Oldfield, 2002). Clustering of gene expression profiles is another area of research attracting much effort (e.g. Shannon *et al.*, 2003). These methods of data mining are often used in combination with each other, either in parallel or as part of a sequential operation.

Data resources written in human language such as the scientific literature pose particular problems for mining techniques. Natural language is ambiguous and the syntax is typically

author-specific. Data may be represented in the main body of the text, in a footnote, in a table or embedded in a graphical illustration. The simplest approach to text-mining is to identify entities that co-occur within abstracts or sentences. As two entities might be mentioned together without being in any way related, most systems use a frequency-based scoring scheme to rank the extracted relationships. However, complex sentences that contain multiple relationships can give rise to erroneous relationships. This approach is also unable to extract directional relationships and has difficulty distinguishing between direct and indirect relationships. These issues are addressed by natural language processing (NLP) approaches, that combine the analysis of syntax and semantics. The text is first 'tokenized' to identify sentence and word boundaries, and a part-of-speech tag (for example, a noun or verb) is assigned to each word. A syntax tree is then derived for each sentence to delineate noun phrases and represent their interrelationships. Simple dictionaries are subsequently used to semantically tag the relevant biological entities (for example, genes and proteins) and other keywords. Finally, a rule set is used to extract relationships on the basis of the syntax tree and the semantic labels. Co-occurrence and NLP methods are reviewed in detail in (Andrade and Bork, 2000).

Text mining can be used to uncover overlooked relationships and to make novel hypotheses by combining information from multiple papers. However, the full discovery potential of such tools will only be realized with the advent of new integrated approaches that combine the literature with other large data sets such as genome sequences, protein–protein interaction screens or microarray expression studies (Jensen *et al.*, 2006).

4.4 Data analysis and presentation

The goal of IMS is not simply to store existing data for efficient querying. They go beyond data integration, to include unique derived data that is computed within the system. For example, similarity data between objects, modules that expand existing data types based on inference, refinement of existing objects, generation of new data types by processing existing data types and other derived data. Close integration of these software protocols into a fully automatic ensemble is necessary to enable smooth operation, minimizing the necessity for the operator's special knowledge of the underlying methods. Some efforts are now being made to develop software protocols and models to facilitate the automatic integration of different biological data and applications. The complexity of the systems encourages using object-oriented models and implementation technologies. The process is made even more complex by the need to exchange data amongst the distributed resources on a real-time basis in order to achieve optimal synchronization. Protocols, such as Corba, DAS (http://www.biodas.org) or the Systems Biology Workbench (Hucka *et al.*, 2002), are therefore required that manage communication among distributed applications. The volumes of data now being generated and the amount of computing needed to process them have also lead to the application of new computational techniques, such as massively parallel supercomputers, or GRID technologies (Foster, 2003) which enable large-scale sharing of data and computational resources across geographically distributed groups. For instance, the European HealthGRID project (http://www.healthgrid.org) covers a range of biomedical information from the molecular level (genetic and proteomic information) over cells and tissues, to the individual and finally the population level (social healthcare).

An example of a biological IMS is the BioSPICE (Kumar and Feidler, 2003) program, whose goal is to create a framework that provides biologists access to the most current

computational tools. Contributions from approximately 20 different laboratories have been integrated under the BioSPICE Dashboard and a methodology for continued software integration has been implemented. A graphical environment is available that combines Open Agent Architecture and NetBeans software technologies in a coherent, biologist-friendly user interface. The current Dashboard permits data sources, models, simulation engines, and output displays provided by different investigators and running on different machines to work together across a distributed, heterogeneous network.

Another example is the Pegasys software system (Shah *et al.*, 2004) that facilitates the execution and integration of biological sequence analyses, such as *ab initio* gene prediction or pairwise and multiple alignment. The software allows users to dynamically create analysis workflows of sequence analyses by manipulating a graphical interface. Results are stored in a relational database management system and can be exported in General Feature Format (GFF) or XML format for import to other tools.

Other systems have been developed for more specific applications, such as the Genome Information Management System (Cornell *et al.*, 2003). GIMS is an object database used to store *Saccharomyces cerevisiae* data that integrates genomic data with data on the transcriptome, protein-protein interactions, metabolic pathways and annotations, such as gene ontology terms and identifiers. The resulting system supports the running of analyses over this integrated data resource, and provides comprehensive facilities for handling and inter-relating the results of these analyses.

4.4.1 Visualisation

An important aspect of these integrated systems will be the accessibility of the results for the biologist. The size and complexity of the data are prohibitive to textual views, whereas a graphical representation allows an intuitive view of complex data. It allows a view of the overall organisation and structural properties and to highlight patterns within the data. Support for interactive data exploration or navigation is also needed for browsing and searching and for manipulation of data structures in order to facilitate analysis from multiple perspectives (Robinson and Flores, 1996).

In bioinformatics, visualisation tools are widely used for displaying specific objects such as phylogenetic trees, protein structures, multiple sequence alignments or protein-protein interaction networks. Recently, more flexible systems have been developed that allow the visualisation of diverse objects in the same interface. For example, Space Explorer (Gilbert *et al.*, 2000) is a system that allows different biological data, such as protein clusters and phylogenetic trees, protein 3D toplogies or gene expression data to be visualized and explored interactively. With this system, objects are directly mapped onto a 1D, 2D or 3D Euclidean space and coordinates are calculated for an optimal rendering of distances between objects. Space Explorer combines 3D visualisation with hierarchical clustering to provide an interactive web-enabled virtual reality environment.

4.5 Conclusions

IMS integrate data from diverse sources, including both experimentally validated and predicted data. To provide reliable information to the biologist, it will be crucial to be able to trace the source of the data and to determine the reliability of the information at all stages of

the knowledge discovery process. Accessibility and ease-of-use for the biologist also need to be taken into account during the IMS design process. Such systems will soon become essential tools for the systematic exploitation of all the data related to a particular research project and will have widespread implications for automatic knowledge discovery and research hypothesis testing.

"Nothing in biology makes sense
except in the light of evolution."
- Theodosius Dobzhansky (1900-1975)

5 The central role of sequence alignments

5.1 Introduction

During evolution, random mutagenesis events take place, which change the gene sequences that encode RNA and proteins. There are several different types of mutation that can occur. Point mutations substitute a single nucleic or amino acid residue for another one. Residue insertions and deletions also occur, involving a single residue up to several hundred residues. Other evolutionary mechanisms at work in nature include genetic recombination, where DNA strands are broken and rejoined to form new combinations of genes. Some of these evolutionary changes will make a protein non-functional, e.g. most mutations of active site residues in an enzyme, or mutations that prevent the protein from folding correctly. If this happens to a protein that carries out an essential process, the cell (or organism) containing the mutation will die. As a result, residues that are essential for a protein's function, or that are needed for the protein to fold correctly, are conserved over time. Occasionally, mutations occur that give rise to new functions. This is one of the ways that new traits and eventually species may come about during evolution. By comparing related sequences and looking for those residues that remain the same in all of the members in the family, we can learn a lot about which residues are essential for function (Lesk 1994). Thus, multiple sequence comparison or alignment 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 not only identify important structural or functional motifs that have been conserved through evolution, but can also highlight particular non-conserved features resulting from specific events or perturbations (Woese and Pace, 1993; Lecompte *et al.*, 2001).

5.1.1 Multiple alignment definitions

In the most general terms, an alignment represents a set of sequences using a single-letter code for each amino acid or nucleotide (figure 5.1). Each horizontal row in the alignment represents a single sequence and structurally, functionally or evolutionarily equivalent residues are aligned vertically. When the sequences are of different lengths, insertion-deletion events are postulated to explain the variation and gap characters are introduced into the alignment.

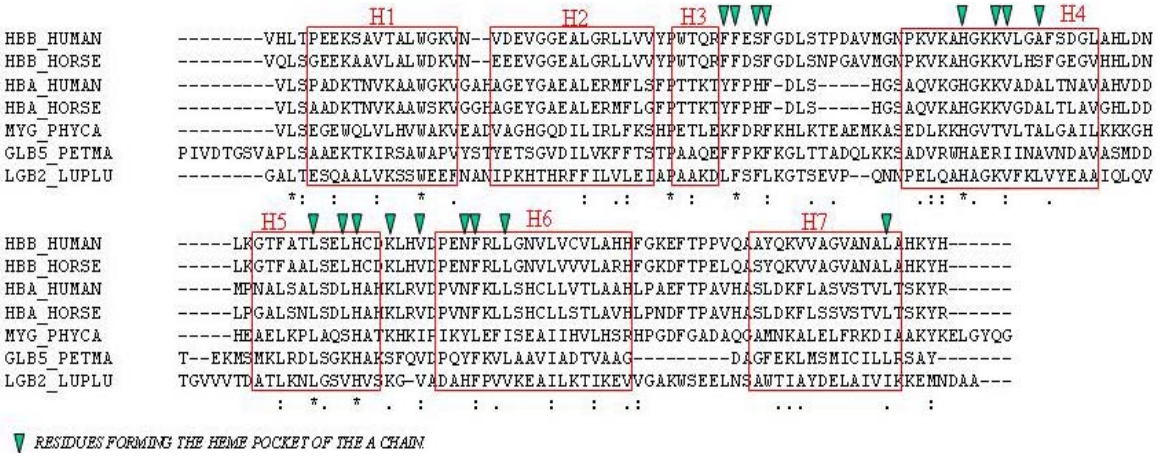


Figure 5.1 Example alignment of a set of 7 hemoglobin domain sequences
*The alignment shows the 7 helical structure (PDB:1a00) and the conserved residues forming the heme pocket of the beta subunit (green triangles). The symbols below the alignment indicate conserved positions: * = fully conserved identical residue, : = fully conserved 'similar' residue, . = partially conserved 'similar' residue.*

Alignments are produced by a wide variety of programs, sometimes as a side-product of the main function of the program. However, at least four different varieties of multiple alignment exist, as illustrated in figure 5.2 (from Lecompte *et al.*, 2001).

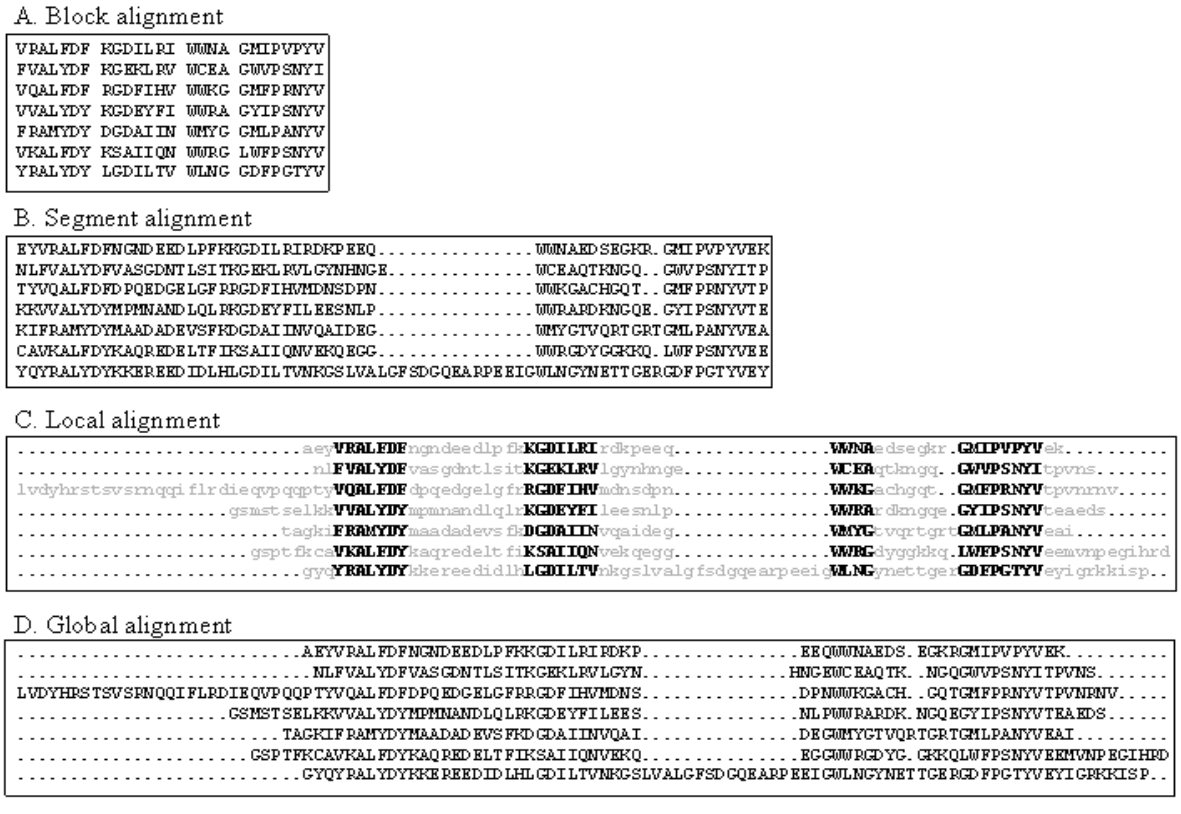


Figure 5.2 Four different types of multiple sequence alignment

The block alignment (figure 5.2A) represents only the conserved motifs and does not contain any gaps. It is used by the Probe program (Neuwald *et al.*, 1997) and in the Blocks database (Henikoff *et al.*, 2000). The segment alignment (figure 5.2B) is used by a number of

database search programs such as Blast (Altschul *et al.*, 1990) and PSI-Blast (Altschul *et al.*, 1997), and in pattern / domain databases such as Pfam (Bateman *et al.*, 2004) or ProDom (Bru *et al.*, 2005). It contains the most similar regions of the sequences and may contain short gaps representing indels. The local and global alignments (figure 5.2C,D) both contain the complete protein sequences and are typically produced by multiple alignment programs such as Dialign (Morgenstein *et al.*, 1996) or ClustalW (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.

5.1.2 Multiple Alignments of Complete Sequences (MACS)

In order to allow the maximum integration of biological information in the context of the complete protein family, a multiple alignment of the full length of the sequences is essential. Global Multiple Alignments of Complete Sequences (MACS) provide an ideal basis for more in-depth analyses of protein family relationships. By placing the sequence in the context of the overall family, the MACS permits not only a horizontal analysis of the sequence over its entire length, but also a vertical view of the evolution of the protein. The MACS thus represents a powerful integrative tool that addresses a variety of biological problems, ranging from key functional residue detection to the evolution of a protein family. The MACS now plays a fundamental role in most areas of modern molecular biology, from shaping our basic conceptions of life and its evolutionary processes, to providing the foundation for the new biotechnology industry.

5.2 Multiple alignment applications

5.2.1 Phylogenetic studies

One of the earliest applications of multiple sequence alignments was in phylogenetic studies. Phylogenetics is the science of estimating the evolutionary past, in the case of molecular phylogeny, based on the comparison of DNA or protein sequences. For example, the accepted universal tree of life, in which the living world is divided into three domains (bacteria, archaea, and eucarya), was constructed from comparative analyses of ribosomal RNA sequences (figure 5.3).

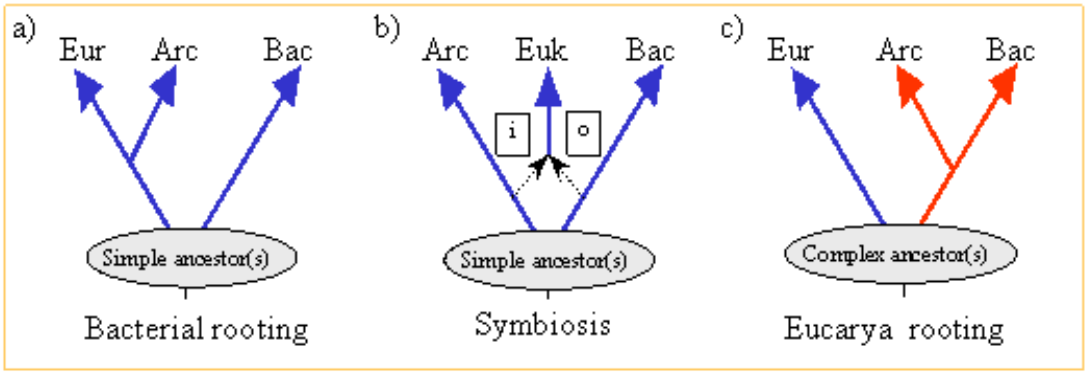


Figure 5.3 Alternative hypotheses for the rooting of the tree of life
In b), i indicates informational proteins and o indicates operational proteins.

According to this rRNA-based tree, billions of years ago a universal common prokaryotic-like ancestor gave rise to the two microbial branches, the archaea and bacteria (collectively called prokarya) and later, the archaea gave rise to the eukarya (Iwabe *et al.*, 1989) (figure 5.3a). More recently, analyses based on whole-genome comparisons have suggested that the eukaryotic lineage arose from metabolic symbiosis between eubacteria and methanogenic archaea (Lopez-Garcia and Moreira, 1999) (figure 5.3b). In this case, early eukaryotes would be a chimera of eubacterial and archaeal genes, in which the operational genes were primarily from the eubacteria, and the informational genes from the archaea. But some important eukaryotic genes have no obvious predecessors in either the archaeal or the bacterial lines, and an alternative has been suggested where prokaryotes would have evolved by simplification of an ancestral eukaryotic-like genome (Forterre and Philippe, 1999; Poole *et al.*, 1999) (figure 5.3c). In a comprehensive study of ribosomal genes in complete genomes from 66 different species, the archaeal ribosome appeared to be a small-scale model of the eukaryotic one in terms of protein composition (Lecompte *et al.*, 2002), which would support the eukaryotic-rooting tree.

The methods for calculating phylogenetic trees fall into two general categories (Page and Holmes, 1998). 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). All of these methods use distance measures based on the multiple sequence alignment and the strategy used to construct the alignment can have a large influence on the resulting phylogeny (Morrison and Ellis, 1997).

5.2.2 Comparative genomics

Of course, in the current era of complete genome sequences, it is now possible to perform comparative multiple sequence analysis at the genome level (Hardison, 2003). As genomes evolve, large-scale evolutionary processes, such as recombination, deletion or horizontal transfer, cause frequent genome rearrangements (Shapiro, 2005). Comparative analyses of complete genomes present a comprehensive view of the level of conservation of gene order, or synteny, between different genomes, and thus provide a measure of organism relatedness at the genome scale (Darling *et al.*, 2004; Elnitski *et al.*, 2005; Ye and Huang, 2005). Examples of such analyses include comparisons among enteric bacteria (McClelland *et al.*, 2000) and between mouse and human (International Mouse Genome Sequencing Consortium, 2002). Comparative genomics is thus an attempt to take advantage of the information provided by the signatures of selection to understand the function and evolutionary processes that act on genomes.

But comparative genomics can also take a medium-resolution view. By identifying all the known genes from one genome and finding their matching genes, if they exist, in another genome, we can determine which genes have been conserved between species and which are unique. The DNA sequences encoding the proteins and RNA responsible for the functions shared between distantly related organisms, as well as the DNA sequences for controlling the expression of such genes, should be preserved in their genome sequences. Conversely, sequences that encode proteins or RNAs responsible for differences between species will themselves be divergent. For example, a comparison of the genomes of yeast, worms and flies revealed that these eukaryotes encode many of the same proteins, but different gene families are expanded in each genome (Rubin *et al.*, 2000). A similar observation was made in a comparison of sixteen complete archaeal genomes, where comparative genomics

revealed a core of 313 genes that are represented in all sequenced archaeal genomes, plus a variable 'shell' that is prone to lineage-specific gene loss and horizontal gene exchange (Makarova and Koonin, 2003).

A number of software tools have been developed for use in comparative genomics, in order to explore the similarities and differences between genomes at different levels. Because of the volume and nature of the data involved, almost all the visualization tools in this field use a web interface to access large databases of pre-computed sequence comparisons and annotations, e.g. Vista (Frazer *et al.*, 2004), Ensembl (Curwen *et al.*, 2004), UCSC (Hsu *et al.*, 2005). For example, figure 5.4 shows an 8 Mb region of the human chromosome 12, together with homologous regions of other vertebrate genomes, displayed using the UCSC genome browser. This particular region was identified by genome-wide SNP-based mapping in families with mutations involved in Bardet-Biedl Syndrome (BBS), a genetically heterogeneous ciliopathy (Stoetzel *et al.*, 2006).

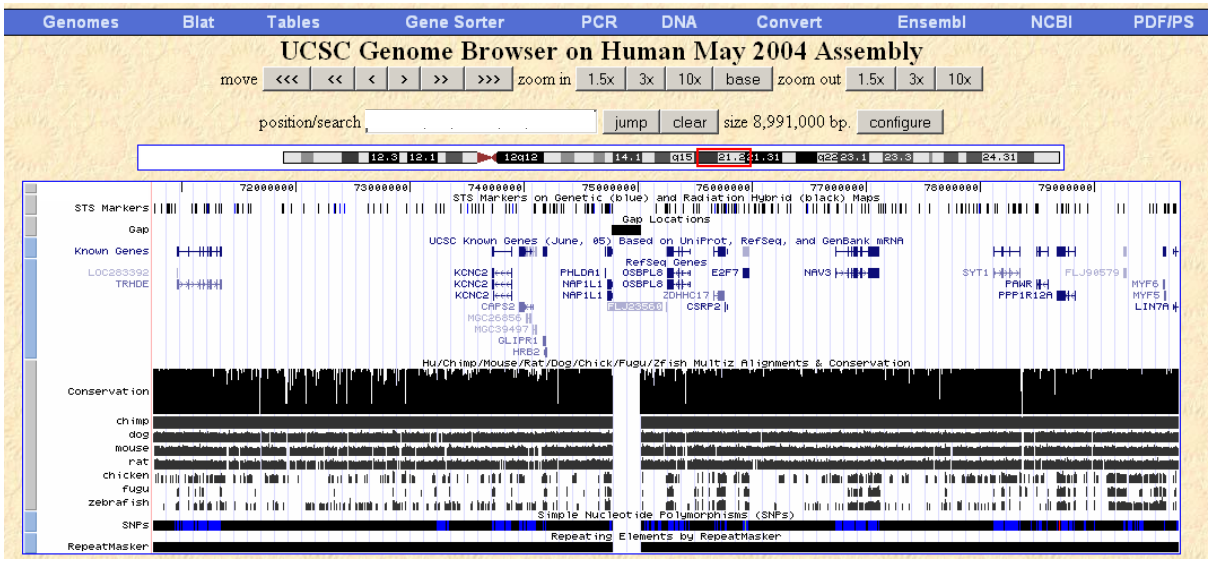


Figure 5.4 UCSC genome browser display
The display shows a 12Mb region of homozygosity that segregated with the disease phenotype in different sibships in families with Bardet-Biedl Syndrome (BBS) mutations. The region contains 23 known genes, including the BBS10 gene, a major locus for BBS. Syntenic regions from chimp, dog, mouse and other organisms are shown at the bottom of the display.

5.2.3 Gene prediction and validation

One important aspect in biotechnology is gene discovery and target validation for drug discovery. At the time of writing, over 1000 genomes (from bacteria, archaea and eukaryota, as well as many viruses and organelles) are either complete or being determined, but biological interpretation, i.e. annotation, is not keeping pace with this avalanche of raw sequence data. There is still a real need for accurate and fast tools to analyze these sequences and, especially, to find genes and determine their functions. Unfortunately, finding genes in a genomic sequence is far from being a trivial problem. It has been estimated that 44% of the protein sequences predicted from eukaryotic genomes and 31% of the HTC (High-throughput cDNA) sequences contain suspicious regions (Bianchetti *et al.*, 2005).

The most widely used approach consists of employing heterogeneous information from different methods, including the detection of a bias in codon usage between coding and non-

coding regions and *ab initio* prediction of functional sites in the DNA sequence, such as splice sites, promoters, or start and stop codons. Most current methods of detection of a signal that may represent the presence of a functional site use position-weight matrices (PWM), consensus sequences or HMM's. The reliability and accuracy of these methods depends critically on the quality of the underlying multiple alignments (for a review, see Mathe *et al.*, 2002). For prokaryotic genomes, these combined methods are highly successful, identifying over 95% of the genes (e.g. Aggarwal and Ramaswamy, 2002), although the exact determination of the start site location remains more problematic because of the absence of relatively strong sequence patterns. The process of predicting genes in higher eukaryotic genomes is complicated by several factors, including complex gene organization, the presence of large numbers of introns and repetitive elements, and the sheer size of the genomic sequence (for a review, see Zhang 2002). It has been shown that comparison of the *ab initio* predicted exons with protein, EST or cDNA databases can improve the sensitivity and specificity of the overall prediction. For example, in the re-annotation of the *Mycoplasma pneumoniae* genome (Dandekar *et al.*, 2000), sequence alignments were used in the prediction of N/C-terminal extensions to the original protein reading frame. This approach has also been implemented in a web server, vALId, developed in our group for automatic protein quality control (Bianchetti *et al.*, 2005).

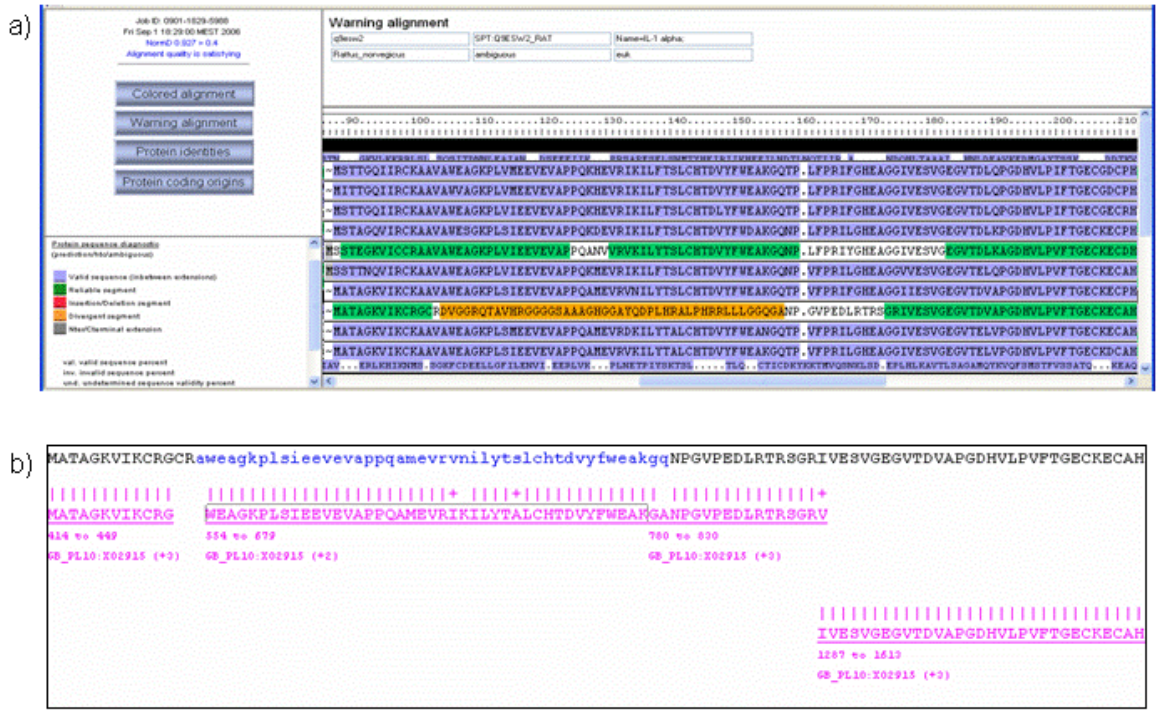


Figure 5.5 vALId display of a multiple alignment of plant alcohol dehydrogenases
a) Multiple alignment display showing reliable sequence segments in green and potential errors in orange. Grey shading represents regions that have been validated by vALId. b) Validation of the predicted error in Q41767_MAIZE by comparison of a chimeric sequence with the original genome sequence.

Taking advantage of high quality MACS, vALId first warns about the presence of suspicious insertions/deletions (indels) and divergent segments, and second, proposes corrections based on transcripts and genome contigs. For example, figure 5.5 shows the vALId analysis of a multiple alignment of plant alcohol dehydrogenases, highlighting a very divergent region in the N-terminal region of the sequence Q41767_MAIZE. Divergent regions are validated by constructing a chimeric sequence, where the suspicious region in the predicted sequence is replaced by the corresponding segment from the closest neighbour in the MACS. A TBlastN search with the chimeric sequence (figure 5.5b) identified an exon

encoding residues that matched the conserved positions in the MACS. The vALId system is described in more detail in section 11.3.1.

5.2.4 Protein function characterisation

In most genome annotation projects, the standard strategy to determine the function of a novel protein is to search the sequence databases for homologues and to propagate the structural/functional annotation from the known to the unknown protein. Recent developments in database search methods have exploited multiple sequence alignments to detect more and more distant homologues e.g. (Altschul *et al.*, 1997: Karplus *et al.*, 1998; Yona and Levitt, 2002). However, most automatic genome projects only use information from the top best hits in the database search, as sequence hits with higher expect values are considered unreliable. This has lead to a certain number of errors in genome annotations. Two types of error have already been identified: those of under- and over-prediction. Under-prediction implies that functional information is not transferred because the chain of propagation is broken, for example, because the top-scoring hits in the database search are all uncharacterised. Over-prediction is perhaps more serious because it introduces incorrect annotations into the sequence databases. Subsequent searches against these databases then cause the errors to propagate to future functional assignments.

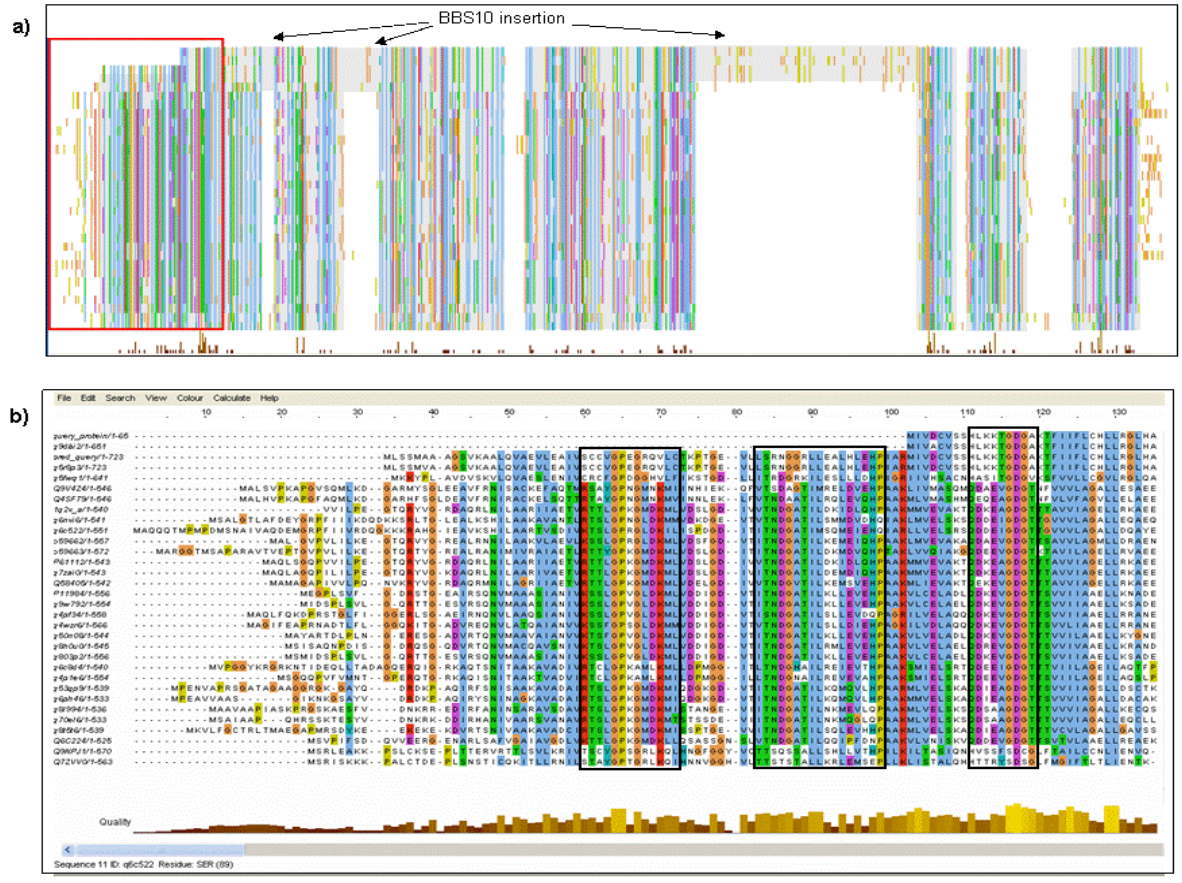


Figure 5.6 Multiple alignment of the BBS10 protein and homologs found in in-depth database searches
a) Overview of complete protein, showing global organisation, including 3 insertions specific to BBS10 and the N-terminal deletion due to an error in the exon prediction of the gene. The red box indicates the region shown in b).Residues are coloured according to the colourrng scheme used in ClustalX (Thompson et al., 1997). b) N-terminal region of the BBS10 alignment. The black boxes indicate the positions of ATP binding site motifs, as defined in the ProSite database.

Another approach is to look for similarities to known domains in pre-compiled databases, such as Interpro (Mulder *et al.*, 2005). These databases contain representations such as profiles or HMM's of individual protein domains based on multiple alignments of known sequences. Genome annotation systems such as Magpie (Gaasterland and Sensen, 1996), Imagene (Medigue *et al.*, 1999), GeneQuiz (Hoersch *et al.*, 2000), Alfresco (Jareborg 2000) now use multiple alignments to reliably incorporate information from more distant homologues and provide a more detailed description of protein function. As an illustration, figure 5.6 shows a MACS of the BBS10 protein (see section 5.2.2). The BBS10 sequence shows some similarity (approx. 11% residue identity) to several chaperonin-like proteins which are found only in vertebrates, although the MACS revealed three BBS10-specific insertions. A 3D homology model based on the crystal structure of the chaperonin from *Thermococcus* (PDB:1q2vA) showed that the 3 insertions are spatially close, suggesting potential interactions and the existence of a new functional domain

5.2.5 Protein 2D/3D structure prediction

Multiple alignments play an important role in a number of aspects of the characterisation of the 3-dimensional structure of a protein. The most accurate *in silico* method for determining the structure of an unknown protein is homology structure modeling. Sequence similarity between proteins usually indicates a structural resemblance, and accurate sequence alignments provide a practical approach for structure modeling, when a 3D structural prototype is available. For models based on distant evolutionary relationships, it has been shown that multiple sequence alignments often improve the accuracy of the structural prediction (Moult *et al.,* 2005). Multiple sequence alignments are also used to significantly increase the accuracy of *ab initio* prediction methods for both 2D (e.g. Lee *et al.*, 2006) and 3D (Al-Lazikani *et al.*, 2001) structures, by taking into account the overall consistency of putative features. Similarly, multiple alignments are also used to improve the reliability of other predictions, such as transmembrane helices (for a review, see Chen *et al.*, 2002). More detailed structural analyses also exploit the information in multiple alignments. For example, binding surfaces common to protein families were defined on the basis of sequence conservation patterns and knowledge of the shared fold (Lichtarge *et al.*, 1996).

More recently, multiple sequence alignments have been used to identify communication pathways through protein folds (Brelivet *et al.*, 2004). Figure 5.7 shows part of a multiple alignment of nuclear receptor (NR) proteins used in this study. Nuclear receptors (NRs) are ligand-dependent transcription factors that control a large number of physiological events through the regulation of gene transcription. Two classes of NRs were identified on the basis of the distribution of differentially conserved residues in the multiple sequence alignment. Differentially conserved residues are defined as those residues that are conserved in one sub-family and that are strictly absent in all the other sequences in the alignment. The two classes of NRs were found to correspond to experimentally verified homodimers and heterodimers. Furthermore, site directed mutagenesis revealed that the differentially conserved residues contribute class-specific communication pathways of salt bridges, confirming the functional importance of these residues for the dimerization process and/or transcriptional activity.

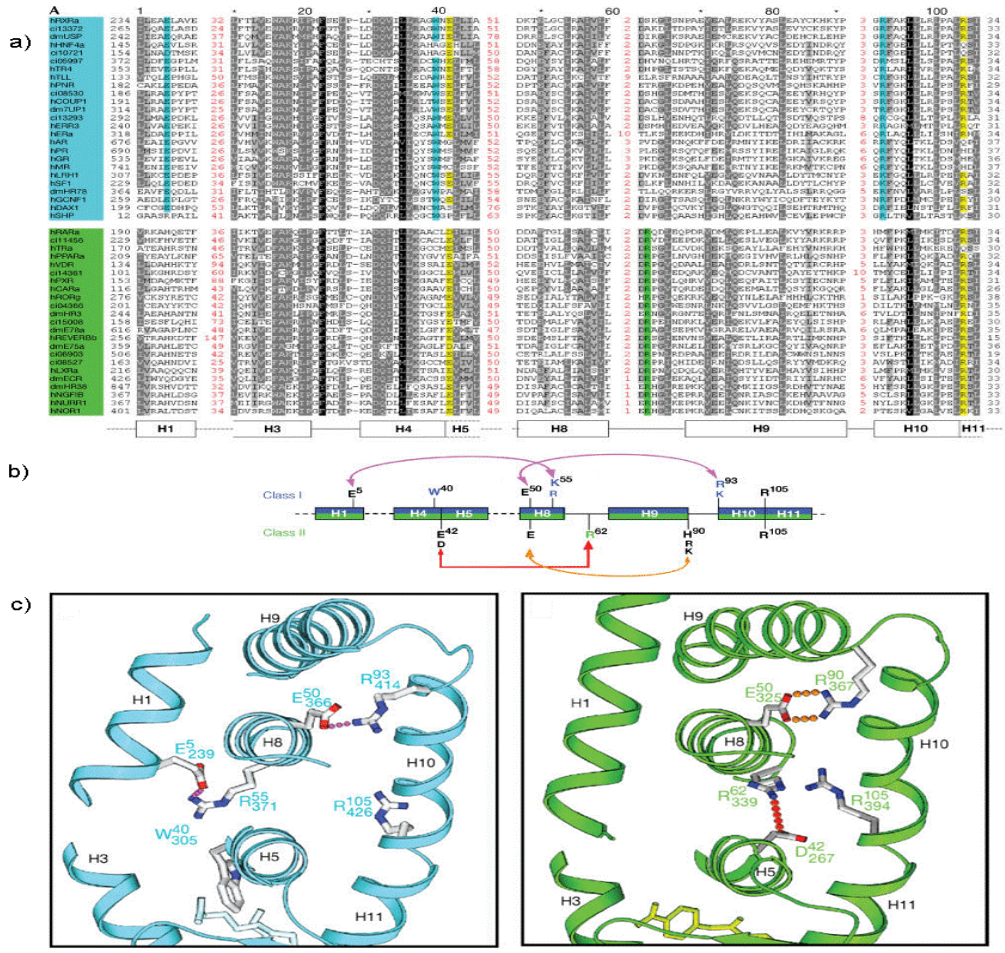


Figure 5.7 Multiple sequence alignment of NR ligand binding domains and class-specific features
a) The differentially conserved residues are highlighted in blue and green for class I and II, respectively. Conserved residues are indicated as follows: 100%, white against black; >80%, white against grey; >60%, black against grey. NR LBD secondary structure elements are indicated. b) Secondary structure diagram showing NR class-specific features. Conserved but not strictly class-specific residues are in black. Arrows indicate the salt bridges in the 3D structures. c) Views of the class-specific residues including the salt bridges forming the class I (blue) and class II (green) communication pathway (from Brelivet et al., 2004).

5.2.6 RNA structure and function

While proteins have been the traditional candidates for detailed structural and functional analyses, RNA secondary and tertiary structure studies remain crucial to the understanding of complex biological systems. Structure and structural transitions are important in many areas, such as post-transcriptional regulation of gene expression, intermolecular interaction and dimerization, splice site recognition and ribosomal frame-shifting. The function of an RNA molecule depends mostly on its tertiary structure and this structure is generally more conserved than the primary sequence. The determination of RNA 3D structure is a limiting step in the study of RNA structure-function relationships because it is very difficult to crystallize and/or get nuclear magnetic resonance spectrum data for large RNA molecules. Currently, a reliable prediction of RNA secondary and tertiary structure from its primary sequence is mainly derived from multiple alignments, searching among members of a family for compensatory base changes that would maintain base-pairedness in equivalent regions. For example, the Sequence to Structure (S2S) tool (Jossinet and Westhof, 2005) proposes a

framework in which a user can display, manipulate and interconnect RNA multiple sequence alignments, secondary and tertiary structures (figure 5.6).

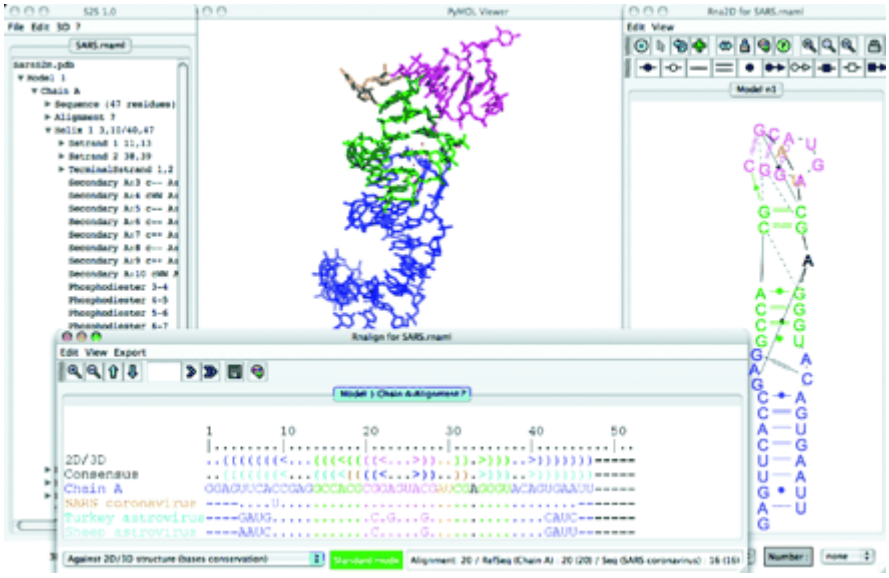


Figure 5.8 S2S display of a multiple alignment of the RNA element conserved in the SARS virus genome
Multiple sequence alignment, secondary and tertiary structure. Inside the multiple alignment, the bracket notation is such that the regular parentheses '('and')' denote the helical Watson–Crick pairs and the '<' and '>' characters specify non-Watson–Crick base-pairs typical of RNA motifs.

These methods have been demonstrated by successful predictions of RNA structures for tRNAs, 5S and 16S rRNAs, RNase P RNAs, small nuclear RNAs (snRNAs) and other RNAs, such as group I introns.

The phylogenetic comparative methods are often supported by complementary, theoretical structure calculations. The most widely used methods are derived from dynamic programming algorithms, such as MFOLD (Zuker, 1989) which predicts on average about 70% of known base-pairs. However, the search for the equilibrium structure by optimization of the global free energy is often insufficient. The biologically functional state of a given molecule may not be the optimal state and moreover, a structured RNA molecule is not a static object. A molecule may pass through a variety of active and inactive states due to the kinetics of folding, to the simultaneity of folding with transcription, or to interactions with extra-molecular factors. To address these problems, integrated systems have been developed that combine traditional thermodynamic calculations with experimental data, e.g. STRUCTURELAB (Shapiro and Kasprzak, 1996). Such systems permit the use of a broad array of approaches for the analysis of the structure of RNA and provide the capability of analysing the data set from a number of different perspectives.

5.2.7 Interaction networks

In the post-genomic view of cellular function, each biological entity is seen in the context of a complex network of interactions. New and powerful experimental techniques, such as the yeast two-hybrid system or tandem-affinity purification and mass spectrometry, are used to determine protein-protein interactions systematically. In parallel with these developments, a number of computational techniques have been designed for predicting protein interactions. The performance of the Rosetta method, which relies on the observation that some interacting

proteins have homologues in another organism fused into a single protein chain, has recently been improved using multiple sequence alignment information and global measures of hydrophobic core formation (Bonneau *et al.*, 2001). A measure of the similarity between phylogenetic trees of protein families has also been used to predict pairs of interacting proteins (Pazos and Valencia, 2001). This method was adapted to consider the multi-domain nature of proteins by breaking the sequence into a set of segments of predetermined size and constructing a separate profile for each segment (Kim and Subramaniam, 2006). Another approach involves quantifying the degree of co-variation between residues from pairs of interacting proteins (correlated mutations), known as the "*in silico* two-hybrid" method. For certain proteins that are known to interact, correlated mutations have been demonstrated to be able to select the correct structural arrangement of two proteins based on the accumulation of signals in the proximity of interacting surfaces (Pazos *et al*., 1997). This relationship between correlated residues and interacting surfaces has been extended to the prediction of interacting protein pairs based on the differential accumulation of correlated mutations between the interacting partners (interprotein correlated mutations) and within the individual proteins (intraprotein correlated mutations) (Pazos and Valencia, 2002).

5.2.8 Genetics

A considerable effort is now underway to relate human phenotypes to variation at the DNA level. Most human genetic variation is represented by single nucleotide polymorphisms (SNPs) and many of them are believed to cause phenotypic differences between individuals (Ramensky *et al.*, 2004). One of the main goals of SNP research is therefore to understand the genetics of human phenotype variation and especially the genetic basis of complex diseases, thus providing a basis for assessing susceptibility to diseases and designing individual therapy. Whereas a large number of SNPs may be functionally neutral, others may have deleterious effects on the regulation or the functional activity of specific gene products. Non-synonymous single-nucleotide polymorphisms (nsSNPs) that lead to an amino acid change in the protein product are of particular interest because they account for nearly half of the known genetic variations related to human inherited disease (Stenson *et al.*, 2003). With more and more data available, it has become imperative to predict the phenotype of a nsSNP *in silico*. Computational tools are therefore being developed, which use structural information or evolutionary information from multiple sequence alignments to predict a nsSNP's phenotypic effect and to identify disease-associated nsSNPs, e.g. (Bao and Cui, 2005).

5.2.9 Drug discovery, design

The structural and functional analyses described above provide an opportunity to identify the proteins associated with a particular disease, that are therefore potential drug targets. Rational drug design strategies can then be directed to accelerate and optimize the drug discovery process using experimental and virtual (computer-aided drug discovery) methods. Recent advances in the computational analyses of enzyme structures and functions have improved the strategies used to modify enzyme specificities and mechanisms by site-directed mutagenesis, and to engineer biocatalysts through molecular reassembly.

For example, vitamin D analogs have been proposed for the treatment of severe rickets caused by mutations in the vitamin D receptor (VDR) gene (Gardezi *et al.*, 2001). The known mutations in the coding regions of the human VDR gene can be divided into two classes, representing two different phenotypes. Mutations in the VDR DNA-binding domain (DBD) prevent the receptor from activating gene transcription, although vitamin D binding is

normal. Patients with this DNA binding-defective phenotype do not respond to vitamin D treatment. In contrast, some patients with mutations in the ligand binding domain (LBD) that cause reduced or complete hormone insensitivity have been partially responsive to high doses of calcium and vitamin D, although this often necessitates long term intravenous infusion therapy. For these patients, an alternative treatment using vitamin D analogs was proposed. Knowledge of the 3D structure of the hormone-occupied VDR LBD (Rochel *et al.*, 2000) and the nature of the amino acid residues that contribute to the functional surface of the receptor allowed the selection of 3 candidate VDR mutations with the potential to interact with the receptor at amino acid contact points that differ from those utilized by the natural ligand, thus restoring the function of mutant VDRs (Gardezi *et al.*, 2001). This example clearly illustrates the importance of polymorphism data that, combined with structural and evolutionary information, can form the basis for biochemical and cellular studies which may eventually lead to new drug therapies.

5.3 Conclusions

Multiple alignments now play a fundamental role in most of the computational methods used in genomic or proteomic projects, ranging from gene identification and the functional characterisation of the gene products to genetics, human health and therapeutics. Since multiple alignments are usually employed at the beginning of the data analysis pipelines, it is crucial that the alignments are of high-quality. Errors in the alignment will lead to further errors in the subsequent analyses and might generate misleading patterns and result in false hypotheses.

Given the pivotal role of multiple alignments, the field has received a lot of attention in recent years. The next chapter will quickly trace the evolution of multiple alignment algorithms from their beginnings in the 1970's to the recent introduction of new integrative and co-operative strategies.

"Emergencies have always been necessary to progress.
It was darkness which produced the lamp.
It was fog that produced the compass.
It was hunger that drove us to exploration."
Victor Hugo (1802 - 1885)

6 Evolution of sequence alignment algorithms

In the face of this growing number of alignment applications, a vast array of diverse algorithms has been developed in an attempt to construct reliable, high-quality multiple alignments within a reasonable time limit that will allow high-throughput processing of large sequence sets.

There exist two main categories of sequence alignment: pairwise alignment (or the alignment of two sequences) and multiple alignment. Pairwise alignments are most commonly used in database search programs such as Fasta (Pearson and Lipman, 1988) and Blast (Altschul *et al.*, 1990) in order to detect homologues of a novel sequence. Multiple alignments, containing from three to several hundred sequences, are more computationally complex than pairwise alignments and in general, simultaneous alignment of more than a few sequences is rarely attempted. Instead a series of pairwise alignments are performed and amalgamated into a multiple alignment. The purpose of any sequence alignment, whether pairwise or multiple, is to show how a set of sequences may be related, in terms of conserved residues, substitutions, insertion and deletion events (described in section 5.1.1).

This chapter describes the development of sequence alignment methods, from the first algorithms for the alignment of two sequences (sections 6.1 and 6.2), via the traditional progressive method for the efficient construction of multiple alignments to the recent introduction of co-operative strategies that combine complementary algorithms or information other than the sequence itself (section 6.3). Finally, section 6.4 contains a brief discussion of the issues related to user access and visualisation of multiple sequence alignments.

6.1 Pairwise alignment scoring and statistics

For any two sequences, there are an exponential number of potential alignments with gaps. Therefore, it is critical to be able to distinguish 'good' alignments from bad ones. A good alignment is one that corresponds to the biologically correct alignment, accurately reflecting the evolutionary, structural and functional relationships between the sequences. Sequence alignment programs have, until recently, used only the primary sequence information to reconstruct these complex relationships. In order to find the best alignment, most alignment programs assign a similarity score to all possible alignments and try to maximize this score. These alignment scores, also known as objective functions, are generally based on scores for aligning single residues with penalties for introducing indels into the sequences.

6.1.1 Scoring matrices

Most alignment programs make comparisons between pairs of bases or amino acids by looking up a value in a scoring matrix. The matrix contains a score for the match quality of every possible pair of residues (figure 6.1).

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X
A	2	-2	0	0	-2	0	0	1	-1	-1	-2	-1	-1	-3	1	1	1	-6	-3	0	0	0	0
R	-2	6	0	-1	-4	1	-1	-3	2	-2	-3	3	0	-4	0	0	-1	2	-4	-2	-1	0	-1
N	0	0	2	2	-4	1	1	0	2	-2	-3	1	-2	-3	0	1	0	-4	-2	-2	2	1	0
D	0	-1	2	4	-5	2	3	1	1	-2	-4	0	-3	-6	-1	0	0	-7	-4	-2	3	3	-1
C	-2	-4	-4	-5	12	-5	-5	-3	-3	-2	-6	-5	-5	-4	-3	0	-2	-8	0	-2	-4	-5	-3
Q	0	1	1	2	-5	4	2	-1	3	-2	-2	1	-1	-5	0	-1	-1	-5	-4	-2	1	3	-1
E	0	-1	1	3	-5	2	4	0	1	-2	-3	0	-2	-5	-1	0	0	-7	-4	-2	3	3	-1
G	1	-3	0	1	-3	-1	0	5	-2	-3	-4	-2	-3	-5	0	1	0	-7	-5	-1	0	0	-1
H	-1	2	2	1	-3	3	1	-2	6	-2	-2	0	-2	-2	0	-1	-1	-3	0	-2	1	2	-1
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5	2	-2	2	1	-2	-1	0	-5	-1	4	-2	-2	-1
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6	-3	4	2	-3	-3	-2	-2	-1	2	-3	-3	-1
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5	0	-5	-1	0	0	-3	-4	-2	1	0	-1
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6	0	-2	-2	-1	-4	-2	2	-2	-2	-1
F	-3	-4	-3	-6	-4	-5	-5	-5	-2	1	2	-5	0	9	-5	-3	-3	0	7	-1	-4	-5	-2
P	1	0	0	-1	-3	0	-1	0	0	-2	-3	-1	-2	-5	6	1	0	-6	-5	-1	-1	0	-1
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2	1	-2	-3	-1	0	0	0
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3	-5	-3	0	0	-1	0
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17	0	-6	-5	-6	-4
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	-2	-3	-4	-2
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4	-2	-2	-1
B	0	-1	2	3	-4	1	3	0	1	-2	-3	1	-2	-4	-1	0	0	-5	-3	-2	3	2	-1
Z	0	0	1	3	-5	3	3	0	2	-2	-3	0	-2	-5	0	0	-1	-6	-4	-2	2	3	-1
X	0	-1	0	-1	-3	-1	-1	-1	-1	-1	-1	-1	-1	-2	-1	0	0	-4	-2	-1	-1	-1	-1

Figure 6.1 PAM-250 matrix.
Substitution scores for amino acids.

The simplest way to score an alignment is to count the number of identical residues that are aligned. When the sequences to be aligned are closely related, this will usually find approximately the correct solution. For more divergent sequences sharing less than 25-30 percent identity, however, the scores given to non-identical residues becomes critically important. More sophisticated scoring schemes exist for both DNA and protein sequences and generally take the form of a matrix defining the score for aligning each pair of residues. For alignments of nucleotide sequences, the simplest scoring matrix would assign the same score to a match of the four classes of bases, ACGT, and 0 for any mismatch. However, transitions (substitution of A-G or C-T) happen much more frequently than transversions (substitution of A-T or G-C) and it is often desirable to score these substitutions differently. More complex matrices also exist in which matches between ambiguous nucleotides are given values whenever there is any overlap in the sets of nucleotides represented by the two symbols being compared. For protein sequence comparisons, scoring matrices generally take into account the biochemical similarities between residues and/or the relative frequencies with which each amino acid is substituted by another. The most widely used scoring matrices are known as the PAM (point accepted mutation) matrices (Dayhoff *et al.,* 1978). The original PAM1 matrix was constructed based on the mutations observed in a large number of alignments of closely related sequences. A series of matrices was then extrapolated from the PAM1. The matrices range from strict ones, useful for comparing very closely related sequences to very 'soft' ones that are used to compare very divergent sequences. For example, the PAM250 matrix corresponds to an evolutionary distance of 250%, or approximately 80% residue divergence. Other matrices have been derived directly from either sequence-based or structure-based alignments. For example, the Blosum matrices are based on the observed residue substitutions in aligned sequence segments from the Blocks database. The proteins in the database are clustered at different percent identities to produce a series of matrices. For example, the Blosum-62 matrix is based on alignment blocks in which all the sequences share at least 62% residue identity. Other more specialized matrices have been developed e.g. for specific secondary structure elements (e.g. Luthy *et al*., 1991) or for the comparison of particular types of proteins such as transmembrane proteins (e.g. Ng *et al*., 2000).

6.1.2 Gap schemes

As well as assigning scores for residue matches and mismatches, most alignment scoring schemes in use today calculate a cost for the insertion of gaps in the sequences. One of the

first gap scoring schemes for the alignment of two sequences charged a fixed penalty for each residue in either sequence aligned with a gap in the other. Under this system, the cost of a gap is proportional to its length. Alignment algorithms implementing such length-proportional gap penalties are efficient, however the resulting alignments often contain a large number of short indels that are not biologically meaningful. To address this problem, linear or 'affine' gap costs are used that define a gap insertion or 'gap opening' penalty in addition to the length-dependent or 'gap extension' penalty. Thus, a smaller number of long gaps is favoured over many short ones. Fortunately, algorithms using affine gap costs are only slightly more complex than those using length-proportional gap penalties, requiring only a constant factor more space and time. Again, more complex schemes have been developed, such as 'concave' gap costs (e.g. Benner *et al.,* 1993) or position-specific gap penalties (e.g. Thompson, 1995). Most of these are attempts to mimic the biological processes or constraints that are thought to regulate the evolution of DNA or protein sequences.

6.1.3 Alignment statistics

An important aspect of sequence alignment is to establish how meaningful a given alignment is. It is always possible to construct an alignment between a set of sequences, even if they are unrelated. The problem is to determine the level of similarity required to infer that the sequences are homologous, *i.e*. that they descend from a common ancestor. A simple rule-of-thumb for protein sequences states that if two sequences share more than 25% identity over more than 100 residues, then the two sequences can be assumed to be homologous. However, many proteins sharing less than 25% residue identity, said to be in the 'twilight zone' (Doolittle, 1986), do still have very similar structures. The measure of the percent identity or similarity of the sequences is generally not sensitive enough to distinguish between alignments of related and unrelated sequences. Much work has been done on the significance of both ungapped and gapped pairwise local alignments (Altschul and Gish, 1996; Pearson, 1998), although the statistics of global alignments or alignments of more than two sequences are far less well understood. The aim of the statistical analysis is to estimate the probability of finding by 'chance' at least one alignment that scores as high as or greater than the given alignment. For ungapped local alignments, these probabilities or P-values may be derived analytically. For alignments with gaps, empirical estimates are used based on the scores obtained during a database search, or from randomly generated sequences. For database search programs, the significance of an alignment between the query sequence and a database sequence is often expressed in terms of Expect- or E-values. The E-value specifies the number of matches with a given score that are expected to occur by chance in a search of a database. An Expect-value of zero, with a given score, would indicate that no matches with this score are expected purely by chance.

6.2 Pairwise alignments

6.2.1 Optimal alignment

The comparison or alignment of biological sequences began in the early seventies, with the first dynamic programming algorithm for the global (or full-length) alignment of two sequences (Needleman and Wunsch, 1971). This recursive algorithm for the alignment of two sequences $X=x_1,\ldots,x_n$ and $Y=y_1,\ldots y_m$ may be summarised as follows:

$$H_{i,j} = MAX\begin{Bmatrix} H_{i-1,j-1} + S_{i,j} \\ MAX\left(H_{i-k,j} - g - hk\right) \\ MAX\left(H_{i,j-1} - g - hl\right) \end{Bmatrix}$$

where $S_{i,j}$ is the score for aligning residues x_i and y_j,
$H_{i,j}$ is the score of the optimal alignment of subsequences $x_1,...x_i$ and $y_1,\ldots y_j$
g is the penalty for opening a gap
h is the penalty for extending a gap by one residue
k,l are the lengths of the gaps in sequences X and Y respectively

The optimal local alignment between a pair of sequences, in which only the highest-scoring sub-segments of the two sequences are aligned, involves a simple modification to the Needleman-Wunsch method (Smith and Waterman, 1981). The additional constraint, Hij >= 0, is included in the recursive algorithm, such that the alignment can start or end at any pair of residues.

Dynamic programming is a rigorous mathematical technique that is guaranteed to find the maximal scoring alignment for any two sequences. It does this by constructing a two-dimensional alignment matrix or path graph of partial alignment scores (figure 6.2).

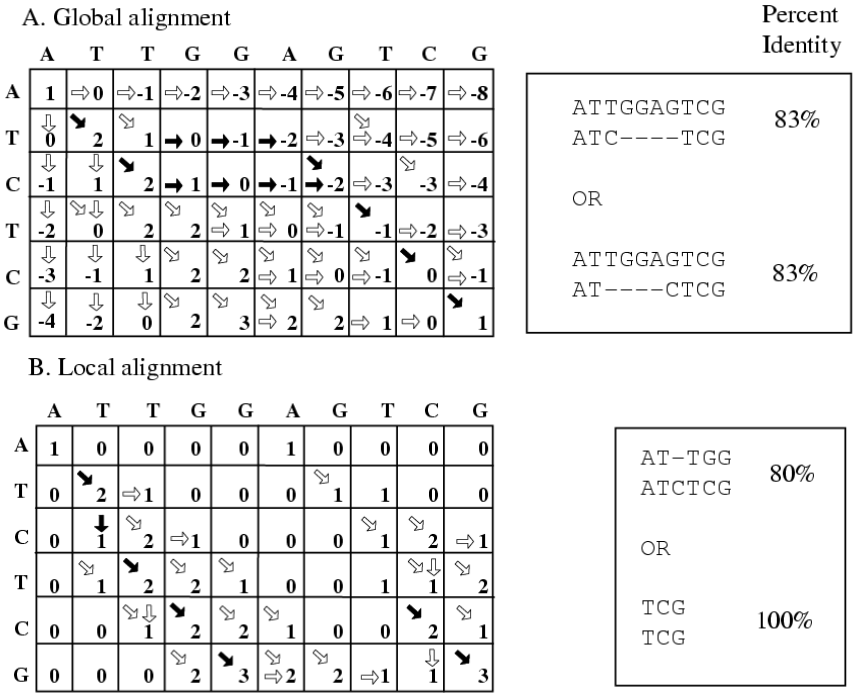


Figure 6.2 Dynamic programming matrices for global and local alignments of two DNA sequences.
Percent identity scores for each alignment are calculated by dividing the number of identical residues aligned by the total number of residues aligned.

Each position in the matrix contains the score for the best partial alignment that ends at that position. The best scoring partial alignment will be extended to subsequent positions in the matrix by either aligning one residue from each sequence or by inserting a gap into one or other of the sequences. In this way all possible alignments are considered and the final

alignment is thus the best scoring alignment possible. The optimal global alignment score is given in the bottom, right-hand corner of the alignment matrix, while the optimal local alignment score is defined as the highest scoring position anywhere in the alignment matrix.

6.2.2 Dot plots

A dot plot is a visual method for comparing two complete sequences that provides a global view of all possible regions of similarity between the two sequences. (For a detailed review, see States and Boguski 1991). Dot plot programs often provide an interactive environment in which the user can select significant sequence segments in order to guide the final alignment. In the dot plot in figure 6.3, the X and Y axes of the plot correspond to the two sequences to be compared.

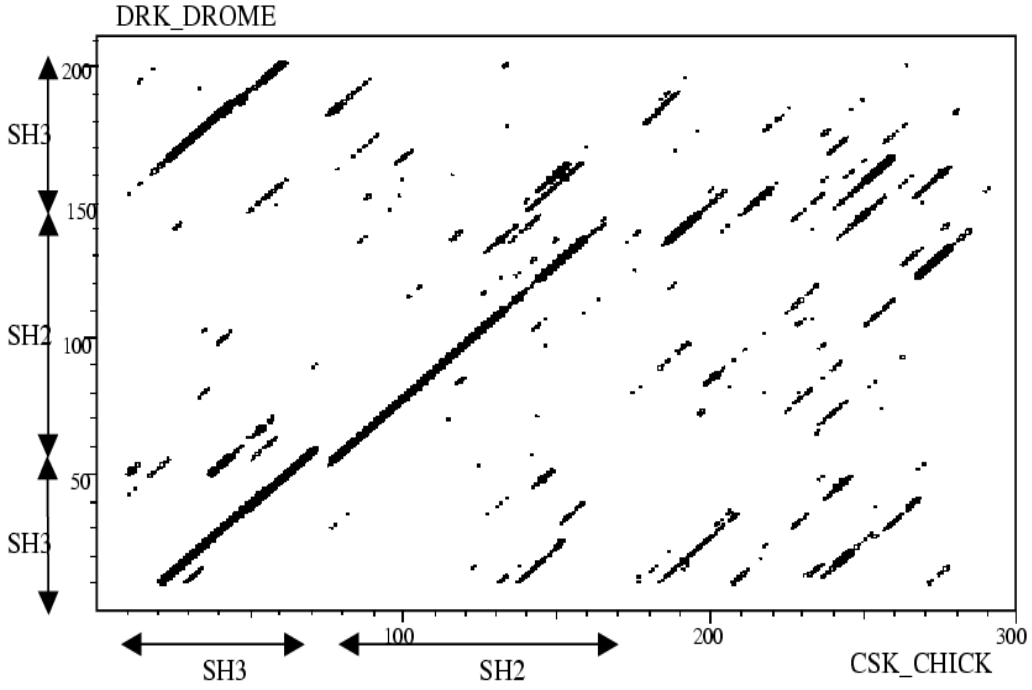


Figure 6.3 Dot plot of a tyrosine-protein kinase protein compared to a SH2-SH3 adaptor protein
On the x-axis: chicken tyrosine-protein kinase (CSK_CHICK) and on the y-axis: drosophila SH2-SH3 adaptor protein (DRK_DROME)

The dots represent all the possible matches of identical residues in the two sequences. Any region of similar sequence appears as a diagonal row of dots. Isolated dots not on the diagonal represent random matches, which are probably not related to any significant alignment. Visualization of matching regions may be improved by filtering out these random matches using a sliding window calculation. Instead of comparing single sequence positions in the two sequences, the average score in a window of adjacent positions is calculated, and a dot is printed only if the score for the window is above a certain average score. Scoring matrices such as the PAM or Blosum matrices may be used instead of residue identities. Dot plots are particularly valuable for finding repeats or inversions in protein and DNA sequences, and for predicting regions in RNA that are self complementary and that, therefore, might form a double-stranded region or secondary structure.

6.2.3 Heuristic methods

A different approach to the local alignment problem involves the use of heuristics or 'approximate' methods, which do not guarantee an optimal alignment solution but are less time-consuming than the rigorous dynamic programming techniques. These approximate alignment algorithms are used in programs such as Fasta (Pearson and Lipman, 1988) and Blast (Altschul *et al.*, 1990) to search the nucleic acid and protein sequence databases for homologues of a target sequence. The general approach involves comparing the target or 'query' sequence to all the sequences in a specified database in a pairwise fashion. Each comparison is given a score reflecting the degree of similarity between the query and the sequence being compared. The higher the score, the greater the degree of similarity. The similarity is measured and shown by aligning the two sequences. The heuristics used involve finding patches of regional similarity, rather than trying to find the best alignment between the entire query and an entire database sequence. Fasta uses a two step pairwise alignment algorithm. The first step consists of a search for exactly matching strings or 'words' that are common to both sequences. This is done in order to identify regions in a two dimensional table similar to that shown for the dynamic programming algorithm above that are likely to correspond to highly similar segments shared by the two sequences. These regions will consist of a diagonal or a few closely spaced diagonals in the table which have a high number of word matches between the sequences. The second step involves a Smith-Waterman local alignment centered on these regions. The speed up achieved by a Fasta alignment relative to a full Smith-Waterman alignment is due to the restriction of the dynamic programming algorithm to only the high-scoring regions. The Blast program works by first making a look-up table of all the short subsequences, known as 'words' and neighboring words, i.e., similar words in the query sequence. The sequence database is then scanned for these matching segments and the high scoring segments found are extended in both forward and backward directions to generate an alignment that continues until the sequence ends, or the alignment becomes non-significant. In both Fasta and Blast, in addition to the alignment scores, the significance of each alignment is computed as a P value or an E value (see section on Alignment Statistics), based on the alignment scores expected by chance in the total sequence space.

6.3 Multiple sequence alignment

The first formal algorithm for multiple sequence alignment (Sankoff, 1975) was developed as a direct extension of the pairwise dynamic programming algorithm. However, the optimal multiple alignment of more than a few sequences (more than 10) remains impractical due to the intensive computer resources required, despite some space and time improvements (e.g. Lipman *et al.*, 1989). Therefore, in order to multiply align larger sets of sequences, most programs in use today employ some kind of heuristic approach to reduce the problem to a reasonable size.

6.3.1 Progressive multiple alignment

Traditionally the most popular method has been the progressive alignment procedure (Feng and Doolittle, 1987), 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. An example using five immunoglobulin-like domains is shown in figure 6.4.

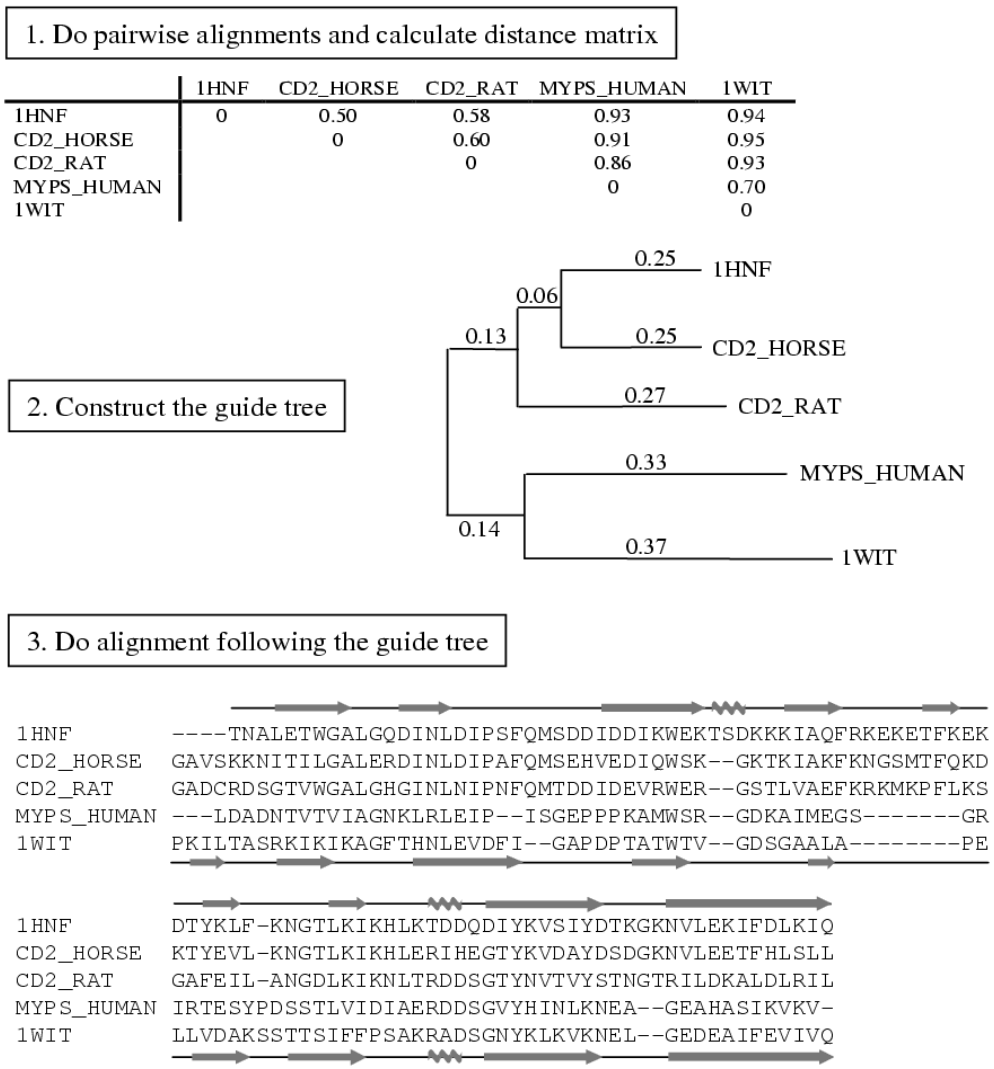


Figure 6.4 The basic progressive alignment procedure
The algorithm is illustrated using a set of five immunoglobulin-like domains. The sequence names are from the Swissprot or PDB databases: 1HNF: human cell adhesion (CD2) protein, CD2_HORSE: horse cell adhesion protein, CD2_RAT: rat cell adhesion protein, MYPS_HUMAN: human myosin-binding protein, 1WIT: nematode twitchin muscle protein. The secondary structure elements of the immunoglobulin-like domains from the human CD2 (1HNF) and the nematode twitchin (1WIT) proteins are shown above and below the alignment (right arrow = beta sheet, coil = alpha helix).

The first step involves aligning all possible pairs of sequences in order to determine the distances between them. A guide tree is then created and is used to determine the order of the multiple alignment. The two closest sequences are aligned first and then larger and larger sets of sequences are merged, until all the sequences are included in the multiple alignment. In the example, the human and horse CD2 sequences are aligned first. These two sequences are then aligned with the rat CD2 sequence. Finally, the myosin-binding protein sequence is aligned with the twitchin sequence, before being merged with the alignment of the three CD2 sequences. This procedure works well when the sequences to be aligned are of different degrees of divergence. Pairwise alignment of closely related sequences can be performed very accurately. By the time the more distantly related sequences are aligned, important

information about the variability at each position is available from those sequences already aligned. A number of different alignment programs based on this method exist, using either a global alignment method to construct an alignment of the complete sequences, or a local algorithm to align only the most conserved subsegments of the sequences (figure 6.5). For example, Multalign (Barton and Sternberg, 1987), Multal (Taylor 1988), Pileup (Wisconsin Package, Genetics Computer Group, Madison, WI), ClustalW/X (Thompson *et al.*, 1994; Thompson *et al.*, 1997) are all based on the global Needleman-Wunsch algorithm. The main difference between these programs lies in the algorithm used to determine the final order of alignment. For example, Multal uses a sequential branching algorithm to identify the two closest sequences first and subsequently align the next closest sequence to those already aligned. Multalign and Pileup use a simple bottom-up data clustering method, known as the Unweighted Pair Grouping Method with Arithmetic means (UPGMA) (Sneath and Sokal, 1973), to construct a phylogenetic tree that is then used to guide the progressive alignment step. ClustalW/X uses another phylogenetic tree construction method, called neighbour-joining (NJ) (Saitou and Nei, 1987). Although the NJ method is less efficient than the UPGMA, it has been extensively tested and usually finds a tree that is quite close to the optimal tree. In contrast to the global alignment methods, the Pima program (Smith and Smith, 1992) uses the Smith-Waterman algorithm to find a local multiple alignment.

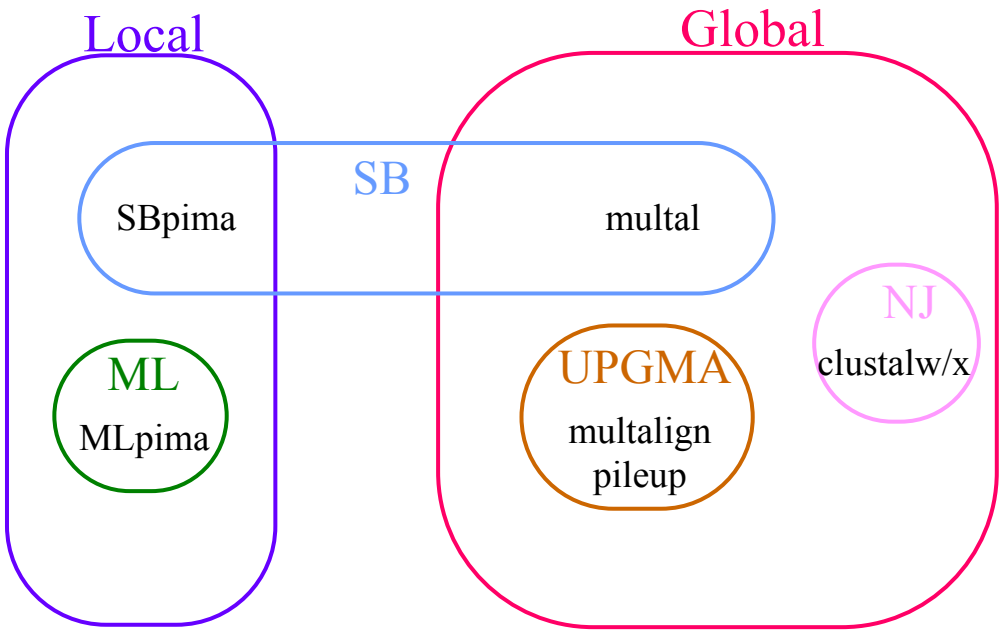


Figure 6.5 Overview of different progressive alignment algorithms
SB=sequential branching, ML=maximum likelihood, NJ=neighbour joining, UPGMA=Unweighted Pair Grouping Method with Arithmetic mean.

Since then, the sensitivity of the progressive multiple sequence alignment method has been somewhat improved with the introduction of several important enhancements to the basic method. For example, Treealign (Hein, 1990) extends the progressive alignment process by adding a parsimony step: an initial alignment is constructed and used to build a parsimony tree which in turn is used to direct the final alignment algorithm. ClustalX (Thompson *et al.*, 1997) reduces the problem of the over-representation of certain sequences by incorporating a sequence weighting scheme that downweights near-duplicate sequences and upweights the most divergent ones. In addition, position-specific gap penalties encourage the alignment of new gaps on existing gaps introduced earlier in the multiple alignment. Most of the alignment programs mentioned above use one residue scoring matrix and two gap

penalties (one for opening a new gap and one for extending an existing gap). When identities dominate an alignment, almost any set of parameters will find approximately the correct solution. With very divergent sequences, however, the scores given to non-identical residues will become critically important. Also, the exact values of the gap penalties become important for success. Thus, the choice of alignment parameters remains a decisive factor affecting the quality of the final alignment.

6.3.2 Iterative strategies

The next generation of multiple alignment algorithms used iterative strategies to refine and improve the initial alignment. The PSI-Blast program builds multiple alignments by aligning the homologous segments detected by a Blast database search to the query sequence. Hidden Markov Models (HMM's) have been used in a number of programs HMMT (Eddy 1998) or SAM (Karplus *et al.*, 1998) to build multiple alignments and have been employed notably to create large reference databases of sequence alignments such as Pfam and ProSite. The flexibility and efficiency of stochastic techniques such as Gibbs Sampling (Lawrence *et al.*, 1993) and Genetic Algorithms (Notredame and Higgins, 1996) have also been exploited in the search for more accurate alignments. Iteration techniques have also been used to refine an initial multiple alignment built using the traditional progressive alignment algorithm in PRRP (Gotoh, 1996). An alternative to the global alignment approach is the 'segment-to-segment' alignment method used in Dialign (Morgenstein *et al.*, 1996). Segments consisting of locally conserved residue patterns or motifs, rather than individual residues, are detected and then combined to construct a local multiple alignment of only the most conserved regions of the sequences.

6.3.3 Co-operative strategies

The complexity of the multiple alignment problem has lead to the combination of different alignment algorithms and the incorporation of biological information other than the sequence itself. A comparison of a number of local and global protein alignment methods based on the BAliBASE benchmark (Thompson *et al.*, 1999a) showed that no single algorithm was capable of constructing accurate alignments for all test cases. A similar observation was made in another study of RNA alignment programs (Gardner *et al.*, 2005), where algorithms incorporating structural information outperformed pure sequence-based methods for divergent sequences. Therefore, recent developments in multiple alignment methods have tended towards an integrated system bringing together knowledge-based or text-mining systems and prediction methods with their inherent unreliability. Some of the most widely used or more innovative methods are described below:

- *DbClustal* (Thompson *et al.*, 2000) exploits information available in the public databases to improve the accuracy of global multiple alignments. Conserved motifs are extracted from the top sequences detected by a BlastP database search (Altschul *et al.*, 1997) using the Ballast program (Plewniak *et al.*, 2000). This local information is incorporated into a ClustalW global alignment in the form of a list of anchor points between pairs of sequences.

- *T-Coffee* (Notredame *et al.*, 2000) uses information from a pre-compiled library of different pairwise alignments including local, global or structural alignments. This strategy has been extended recently to combine alternative multiple alignments (Wallace *et al.*, 2006).

- *MAFFT* (Katoh *et al.*, 2002) and *MUSCLE* (Edgar, 2004) are efficient methods, that include fast pairwise alignments, using a fast Fourier transform (for MAFFT) or using k-

mer counting (for MUSCLE), together with a progressive multiple alignment method and iterative refinement.

- *PMComp* (Hofacker *et al*., 2004) aligns RNA sequences by first computing base pairing probability matrices and then aligning the common secondary structure in order to deduce a multiple sequence alignment.
- *Praline* (Simosis and Heringa, 2005) exploits protein secondary structure information either from 3D structures or from computational predictions to increase alignment sensitivity.
- *POA* (Lee *et al.*, 2002b) and *RAlign* (Sammeth and Heringa, 2006) use local algorithms that are suitable for multi-domain proteins, that may contain repeated or shuffled elements.
- *Rascal* (Thompson *et al.*, 2003) is an alternative, knowledge-based program designed to improve an existing multiple alignment constructed using any of the above methods. It uses information from clustering algorithms (Wicker *et al.*, 2001; Wicker *et al.*, 2002) and residue conservation analysis (Thompson *et al.*, 2001) in a two-step refinement process to detect and correct local alignment errors.
- *Probcons* (Do *et al.*, 2005) uses HMM-derived posterior probabilities and three-way alignment consistency in a global, progressive alignment, together with an iterative refinement step.

6.4 User access and visualisation

Thanks to the recent developments in multiple alignment algorithms, it is now possible to build accurate and reliable multiple alignments of large sequence sets, with the throughput time required by large scale projects. A crucial factor is the ease-of-use of the new complex systems software currently being developed. Some software is difficult to operate for biologists with limited computer training. Programs that have non-graphical, command-line driven interfaces are not intuitive because they require the use of exact command syntax, including all possible options. In contrast, graphical interfaces such as Modview (Ilyin *et al.*, 2003), Jalview (Clamp *et al.*, 2004) or VISSA (Li and Godzik., 2006) allow visualization of multiple protein sequences and structures with highlighting of features such as conserved residues, active sites, fragments or domains. In addition, some programs are designed to run on specific platforms with specific operating systems (e.g. Unix). Users who are not familiar with an operating system may have difficulty in installing and using these programs. One solution is to use a Web interface, which allows the user to access data files as well as analysis programs in an integrated fashion regardless of client platforms. One example is W2H (Senger *et al*., 1998), a Web-based interface to the popular GCG Sequence Analysis Software Package (Wisconsin Package). Some systems running a number of different automatic bioinformatics analyses e.g. Pfaat (Johnson *et al.*, 2003) also allow expert knowledge to be manually incorporated in the results.

"Quality is never an accident;
it is always the result of
intelligent effort."
John Ruskin (1819 - 1900)

7 Multiple alignment quality

Since the introduction of the first sequence comparison methods in the 1970s, a vast number of alignment methods have been developed that use very different algorithms, ranging from traditional optimal dynamic programming or progressive alignment strategies to the application of algorithms such as simulated annealing, Hidden Markov Models or genetic algorithms. Since the year 2000, the new challenges posed by the post-genomic era have lead to an explosion of new methods. In the search for more accurate alignments, most state of the art methods now often use a combination of complementary techniques, such as local/global alignments or sequence/structure information. Although much progress has been achieved, the latest methods are not perfect and misalignments can still occur. If these misalignments are not detected, they will lead to further errors in the subsequent applications that are based on the multiple alignment (see Chapter 5). The assessment of the quality and significance of a multiple alignment has therefore become a critical task, particularly in high-throughput data processing systems, where a manual verification of the results is no longer possible.

A number of quality issues can be distinguished. First, given a set of sequences, how to evaluate the quality of a multiple alignment of those sequences. The most reliable is probably to compare the alignment to a reference alignment, e.g. 3D structural superposition. In the absence of a known reference, a score is calculated, known as an objective function that estimates how close the alignment is to the correct or optimal solution. Objective scoring functions are discussed in section 7.1. In general though, most multiple alignments contain regions that are well aligned and regions that contain errors. Section 7.2 describes methods that can distinguish reliable from unreliable regions. Even if the alignment is optimal, this does not mean that the sequences are actually homologous. Most multiple alignment methods available today will produce an alignment even if the sequences are unrelated. Section 7.3 discusses methods to detect unrelated sequences. Finally, section 7.4 describes the most widely used benchmarks that are used to compare multiple alignment methods and evaluate the improvements obtained by the new methods.

7.1 Multiple alignment objective scoring functions

Given a particular set of sequences, an objective score is needed that describes the optimal or "biologically correct" multiple alignment. Sub-optimal or incorrect alignments would then score less than this maximal score. Such measures, also known as objective functions, are currently used to evaluate and compare multiple alignments from different sources and to detect low-quality alignments. They are also used in iterative alignment methods to improve the alignment by seeking to maximize the objective function.

One of the first scoring systems was the Sum-of-Pairs score (Carrillo and Lipman, 1988). For each pair of sequences in the multiple alignment a score is calculated based on the percent identity or the similarity between the sequences. (Pairwise alignment scores are discussed in detail in Chapter 6). The score for the multiple alignment, S(m), is then taken to be the sum of all the pairwise scores:

$$S(m) = \sum_{i<j, j<N} s(i, j)$$

where *s(i, j)* is the score of the pairwise alignment between sequences *i* and *j* and N is the total number of sequences in the alignment

Pairwise scores are also used in the COFFEE objective function (Notredame *et al.*, 1998), which reflects the level of consistency between a multiple sequence alignment and a library containing pairwise alignments of the same sequences. This method was shown to be a good estimation of the accuracy of the multiple alignment when high quality pairwise alignments, such as 3D structural superpositions, are available as reference. One problem with multiple alignment scores based on pairwise sequence comparisons is that they assume that substitution probabilities are uniform and time-invariant at all positions in the alignment. This is unrealistic as the variability may range from total invariance at some positions to complete variability at others, depending on the functional or structural constraints of the protein.

For this reason, more recent work has concentrated on column statistics. One approach uses an Information Content statistic, assuming that the most interesting alignments are those where the frequencies of the residues found in each column are significantly different from a predefined set of a priori residue probabilities (Hertz and Stormo, 1999). The Information Content of a multiple alignment is defined as:

$$I(m) = \sum_{j=1}^{L} \sum_{i=1}^{A} f_{i,j} \ln \frac{f_{i,j}}{p_i}$$

where $f_{i,j} = \frac{n_{i,j}}{N}$ is the frequency that letter *i* occurs at position *j*

A is the total number of letters in the alphabet
L is the total number of positions in the alignment
N is the total number of sequences in the alignment
p_i is the *a priori* probability of letter i
$n_{i,j}$ is the frequency that letter i occurs at position j

One disadvantage of this measure is that it considers only the frequencies of identical residues in each column and does not take into account similarities between residues. For this reason, another column-based measure, norMD, was introduced (Thompson *et al.*, 2001), based on the Mean Distance (MD) column scores implemented in ClustalX (Thompson *et al.*, 1997). The MD scores are summed over the full-length of the alignment and the total score is then normalized to take into account the number, length and similarity of the sequences in the alignment, and the presence of gaps.

These different objective functions can be evaluated using the multiple alignments in the BAliBASE benchmark database (see below for more details), as shown in figure 7.1 The Sum-of-Pairs score increases proportionally with the number of the sequences in the alignment (figure 7.1A). Thus, an alignment containing many sequences will score higher than an alignment of fewer sequences, regardless of the respective quality. The Information Content measure solves the problem of the number of sequences, as all columns will score between 0 and 1. However, the scores increase proportionally with the length of the alignment (figure 7.1B).

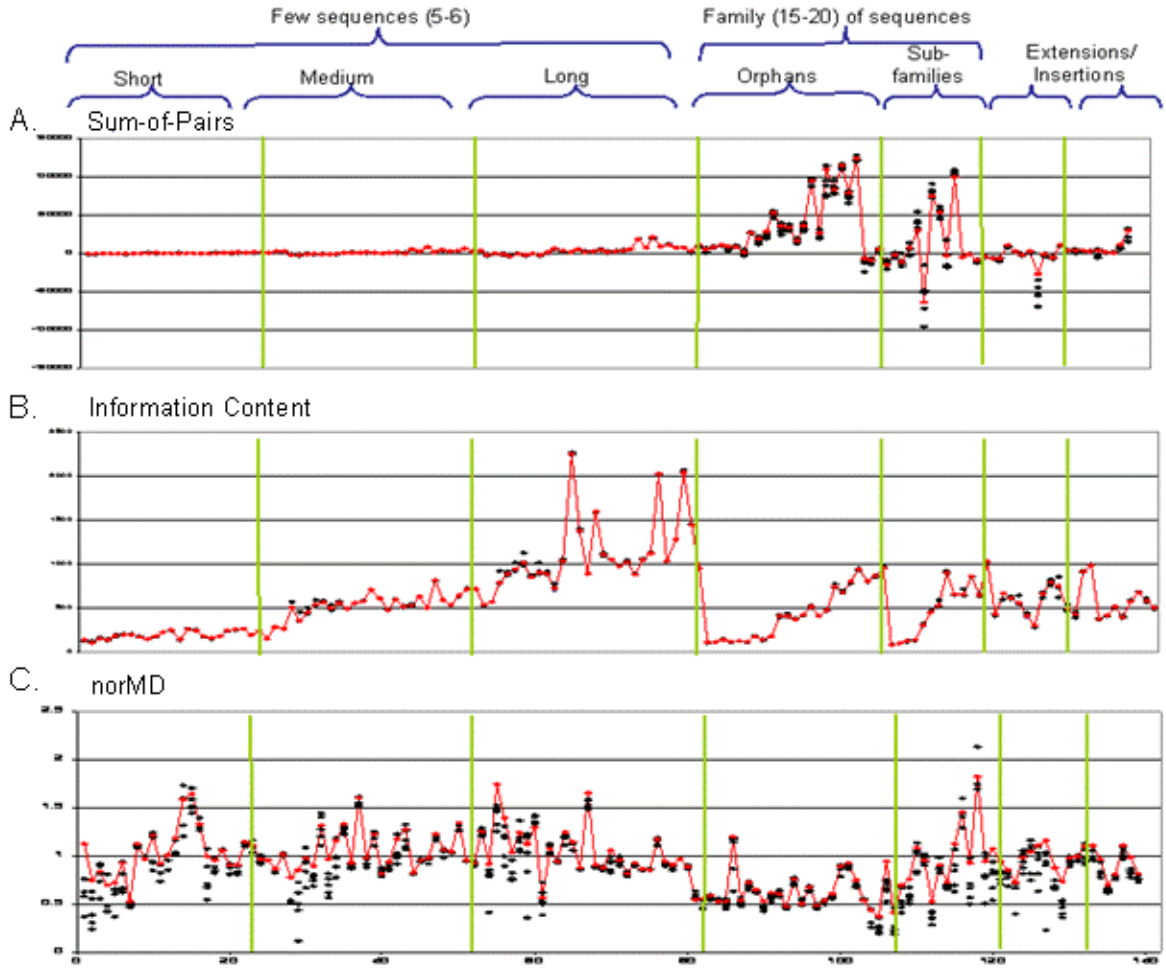


Figure 7.1 Comparison of three objective functions: sum-of-pairs, relative entropy and norMD
Comparison based on the BAliBASE multiple alignment benchmark.Red points indicate scores for the optimal alignment based on 3D superposition. Black points indicate scores for sub-optimal alignments constructed by automatic multiple alignment programs (from Thompson et al., 2001).

The norMD score partially resolves these problems, and can be used to estimate the quality of the alignment even when the optimal alignment score is unknown. As shown in figure 7.1C, most of the alignments scoring higher than the threshold score of 0.5 are correct, while alignments scoring less than 0.3 are generally of poor quality. In addition, the relative difference between the scores for the optimal reference alignment and the sub-optimal alignments produced by different automatic multiple alignment programs is larger for norMD than for either Sum-of-Pairs or Information Content. Nevertheless, a twilight zone still exists for norMD scores between 0.3 and 0.5, where no distinction can be made between good and bad alignments.

7.2 Determination of reliable regions

The objective functions described above calculate a global score that estimates the overall quality of a multiple alignment. However, even when misalignments occur, it is not necessarily true that all of the alignment is incorrect. Useful information could still be extracted if the reliable regions in the alignment could be distinguished from the unreliable

regions. The prediction of the reliability of specific alignment positions has therefore been an area of much interest. One of the first automatic methods for the analysis of position conservation was the AMAS program (Livingstone and Barton, 1993), which was based on a set-based description of amino acid properties. Since then, a large number of different methods have been proposed. For example, Al2Co (Pei and Grishin, 2001) calculates a conservation index at each position in a multiple sequence alignment using weighted amino acid frequencies at each position. The DIVAA method (Rodi *et al.*, 2004) is based on a statistical measure of the diversity at a given position. The diversity measures the proportion of the 20 possible amino acids that are observed. If the position is completely conserved (i.e. only one amino acid is observed in all sequences analyzed), the diversity is 0.05 (1/20); if it is populated by equal proportions of all amino acids, the diversity is 1.0 (20/20). Diversity (as defined here) is inversely and non-linearly related to the measure of sequence information content described above, with a highly conserved position exhibiting relatively low diversity and high information content. For nucleic acid sequences, the ConFind program (Smagala *et al.*, 2005) identifies regions of conservation in multiple sequence alignments that can serve as diagnostic targets and is designed to work with a large number of highly mutable target sequences such as viral genomes.

An alternative approach has been implemented recently in the MUMSA program (Lassmann and Sonnhammer, 2005), based on the comparison of several alignments of the same sequences. The idea is to search for regions which are identically aligned in many alignments, assuming that these are more reliable than regions differently aligned in many alignments. The method also results in a score for a given alignment. A high quality alignment in this case, is one that shares more aligned residues with other alignments. The choice of multiple alignment methods used as input is therefore crucial, in order to avoid a bias towards one particular algorithm. Ideally, different algorithms should be used, such as local and global methods, algorithms designed for transmembrane sequences, repeats, etc. In tests on BAliBASE, the MUMSA scores correlate higher with true alignment quality than the norMD scores. However, a major drawback of the MUMSA method is that several multiple alignments of the same set of sequences have to be constructed for the purpose of comparison, which is not always computationally feasible.

An alternative approach to the calculation of position conservation scores is to use a graphical representation for displaying the patterns in a set of aligned sequences, known as sequence logos, first introduced in 1990 (Schneider and Stephens, 1990). Figure 7.2 shows an example display created using the WebLogo server (Crooks *et al.*, 2004).

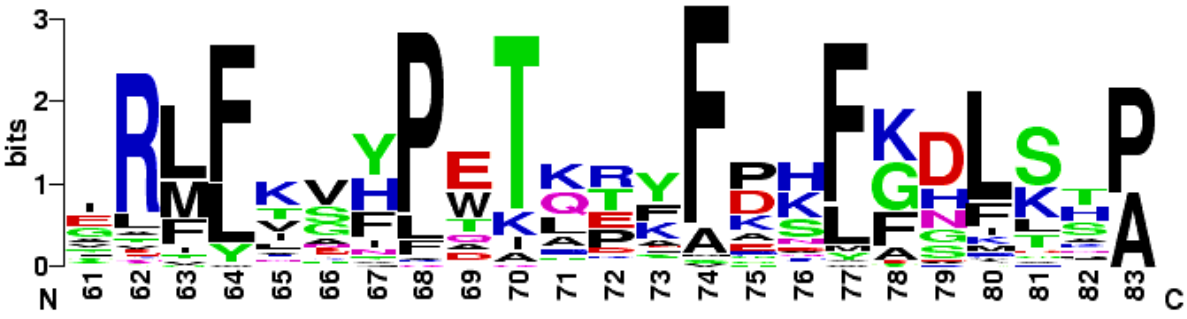


Figure 7.2 An example sequence logo for displaying patterns in aligned sequences
The logo is of the conserved packing and sliding contacts at the end of the B through the beginning of the D helices of the globins (Dickerson and Geis, 1983).

The characters representing the sequence are stacked on top of each other for each position in the aligned sequences. The height of each letter is made proportional to its frequency, and the letters are sorted so the most common one is on top. The height of the entire stack is then adjusted to signify the information content of the sequences at that position.

7.3 Estimation of homology

As seen in chapter 5, many applications have been developed that predict or propagate biological information between the sequences in a multiple alignment based on a presumed homology. The hypothesis is that homologous sequences, *i.e.* sequences that have descended from the same ancestor, often share the same structure and function. A fundamental step in these so-called 'homology-based' methods is the determination of the extent of similarity between the aligned sequences. Without this initial crucial step, the subsequent applications that rely on an accurate multiple alignment cannot be expected to yield high-quality results.

This particular problem has been addressed by a number of groups. The degree to which the sequences in a multiple alignment are related can be estimated by an analysis of positional conservation or by measuring the statistical significance of the alignment (Hertz and Stormo, 1999). Cline *et al.* in 2002 tested four different predictors of alignment position reliability and concluded that near-optimal alignment information was the best predictor, removing 70% of the substantially misaligned positions. Thompson *et al.* (2001) used the NorMD objective function to remove unrelated or badly aligned sequences from multiple alignments. Errami *et al*. (2003) analysed the agreement between predicted secondary structures of the aligned sequences to detect and discard unrelated sequences. Tress *et al*. (2003) used sequence profiles generated from PSI-BLAST alignments to predict reliable regions between remotely related pairs of proteins. These methods work well when the sequences to be compared are homologous over their full lengths, but large multi-domain proteins are becoming more and more prevalent in the sequence databases, with the arrival of numerous new genome sequences, in particular from eukaryotic organisms. In the face of these highly complex proteins, the definition of 'homologous sequences' needs to be more detailed. Two sequences can share one or more homologous domains without being homologous over their full-lengths.

A more recent method, LEON (Thompson *et al.*, 2004) has been developed to determine the extent of homology between proteins based on the MACS. LEON incorporates some of the latest developments in multiple alignment analysis, including sequence clustering (Wicker *et al.*, 2001) and the identification of locally conserved motifs or 'core blocks' (Thompson *et al.*, 2003). In LEON, weak signals from distantly related proteins can be considered in the overall context of the family and intermediate sequences and the combination of individual weak matches are used to increase the significance of low-scoring regions. Residue composition is also taken into account by the incorporation of several existing methods for the detection of compositionally biased sequence segments. LEON can be used to reliably identify the complex relationships between large multi-domain proteins and should be useful for automatic high-throughput genome annotations, 2D/3D structure predictions, protein–protein interaction predictions.

7.4 Multiple alignment benchmarks

The methods described above are used to determine the quality and reliability of a given multiple alignment. In computer science, the quality of an algorithm is often estimated by comparing the results obtained with a pre-defined benchmark or 'gold standard'. Clearly, the tests need to be of high-quality. Errors in the benchmark will lead to biased or erroneous results. The tests in the benchmark need not be comprehensive, but must be representative of ones that the system is reasonably expected to handle in a natural (meaning not artificial) setting and the performance measure used must be pertinent to the comparisons being made. Enough tests need to be included in order to obtain statistical differences between programs tested. It should be possible to complete the task domain sample and to produce a good solution. A task that is too difficult for all or most tools yields little data to support comparisons. A task that is achievable, but not trivial, provides an opportunity for systems to show their capabilities and their shortcomings (Sim *et al.*, 2003).

One of the first studies to compare the quality of different methods was performed in 1994, when McClure *et al.* compared several progressive alignment methods, including both global and local algorithms. They concluded that global methods generally performed better. However, the number of suitable test sets available at that time was somewhat limited and this was therefore not a comprehensive test.

7.4.1 BAliBASE

One of the first large scale benchmarks specifically designed for multiple sequence alignment was BAliBASE (Thompson *et al.*, 1999a; Bahr *et al.*, 2001). The alignment test cases in BAliBASE are based on 3D structural superpositions that are manually refined to ensure the correct alignment of conserved residues. The alignments are organised into reference sets that are designed to represent real multiple alignment problems. The first version of BAliBASE consisted of 5 reference sets representing many of the problems encountered by multiple alignment methods at that time, from a small number of divergent sequences, to sequences with large N/C-terminal extensions or internal insertions (see figure 7.3).

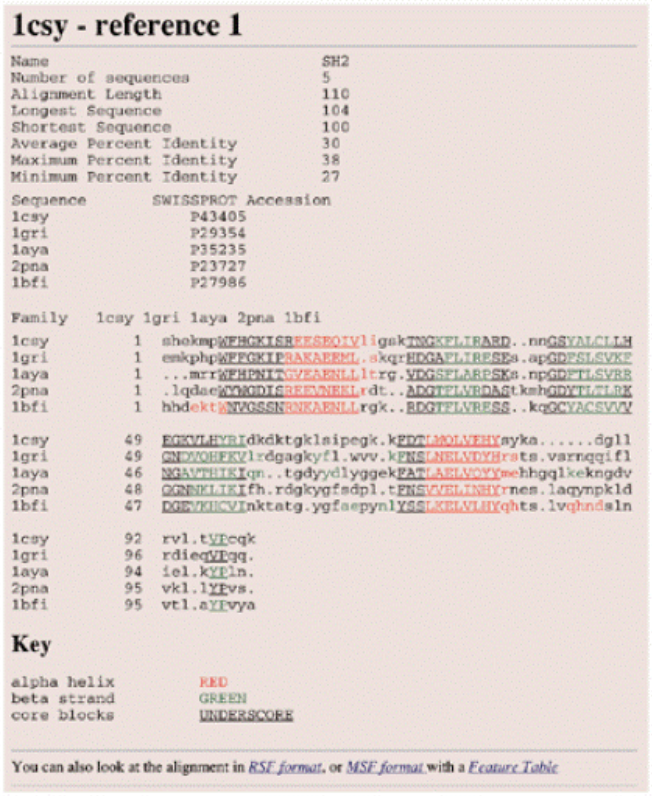
In version 2 (Bahr *et al*., 2001), three new Reference sets were included, devoted to the particular problems posed by sequences with transmembrane regions, repeats and inverted domains. In each reference alignment, core blocks are defined that exclude the regions for which the 3D structure superpositions are unreliable, for example, the borders of secondary structure elements or in loop regions.

a)

Reference 1	<100 residues	200-300 residues	>500 residues
<25% identity	7	8	8
20-40% identity	10	9	10
>35% identity	10	10	8
Reference 2	9	8	7
Reference 3	5	3	5

Reference 4	12
Reference 5	12

b)



1csy - reference 1

```
Name                              SH2
Number of sequences               5
Alignment Length                  110
Longest Sequence                  104
Shortest Sequence                 100
Average Percent Identity          30
Maximum Percent Identity          38
Minimum Percent Identity          27

Sequence        SWISSPROT Accession
1csy                P43405
1gri                P29354
1aya                P35235
2pna                P23727
1bfi                P27986

Family   1csy 1gri 1aya 2pna 1bfi

1csy        1  shekmpWFHGKISREESEQIVligskTNGKFLIRARD..nnGSYALCLLH
1gri        1  emkphpWFFGKIPRAKAEEML.skqrHDGAFLIRESE..apGDFSLSVKF
1aya        1  ...mrrWFHPNITGVEAENLLltrg.VDGSFLARPSK..npGDFTLSVRR
2pna        1  .lqdaeWYWGDISREEVNEKLrdt..ADGTFLVRDAStkmhGDYTLTLRK
1bfi        1  hhdektWNVGSSNRNKAENLLrgk..RDGTFLVRESS..kqGCYACSVVV

1csy       49  EGKVLHYRIdkdktgklsipegk.kFDTLWQLVEHYsyka......dgll
1gri       49  QNDYDHFKVlrdgagkyfl.wvv.kFNSLNELVDYHrts.vsrnqqifl
1aya       46  NGAVTHIKIqn..tgdyydlyggekFATLAELVQYYmehhgqlkekngdv
2pna       48  QGRNKLIKIfh.rdgkygfsdpl.tFNSVVELINHYrnes.laqynpkld
1bfi       47  DGEVKHCVInktatg.ygfsepynlYSSLKELVLHYqhts.lvqhndsln

1csy       92  rvl.tYPcqk
1gri       96  rdieqYPqq.
1aya       94  iel.kYPln.
2pna       95  vkl.lYPvs.
1bfi       95  vtl.aYPvya

Key

alpha helix          RED
beta strand          GREEN
core blocks          UNDERSCORE
```

You can also look at the alignment in *RSF format*, or *MSF format* with a *Feature Table*

Figure 7.3 Version 1 of the BAliBASE benchmark alignment database
a) The number of alignments in each reference set. Reference 1 contains alignments of (<6) equi-distant sequences. Reference 2 aligns up to three 'orphan' sequences (<25% identical) from reference 1 with at least 15 closely related sequences. Reference 3 consists of up to four subgroups with <25% residue identity between groups. References 4 and 5 contain sequences with N/C-terminal extensions or insertions respectively. b) The Web display of an alignment, showing secondary structure elements and the conserved core blocks (adapted from Thompson et al., 1999).

In order to assess the accuracy of a multiple alignment program, the alignment produced by the program for each BAliBASE test case is compared to the reference alignment. Two scores are used to evaluate the alignment :

- the SP (sum of pairs) score calculates the percentage of pairwise residues aligned the same in both alignments
- the CS (column score) calculates the percentage of complete columns aligned the same. These scores are calculated in the core block regions only.

A comparison of some of the alignment methods described in chapter 6 (Thompson *et al.*, 1999b), based on BAliBASE (version 1.0) showed that there was no single algorithm that was consistently better than the others.

The study revealed a number of specificities in the different algorithms (see figure 7.2). For example, while most of the programs successfully aligned sequences sharing >40% residue identity, an important loss of accuracy was observed for more divergent sequences with <20% identity. Another important discovery was the fact that global alignment methods in general performed better for sets of sequences that were of similar length, although local algorithms were more successful at identifying the most conserved motifs in sequences containing large extensions and insertions. Of the local methods, Dialign (Morgenstein *et al.*, 1996) was the most successful. The iterative methods, such as PRRP (Gotoh, 1996) or SAGA (Notredame *et al.*, 1996) were generally more accurate than the traditional progressive methods, although at the expense of a large time penalty.

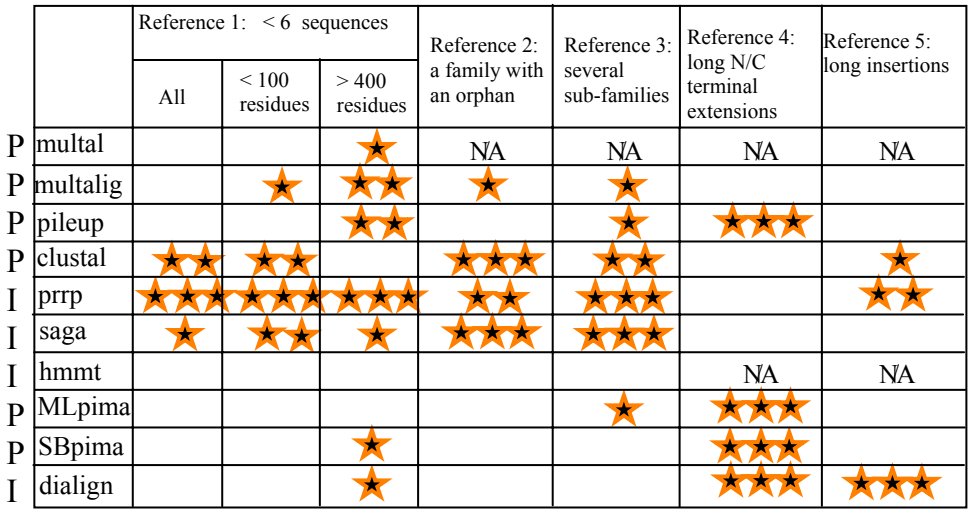


Figure 7.4 Comparison of multiple alignment programs using the alignments in the BAliBASE benchmark
Stars were assigned to the top ranking programs that were significantly different according to a Friedman rank test.

7.4.2 OxBench

The OXBench benchmark suite from the Barton group (Raghava *et al.*, 2003), provides multiple alignments of protein domains that are built automatically using structure and sequence alignment methods. The automatic construction means that a large number of tests can be included, however the benchmark results will be biased towards sequence alignment programs using the same methodology as that used to construct the reference. The benchmark is divided into three data sets. The master set currently consists of 218 alignments of sequences of known 3D structure, with from 2 to 122 sequences in each alignment. The extended data set is constructed from the master set by including sequences of unknown structure. Finally, the full-length data set includes the full-length sequences for the domains in the master data set.

A number of different scores are included in the benchmark suite, in order to evaluate the accuracy of multiple alignment. The average number of correctly aligned positions is similar to the column score used in BAliBASE. This can be calculated over the full alignment or over Structurally Conserved Regions (SCR). The Position Shift Error measures the average magnitude of error, so that misalignments that cause a small shift between two sequences are penalised less than large shifts. Two other measures are also provided that are independent of the reference alignment, and are calculated from the structure superposition implied by the test alignment.

The OxBench suite was used to compare 8 different alignment programs, including many of those in the BAliBASE study, together with the AMPS program (Barton and Sternberg, 1987). The MSA (Lipman *et al.*, 1989) and T-COFFEE (Notredame *et al.*, 2000) programs were also tested, although the tests were restricted to a smaller data set because these methods were unable to align the largest test sets due to prohibitive space and time requirements. The AMPS program was shown to perform as well or better than the other progressive alignment methods in most tests. The T-COFFEE method which incorporates both local and global pairwise alignment algorithms was shown to outperform the other methods on the smaller data set. Another important result was that the rigorous dynamic programming method used in the MSA program did not perform better than the heuristic progressive methods in this study, supporting the hypothesis that the optimal sum-of-pairs score does not always correspond to the biologically correct alignment (see discussion in section 7.1).

7.4.3 PREFAB

The PREFAB (Edgar, 2004) benchmark was constructed using a fully automatic protocol and currently contains 1932 multiple alignments. Pairs of sequences with known 3D structures were selected and aligned using two different 3D structure superposition methods. A multiple alignment was constructed for each pair of structures, by including 50 homologous sequences detected by sequence database searches.

The accuracy of an alignment program is estimated by comparing the alignment of the structure pair in the test multiple alignment with the reference superposition in each test case. Only positions that are aligned the same by the two different superposition methods are considered. The PREFAB benchmark was used to compare the MUSCLE program (Edgar, 2004) with MAFFT (Katoh *et al.*, 2002), T-COFFEE and ClustalW and showed that the MUSCLE program performed significantly better than the other methods. The programs were also compared with the BAliBASE benchmark, where a similar ranking of programs was obtained although the difference between MUSCLE and T-COFFEE was not significant in this case.

7.4.4 SABmark

SABmark (van Walle *et al.*, 2005) contains reference sets of sequences derived from the SCOP protein structure classification, divided into 2 sets, twilight zone (Blast E-value>=1) and superfamilies (residue identity<=50%). Pairs of sequences in each reference set are then aligned based on 3D structure superpositions. To evaluate the quality of a multiple alignment program, multiple alignments of each reference set are constructed. Pairwise alignments are then extracted from the multiple alignment and compared to the structure superpositions. Although the benchmark covers most of the known protein fold space, the major

disadvantage of this benchmark is that only pairwise reference alignments are considered and no multiple alignment solution is provided.

In a comparison of 4 different alignment methods using SABmark (van Walle *et al.*, 2005), two different scores were used. The first, f_D is similar to the SP score and is defined as the ratio of the number of correctly aligned residues divided by the length of the reference alignment, and may be thought of as a measure of sensitivity. The f_M score measures the specificity and is defined as the ratio of the number of correctly aligned residues divided by the length of the test alignment. At the SCOP family level, T-COFFEE and ClustalW were shown to perform better, while Align-m (van Walle *et al.*, 2004) was more successful at constructing pairwise alignments at the SCOP superfamily level.

7.4.5 Homstrad

HOMSTRAD (Mizuguchi *et al.*, 1998) is a database of protein families, clustered on the basis of sequence and structural similarity. It was not specifically designed as a benchmark database, although it has been employed as such by a number of authors

7.4.6 BRAliBASE

BRAliBASE (Gardner *et al.*, 2005) includes four diverse structural RNA datasets of Group II introns, 5S rRNA, tRNA and U5 spliceosomal RNA. The sequences and the reference alignments were obtained from the Rfam v5.0 database. Approximately 100 sub-alignments were also generated for each of the four families. The sub-alignments contained five sequences each and encompassed a range of sequence identities. This large dataset was divided into high (≥75% sequence identity, 73 alignments), medium (55-75% sequence identity, 73 alignments) and low (<55% sequence identity, 242 alignments) sequence homology groups. An additional tRNA dataset was also generated with just two sequences to each alignment.

The datasets were used to evaluate both sequence and structure alignment methods. It was found that sequence alignment alone, using the current algorithms, was generally inappropriate for <50–60% sequence identity. Below this limit, algorithms incorporating structural information outperformed pure sequence-based methods. However, these algorithms are computationally demanding which severely limits their use in practice.

7.4.7 Comparison of multiple alignment benchmarks

A comparison of a number of benchmarks for protein sequence alignment algorithms, including those described above, has been performed recently (Blackshields *et al.*, 2006). They concluded that, although SABmark boasts full coverage of the known fold space, there are only pairwise references for each group, so multiple alignment assessment becomes complicated depending on how the results are treated. The importance of core region annotation was also stressed by the authors. HOMSTRAD is often used as a benchmark though it lacks this annotation. Finally, they recommended that several benchmarks be used for program comparison, although this can become time-consuming and confusing. Oxbench, PREFAB and BAliBASE all contain difficult cases containing full-length sequences of low sequence identity. The authors noted that BAliBASE has the advantage that several distinct problem areas are explicitly addressed. It is smaller than the other test sets, but nevertheless

has a large enough range of representative examples from the known fold-space to evaluate relative performance.

7.5 Multiple alignment revolution

The objective evaluation of alignment quality and the introduction of large-scale alignment benchmarks have clearly had a positive effect on the development of multiple alignment methods (see figure 7.3). From their beginnings in 1975, until 1994 when McClure first compared different methods systematically, the main innovation was the introduction of the heuristic progressive method that allowed the multiple alignment of larger sets of sequences within a practical time limit.

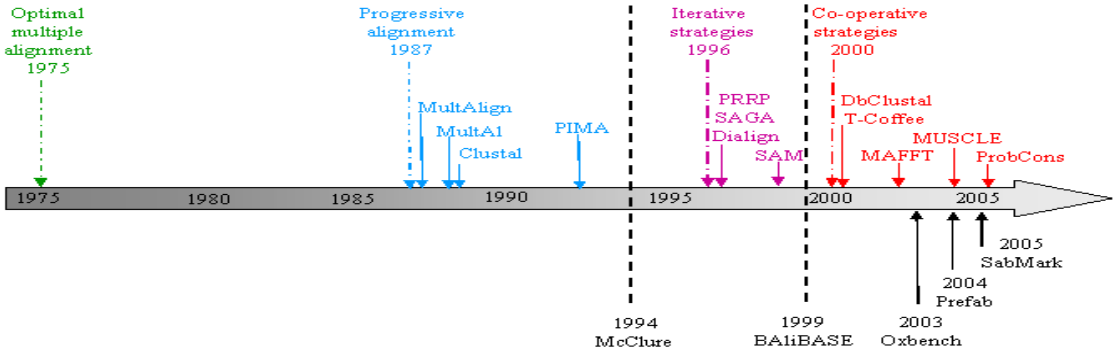


Figure 7.5 The simultaneous development of multiple alignment algorithms and alignment benchmarks
The development of multiple alignment algorithms (above the dateline) is shown in the context of the analogous developments of alignment benchmarks (below the dateline). The dotted arrows represent the introduction of a novel approach to the multiple alignment problem.

Soon after this initial comparison, various new methods were introduced that exploited novel algorithms, such as Iterative refinement, Hidden Markov Models or Genetic algorithms. These new approaches significantly improved alignment quality, as shown in the comparison of these methods by Thompson *et al.*, in 1999b. Nevertheless, this study highlighted the fact that no single algorithm was capable of constructing high quality alignments for all test cases. In particular, global methods (e.g. ClustalW) were shown to perform well when the sequences were homologous over their entire lengths, while local methods (e.g. DiAlign) were shown to perform better when the sequence set contained large insertions or N/C terminal extensions.

As a consequence, the first methods were introduced that combined both global and local information in a single alignment program, such as dbClustal, T-Coffee, MAFFT, MUSCLE or ProbCons. Table 7.1 shows the scores obtained using most of these new methods for the different multiple alignment benchmarks described above.

	SABmark	PREFAB		OxBench		BAliBASE (version 2)					
	25-50%	0-20%	20-40%	20-40%	40-60%	Ref1	Ref2	Ref3	Ref4	Ref5	Time (hrs:mins)
DiAlign	41.9	34.1	78.1	34.9	66.7	70.9	35.9	34.4	**76.2**	84.3	2:53
ClustalW	41.2	41.1	78.3	38.7	66.2	77.3	56.8	46.0	52.2	63.8	1:07
T-Coffee	44.4	44.8	81.8	38.3	73.3	77.4	56.1	48.7	73.0	90.3	21:31
MAFFT	45.2	48.6	83.8	38.4	71.5	78.1	50.2	50.4	72.7	85.9	1:18
MUSCLE	**50.6**	46.0	83.0	**39.9**	72.1	80.8	56.3	56.4	60.9	90.2	**1:05**
ProbCons	48.4	**49.0**	**85.2**	39.7	**74.4**	**82.6**	**61.3**	**61.3**	72.3	**91.9**	5:32

Table 7.1 Current state of the art for multiple sequence alignment methods
All scores shown are column scores. For PREFAB, the score is calculated in the superposable regions. For OxBench, the full alignments were used and the scores were calculated in structurally conserved regions only. For BAliBASE, the scores are for core block regions only.

The new combined strategies certainly improve alignment quality for a wide range of alignment problems. However, using the existing multiple alignment benchmarks it is becoming more and more difficult to make clear distinctions between the more recent methods. Therefore, the benchmarks must now evolve if they are to keep pace with the multiple alignment revolution. Hopefully, new benchmarks with larger, more complex test sets will stimulate the development of new alignment algorithms and vice versa.

8 Material and Methods

The algorithms and methods for multiple alignment construction, evaluation and analysis described in chapters 9-12 of this thesis were developed using the existing infrastructure and computer resources of the Laboratoire de Biologie et Genomique Structurales (LBGS) and the Plate-forme Bio-informatique de Strasbourg (BIPS). The BIPS is a high-throughput platform for comparative and structural genomics, which was identified in 2003 as a national inter-organisational technology platform (*Plate-forme Nationale RIO).* The BIPS is also part of the national genomics and biotechnology network (*Génopôle Grand-Est* - "*from gene to drug*").

8.1 Computing resources

8.1.1 Servers

Three central servers are currently available for program development and computational data analyses:

(i) Interactive and web services : Sun Enterprise 450 (Solaris 9). 4 processors with 1 Gb shared memory.
(ii) Computational servers:
- 6 Compaq ES40 cluster (Tru64 UNIX). 6 x 4 EV67 processors. Of the 6 machines in this cluster, 5 have 4 Gb memory each, and the sixth has 16 Gb.
- 6 Sun Enterprise V40z server (2 x Solaris 10 and 4 x RedHat Enterprise Linux 4). 6 x 4 Opteron processors with 2 x 32 Gb and 4 x 16 Gb memory.

(iii) Disk server: Sun V480 (Solaris 9) providing 8 Tera-bytes on Raid5 disks shared with other servers using NFS.

8.1.2 Databases

A number of general as well as some more specialist databases are installed and updated regularly on the LBGS servers. These databases are available in GCG format (Butler, 1998) and can also be queried using the SRS (Sequence Retrieval Software) (Etzold and Argos, 1993).

Generalist databases:

The main public sequence and structure databases have been installed locally on the IGBMC servers. The protein sequence database Uniprot (Wu *et al.*, 2006), consists of both SwissProt and SpTrEMBL databases (Boeckmann *et al.*, 2003). The SpTrEMBL sequences are produced by automatic translation of the coding sequences from the EMBL nucleotide sequence database. After validation and annotation by experts, the sequences in SpTrEMBL are incorporated in the SwissProt database. The protein 3D structure database PDB (Protein Data Bank) (Kouranov *et al.*, 2006), includes structures determined by X-ray crystallography or by NMR. The amino acid sequences of the proteins or domains in PDB are also available.

Specialist databases:

In addition to these generalist databases, a number of specialist databases are maintained locally. In particular, the InterPro database (Mulder *et al.*, 2005) contains information on protein families, protein domains and functional sites. InterPro is a collaboration between a number of different protein signature databases, including the protein domain databases: Pfam (Bateman *et al.*, 2004), Prodom (Bru *et al.*, 2005), Smart (Letunic *et al.*, 2006) and the protein pattern databases: Prints (Attwood, 2002) and Prosite (Hulo *et al.*, 2006). Protein signatures are manually integrated into InterPro database entries and are then curated to provide reliable biological and functional information. InterPro also provides links to other specialised databases, including the Gene Ontology (Ashburner *et al.*, 2000).

8.1.3 GCG package

The GCG package (Wisconsin Package Version 10.2, Genetics Computer Group Madison, Wisc.) is a software suite containing diverse sequence analysis programs. Version 10 allows the manipulation, visualization, analysis and comparison of sequences in the GCG format databases installed locally. In particular, the SEQLAB multiple alignment editor was used here for the visualization and manual correction of the various multiple sequence alignments produced during the course of this work. The SEQLAB editor provides an easy-to-use, graphical interface with many facilities for editing alignments.

8.1.4 Sequence Retrieval Software (SRS)

The Sequence Retrieval Software (SRS) (Etzold and Argos, 1993) is an integration system that serves as a gateway to many major databases in the field of molecular biology. SRS currently allows access to more than 150 biological databases, including nucleic acid and protein sequences and structures, protein domains and metabolic pathways. It is designed for the extraction of semi-structured data, i.e. textual data with a pre-defined structure that may include redundancies or irregularities. The textual data is stored in flat files containing all the entries of a database. The flat files are organised into structured fields, which may be different depending on the databases. SRS performs a grammatical parsing of the information contained in the flat files and then indexes the different fields associated with each entry. This indexing allows a rapid access to the entry fields via complex queries, as well as cross-referenced queries which exploit the links between the different databases. Version 6 of SRS is currently installed at the LBGS. Database queries can be performed interactively (http://bips.u-strasbg.fr/srs6/) or using UNIX commands.

8.2 The GScope platform

Gscope is an in-house platform, specifically designed for high-throughput data analyses, either for complete genomes or for large sets of genes or proteins. The platform is developed by Raymond Ripp (manuscript in preparation), using the Tcl/Tk language and outputs the results in ASCII flat files. GScope uses both internal and external programs to perform a wide range of data analyses, ranging from DNA sequence processing, such as open reading frame (ORF) identification and sequence database searches to data clustering, cross-validation and predictions at the genome level (figure 8.1).

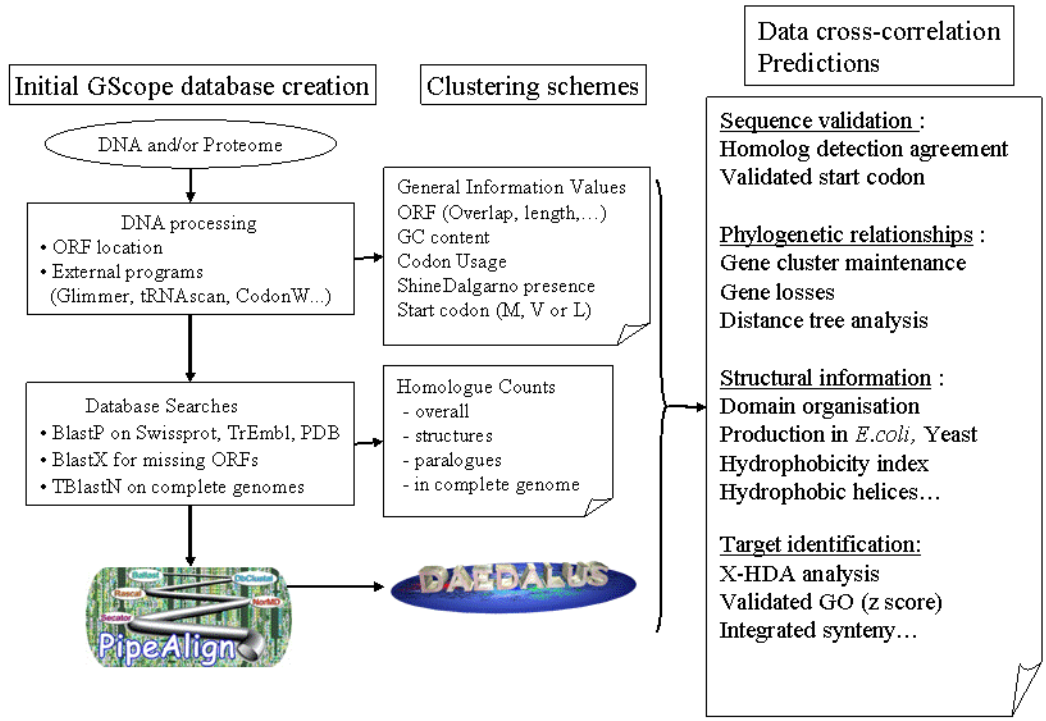


Figure 8.1 Schematic overview of the Gscope high throughput platform processing pipeline

In this work, the GScope platform was used to perform protein sequence and structure database searches and to construct the multiple sequence alignments for the large scale test sets and applications described in chapters 9-12.

8.2.1 Sequence and structure database searching

The sequence database searches were performed using the BlastP program (Altschul *et al.*, 1997). Given a query sequence of interest, BlastP provides an efficient, sensitive means of searching the sequence databases to find sequences that resemble the query. Here, BlastP searches were performed in the UniProt and PDB databases, using the default parameters. The output from BlastP includes a list of the top-scoring database sequences and a local pairwise alignment for each database hit of the similar regions identified.

The output of the initial BlastP search was then used to perform a more in-depth search for proteins with known 3D structures. Protein structures are often resolved for a given region of a particular sequence of interest that may not correspond to the complete protein. The partial sequences in the PDB database are penalised by the length correction parameter, which is part of the default parameters of the BlastP program. This can cause problems in large sequence families because the number of database hits is limited and proteins that are highly similar to the query can saturate the results. In this case, the short PDB sequence fragments may not be included in the BlastP final output file. We therefore implemented a two step approach, similar to that used by Rychlewski *et al.*, 2000 for their profile-profile alignment method. A Position Specific Scoring Matrix (PSSM) was constructed from the BlastP search of the large UniProt database, and then this PSSM was used to search the PDB database with the PsiBlast (Altschul *et al.*, 1997) iterative database search program. The PsiBlast search was stopped after 5 iterations in order to avoid the problems of 'profile drift' (Muller *et al.*,

1999), where the profile moves too far away from the original sequence pattern and increases the chance of spurious matches.

8.2.2 Multiple alignment construction

Multiple alignments of complete sequences (MACS) were built of all the top scoring sequences detected by the BlastP and PsiBlast searches with an Expect value greater than 10.0. The Expect value is a parameter that describes the number of hits one can expect to see by chance when searching a database of a particular size. It decreases exponentially with the score that is assigned to a match between two sequences. For example, an E value of 1 assigned to a database hit can be interpreted as meaning that in a database of the current size one might expect to see 1 match with a similar score by chance. We used an Expect value of 10 as the significance threshold for inclusion in the MACS, in order to include the maximum number of homologous sequences, although this will clearly lead to a number of unrelated sequences being included in the initial MACS. However, these unrelated sequences will be identified in the context of the MACS and will be removed in the subsequent multiple alignment processing.

The multiple alignment processing pipeline implemented in GScope consists of six steps, from homology database searches to the construction of a high quality, hierarchical multiple alignment of complete sequences. The pipeline, known as PipeAlign (Plewniak *et al.*, 2003) is described in detail in section 8.3.

8.3 PipeAlign protein family analysis toolkit

PipeAlign (Plewniak *et al.*, 2003) is a protein family analysis developed in the LBGS.

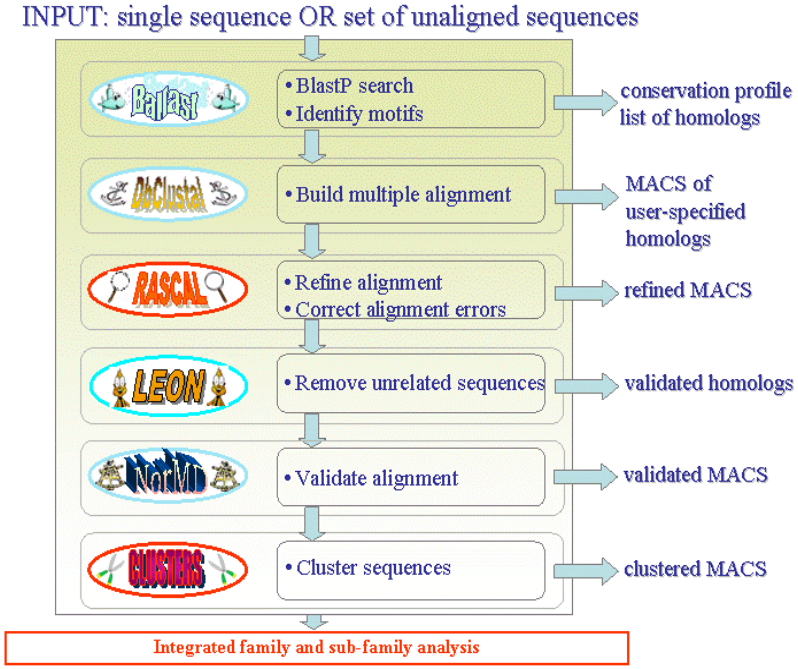


Figure 8.2 Overview of PipeAlign multiple alignment construction pipeline
(adapted from Plewniak et al., 2003)

The pipeline integrates a six step process ranging from the search for sequence homologues in the protein and 3D structure databases to the definition of the hierarchical relationships within and between subfamilies. The complete, automatic pipeline takes a single sequence or a set of sequences as input and performs an initial BlastP search in the UniProt and PDB databases. The database search is followed by a cascade of six different sequence analysis programs, shown in figure 8.2 and described in detail below.

8.3.1 Ballast: post-processing of BlastP results

Ballast (Plewniak *et al.*, 2000) builds a conservation profile of the database hits detected by BlastP. The contribution of each database hit is proportional to its significance, i.e. its E-value. The conservation profile is smoothed and then peaks are detected using the second derivative of the smoothed profile. These peaks define local maximum segments (LMSs) that correspond to sequence segments that are more conserved than their flanking regions. The positions 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 DbClustal.

8.3.2 DbClustal: construction of the MACS

DbClustal (Thompson *et al.*, 2000) integrates the local conservation information from the Ballast LMS or 'anchor' file in the global multiple sequence alignment program, ClustalW (Thompson *et al.*, 1994). ClustalW incorporates the global dynamic programming algorithm developed by Needleman and Wunsch, 1971 (see section 1.6.2.1). The recursive algorithm was modified in DbClustal, such that the score for aligning any pair of residues combines the residue comparison matrix score for the two amino acids (see section 1.6.1.1) and the anchor scores from the Ballast program. An anchor propagation is also incorporated, as Ballast determines anchors for each database sequence relative to the query sequence only. Therefore, DbClustal propagates these anchors between all the sequences. The alignment weighting scheme implemented in DbClustal means that the global alignment is encouraged towards, but not constrained to, the conserved motifs.

8.3.3 RASCAL: rapid scanning and correction of alignment errors

DbClustal is a heuristic algorithm that can sometimes introduce errors into the multiple alignment. The RASCAL program (Thompson *et al.*, 2003) is designed to detect these errors and to correct them. The method implemented in RASCAL is shown in figure 8.3. The multiple alignment output by DbClustal is first divided horizontally and vertically to form a lattice in which well aligned, reliable regions can be differentiated. Potential errors are detected by comparing profiles of the reliable regions. RASCAL then performs a single re-alignment of each badly aligned region using an algorithm similar to that implemented in ClustalW (Thompson *et al.*, 1994).Alignment correction is restricted to the less reliable regions only, leading to a more reliable and efficient refinement strategy.

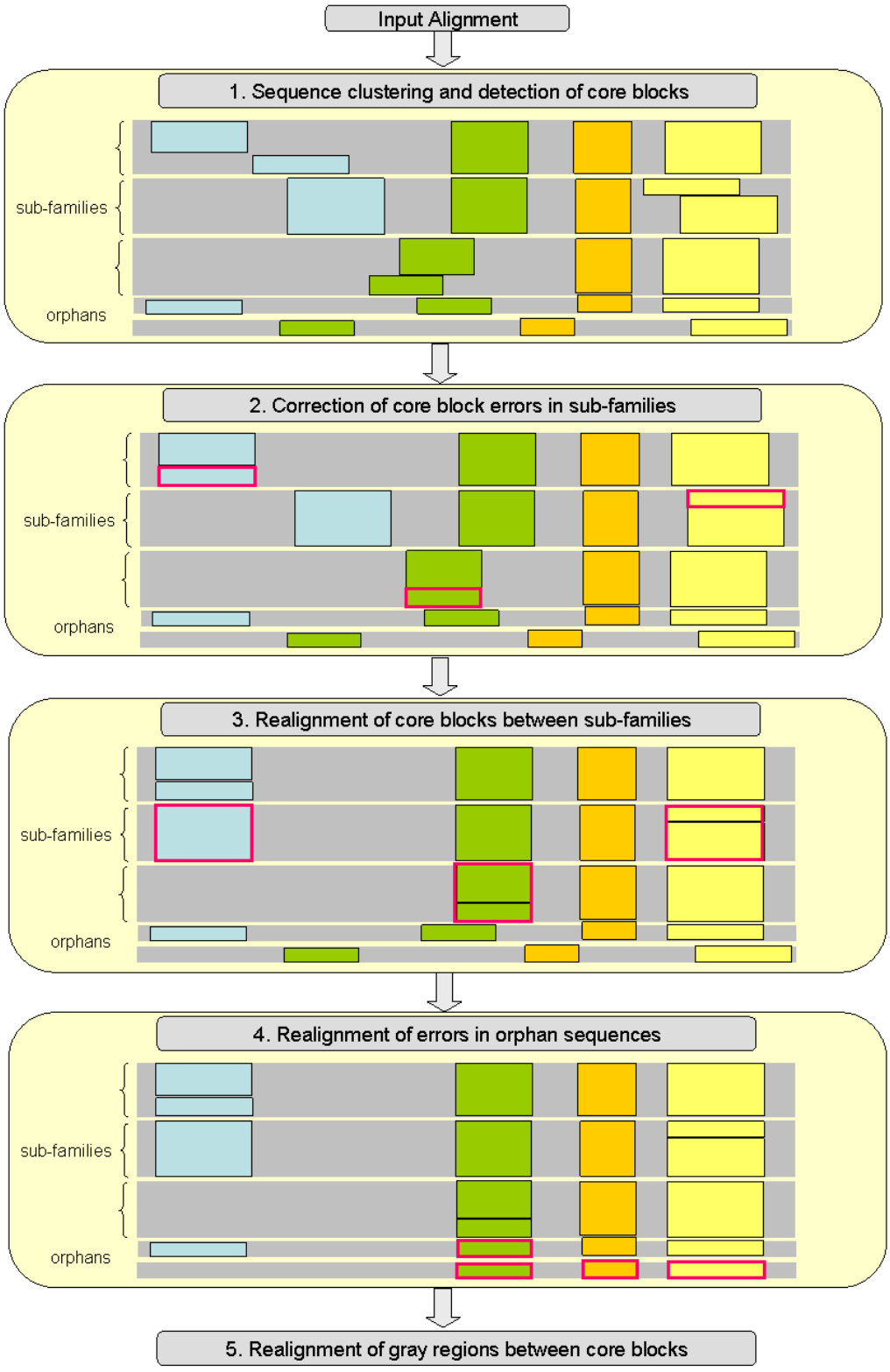


Figure 8.3 Overview of the RASCAL algorithm

1. *The sequences in the input alignment are divided into sub-families and reliable core block regions are identified. 2. For each sub-family, misaligned sequences are realigned. 3. Misaligned core blocks between sub-families are corrected. 4. Divergent 'orphan' sequences are realigned.5. Finally, the regions between core blocks are realigned (shown in gray).*

8.3.4 LEON: multiple alignment-based homology evaluation

The next step in the pipeline is designed to detect the sequences in the MACS that are unrelated to the query sequence. The LEON program (Thompson *et al.*, 2004) uses the reliable regions, or 'core blocks' determined by RASCAL. An overview of the method is shown in figure 8.4.

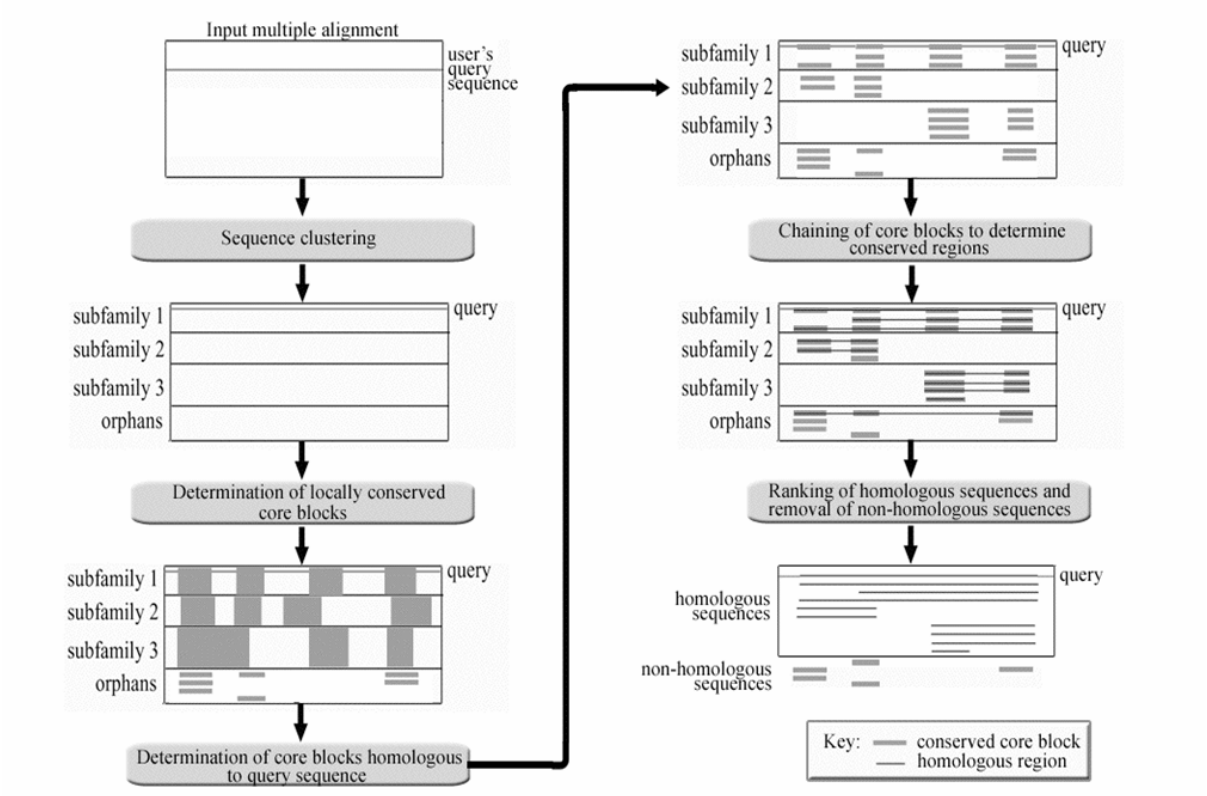


Figure 8.4 Overview of the LEON algorithm

Taking advantage of the transitive nature of homologous relationships, information from intermediate sequences is used to help define the conserved core blocks for the more divergent sequences. The conserved core blocks for each subfamily in the MACS are then chained together to form contiguous regions that are considered to be homologous to the query sequence. The amino acid composition of the sequences is also taken into account by the incorporation of a number of different algorithms for the detection of compositionally biased segments. Finally, any sequences that do not contain any homologous regions are removed from the MACS. The output from LEON is thus a high quality MACS containing only those sequences that share at least one homologous region with the query.

8.3.5 NorMD: MACS quality evaluation

The NorMD objective function (Thompson *et al.*, 2001) is used to evaluate the quality of the MACS produced by the four previous steps. NorMD combines the advantages of a column-scoring technique with the sensitivity of methods incorporating residue similarity scores. NorMD is based on the Mean Distance (MD) scores introduced in ClustalX. A score for each column in the alignment is calculated using the concept of continuous sequence space (Vingron and Sibbald, 1993) and the column scores are summed over the full length of

the alignment. The MD score is then normalized to take into account the number of sequences, the length of each sequence and their estimated similarity, as shown in figure 8.5.

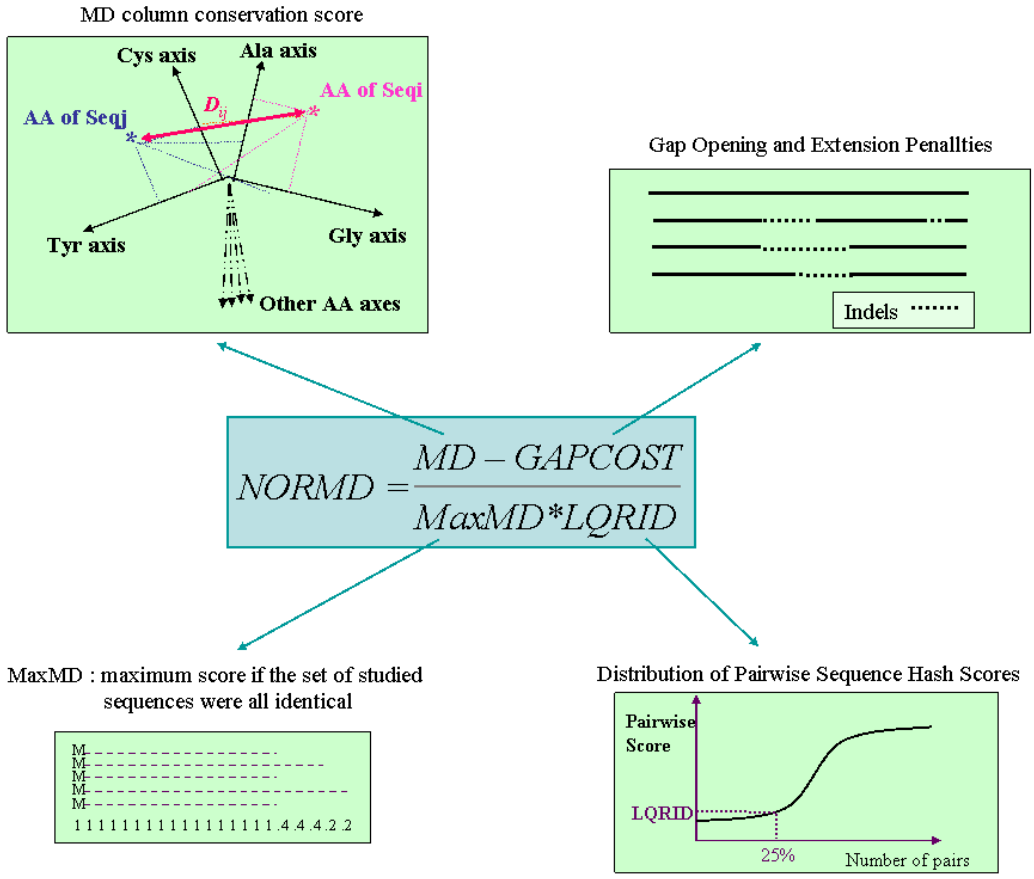


Figure 8.5 Calculation of the norMD score for a multiple sequence alignment
The Mean Distance (MD) score for a multiple alignment is defined as the sum of the conservation scores for each column in the alignment, as defined in ClustalX. The MD score is then normalised to take into account the number of gaps in the alignment (GAPCOST), the maximum score possible for the same set of sequences if they were all identical (MaxMD) and the lower quartile range of the distribution of the pairwise sequence percent indentities (LQRID).

The normalised scores allow us to define a cutoff above which the alignment is probably of high quality. Here, a multiple alignment with a NorMD score greater than 0.5 is considered to be mostly well aligned. The norMD scores were compared with other multiple alignment objective functions in Section 1.7.1.

8.3.6 Secator: sequence clustering

The Secator program (Wicker *et al.*, 2001) clusters the sequences in a multiple alignment into potentially functional subgroups. The number of subgroups is determined automatically by the program. The first step is to create a phylogenetic tree from a distance matrix based on the MACS. Secator then assigns a dissimilarity value to each node in the tree and collapses branches by automatically detecting the nodes joining distant subtrees. The remaining subtrees represent sequence subfamilies in the alignment.

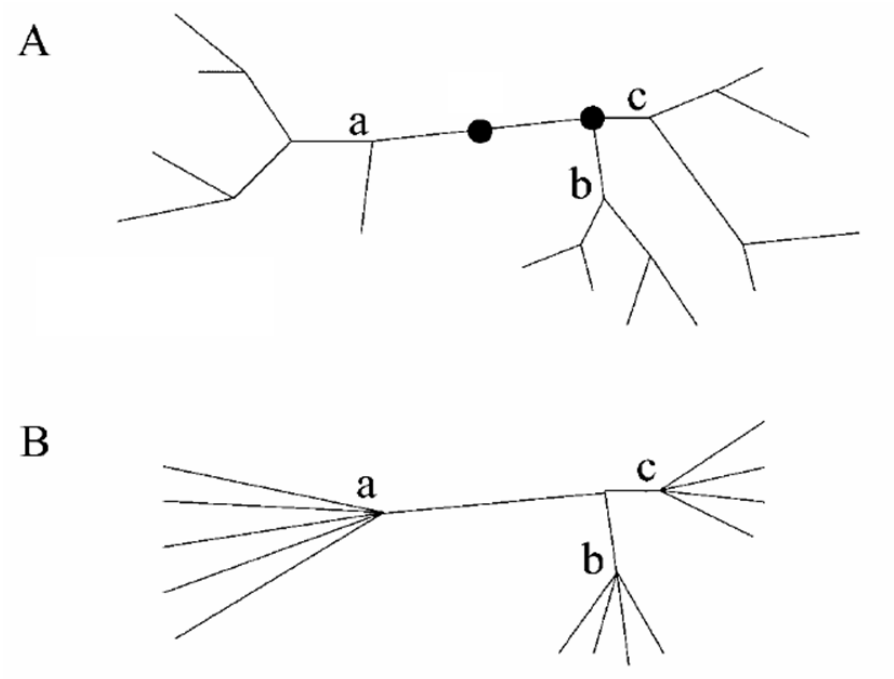


Figure 8.6 Example of Secator sequence clustering by collapsing branches of a tree
Phylogenetic tree (A) before and (B) after collapsing. Branches are collapsed from the leaves up to the internal branches joining distant subtrees (from Wicker et al., 2001).

The final output of the PipeAlign system is a high-quality, validated MACS, in which sequences are clustered into potential functional subgroups. PipeAlign has been implemented in the GScope platform (see section 2.2) for high-throughput processing and is also available for interactive use via the web server at http://bips.u-strasbg.fr/PipeAlign/.

8.4 Other software

8.4.1 Data retrieval

For large sets of sequences, structural and functional information was retrieved from the public databases using the Daedalus system developed in the group by Arnaud Muller. Daedalus allows the combination of personal information from applications such as BlastP (Altschul *et al.*, 1997) or ClustalW (Thompson *et al.*, 1994) with the public databases indexed by SRS (see section 8.1.4). The integration is performed by creating a temporary or 'on-the-fly' database, Daedalus_DB, using the User Owned Databank facility in SRS. This temporary database can then be cross-linked to the other databases, using the SRS indexing system (see figure 8.3).

The structure of the data in the Daedalus_DB and the cross-links to other databases are specified in the SRS programming language, ICARUS (Interpreter of Commands And RecUrsive Syntax). The indexation of Daedalus_DB then allows a simple and efficient access to the user's personal information as well as the information in the public databases. Cross-links can be defined either as direct links between two databases, or as indirect links that take advantage of intermediary databases. For example, the GO Gene Ontology database is accessible from Daedalus_DB via the Swissprot and Sptrembl database links.

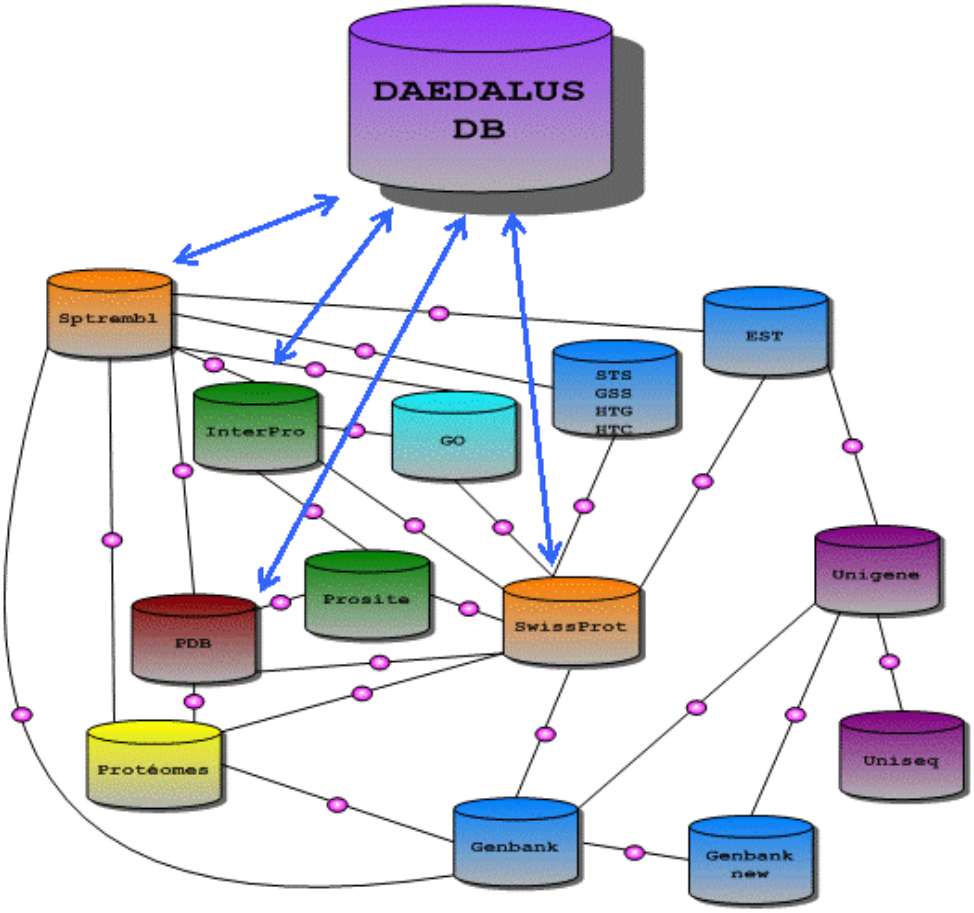


Figure 8.7 Incorporation of the Daedalus_DB temporary database in SRS
Lines with pink dotes indicate SRS cross-links between the different databases. Blue arrows indicate temporary links from Daedalus_DB to standard public databases.

8.4.2 Annotated multiple alignment display

Two complementary programs were used to display multiple alignments with their associated structural and functions annotations.

- *Jalview* (Clamp *et al.*, 1994) is a multiple alignment editor written in Java. The web applet version is used widely in a variety of web pages (e.g. the EBI Clustalw server and the Pfam protein domain database). Jalview allows viewing of multiple alignments of either nucleic acid or protein sequences via a web browser and includes simple editing facilities, tree construction and sequence clustering. Jalview can colour parts of a sequence based on the presence of sequence features, which may be retrieved from database records (such as Uniprot), or may be defined by the user and read from a sequence features file.

- *OrdAlie* (Ordered Alignment Information Explorer) has been developed in the group by Luc Moulinier. OrdAlie is written in Tcl/Tk and is designed to allow interactive analysis and exploration of protein sequence, structure, function and evolution relationships. The multiple sequence alignment is displayed in a graphical window, together with the user-selected sequence features. Sequences can be clustered automatically into sub-families

and a detailed, hierarchical analysis of residue conservation can be performed at the family or sub-family level. Conserved residues can also be visualized in the context of their 3D structural environment, using the RasMol structure viewer (Sayle and Milner-White, 1995).

8.4.3 3D structure superposition and display

Protein 3D structures were superposed automatically using the SAP program (Taylor, 2000). SAP searches for an optimal alignment of two protein structures using dynamic programming (DP). The DP algorithm requires a similarity measure for all pairs of residues, one from each structure to find the optimal alignment. For SAP, the residue similarity is the overlap of the "views" from each of the two residues, where the "view" is the list of distances from the particular residue to all other residues in the same structure. SAP also uses dynamic programming to optimize the overlap of distances in the two distance lists. The procedure is thus dubbed double dynamic programming. Gaps are allowed, but their lengths are limited to improve speed. The native score of SAP is a normalized logarithm of a measure which combines the similarity of the aligned residues (accounting for the length of the alignment) and the number of residues in the smaller protein. In a recent comparison of structure superposition methods (Kolodny *et al.*, 2005), SSAP (a predecessor of SAP) was found to produce the best alignments when using a scoring scheme that emphasised longer alignment length.

Two programs were used for the visualisation of protein 3D structures and the automatic structural superpositions produced by SAP :

- *RasMol* (Sayle and Milner-White, 1995) loads a single structure quickly and directly from standard Brookhaven Protein Data Bank (PDB) files and runs on almost all current computers, including Unix workstations, personal computers running Windows, and Macintoshes. Several different representations are available, including wireframe, spacefill, α-carbon backbone, strands and ribbons. Ligands, active sites, multiple subunits, hydrogen bonds and various parts of the molecule can also be displayed selectively or in a combination of display modes. Once an image has been created, it can be printed directly or translated into a variety of formats for display or alteration by other graphics programs.

- *PyMol* (http://pymol.sourceforge.net) is a molecular graphics system with an embedded Python interpreter designed for real-time visualization and rapid generation of high-quality molecular graphics images and animations. It also runs on almost all current computers, including Unix workstations, personal computers running Windows, and Macintoshes. PyMol can be controlled by commands and can display 3D superpositions of two or more structures, as shown in figure 8.4.

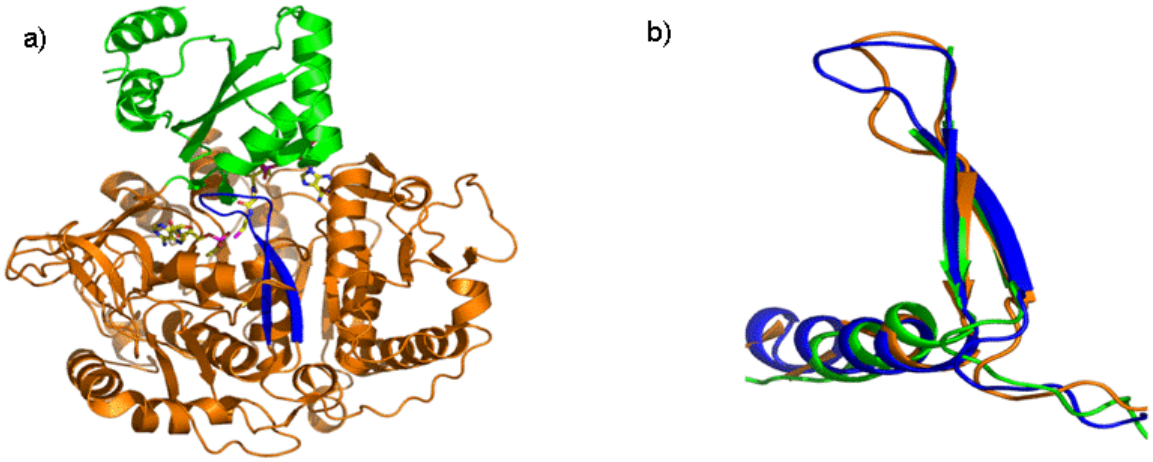


Figure 8.8 3D structure display and superposition with PyMol
Structure of Acetyl coa synthetase bound to CoA and adenosine-5 -propylphosphate (PDB:1pg3). The larger N-terminal domain is shown in brown and the C-terminal domain is shown in green. The putative AMP-binding loop is highlighted in blue. (c) 3D structural superposition of the region containing the AMP-binding signature motif in 1pg3A (blue), 1md9A (brown), 1lci (green). All structure cartoons were produced with Pymol (http://www.pymol.org)..

9 Development of a new multiple alignment benchmark

The first part of this thesis addresses the problem of the assessment of multiple sequence alignment algorithms. We have developed a new version of the BAliBASE benchmark for the objective evaluation and comparison of multiple alignment algorithms, as described in Publication No. 3, included at the end of this chapter.

9.1 Introduction

Multiple sequence alignment is one of the most fundamental tools in molecular biology. It is used not only in evolutionary studies to define the phylogenetic relationships between organisms, but also in numerous other tasks ranging from comparative multiple genome analysis to detailed structural analyses of gene products and the characterisation of the molecular and cellular functions of the protein. The accuracy and reliability of all these applications depend critically on the quality of the underlying multiple alignments. Consequently, a vast array of multiple alignment programs have been developed based on diverse algorithms, from multi-dimensional dynamic programming, via progressive, tree-based programs to the more recent methods combining several complementary algorithms and/or 3D structural information (reviewed in Thompson and Poch, 2006a; see Publication No. 8 in Annex). Comparative evaluation of these different methods has become a crucial task, in order to select the most suitable method for a particular alignment problem (e.g., more efficient, more correct, more scalable); to evaluate the improvements obtained when new methods are introduced; and to identify the strong and weak points of the different algorithms.

In computer science, benchmarking is widely used to compare the performance characteristics of computer systems, compilers, databases and many other technologies. A benchmark is generally made up of two components: a task sample used to compare the performance of alternative tools or techniques and some kind of performance measure that evaluates the fitness for purpose. Within a scientific discipline, a benchmark captures the community consensus on which problems are worthy of study, and determines what are scientifically acceptable solutions. Control of the task sample is used to reduce variability in the results because all tools and techniques are evaluated using the same tasks and experimental materials. Another advantage of benchmarking is that replication is built into the method. Since the materials are designed to be used by different laboratories, people can perform the evaluation on various tools and techniques, repeatedly, if desired. Also, some benchmarks can be automated, so the computer does the work of executing the tests, gathering the data, and producing the performance measures. The resulting evaluations allow developers to determine where they need to improve and to incorporate new features in their programs with the aim of increasing specific aspects of performance. During deployment, the results from different technologies are compared, which requires researchers to look at each other's contributions. Researchers become more aware of one another's work and ties between researchers with similar interests are strengthened. Consequently, the creation and widespread use of a benchmark within a research area is frequently accompanied by rapid technical progress (Sim *et al.*, 2003).

For multiple sequence alignment algorithms, several benchmarks are now available (reviewed in chapter 7.4), whose primary goal is to assess the quality of the different

programs. These benchmarks have been used in the past to compare different multiple alignment programs and have led to significant progress (see chapter 7.5). For example, the comparison study based on BAliBASE (Thompson *et al.*, 1999) showed for the first time, that no single algorithm was capable of producing high quality alignments for all the test cases studied. In particular, the results obtained for the different reference sets in BAliBASE highlighted the complementary of the local and global alignment approaches. As a result, new methods were developed that combined the advantages of the two methods. BAliBASE was identified recently in an independent study (Blackshields *et al.*, 2006) as one of the most useful benchmarks available. The organisation of the alignments into different Reference Sets means that several distinct problem areas are explicitly addressed. In addition, reliable 'core blocks' are defined that exclude the non-superposable regions of the alignment. As a consequence, the behaviour of different alignment programs can be accurately determined with respect to different alignment conditions. Furthermore, it is inherently difficult to over-train methods on this benchmark.

In the post-genomic era, the ever-increasing amount of sequence and structure information available in the public databases means that the size and complexity of the data sets that need to be routinely analyzed are increasing. The alignment benchmarks also need to evolve in order to provide new larger test cases, which are representative of the new alignment requirements.

9.1.1 Criteria for benchmark development

The process of constructing a benchmark implies the rigorous definition of both what is to be measured (for example, the quality of a solution or the time required to produce it) and how it should be measured. A number of requirements for successful benchmarks have been identified previously (Sim *et al.*, 2003), which can be used as design goals when creating a benchmark or as dimensions for evaluating an existing one:

- *Relevance*. The task set out in the benchmark should be representative of ones that the system is reasonably expected to handle in a natural (*i.e.* not artificial) setting and the performance measure used should be pertinent to the comparisons being made.

- *Solvability*. It should be possible to complete the task sample and to produce a good solution. A task that is too difficult for all or most tools yields little data to support comparisons. A task that is achievable, but not trivial, provides an opportunity for systems to show their capabilities and their shortcomings.

- *Scalability*. The benchmark tasks should scale to work with tools or techniques at different levels of maturity. This property influences the size of task: it should be sufficiently large to showcase the more mature techniques, but not too large to test techniques currently being researched.

- *Accessibility*. The benchmark needs to be easy to obtain and easy to use. The test materials and results need to be publicly available, so that anyone can apply the benchmark to a tool or techniques and compare their results with others.

- *Evolution*. Continued evolution of the benchmark is necessary to prevent researchers from making changes to optimise the performance of their contributions on a particular set of

tests. Too much effort spent on such optimisations indicates stagnation, suggesting that the benchmark should be changed or replaced.

Benchmarks that are designed according to these conditions will lead to a number of benefits, including a stronger consensus on the community's research goals, greater collaboration between laboratories, more rigorous examination of research results, and faster technical progress.

9.2 BAliBASE multiple alignment benchmark

There are two main issues involved in the definition of a multiple alignment benchmark. First, what is the 'correct' alignment of the sequences included in the tests? Second, which alignment problems should be represented in the benchmark, and how many test cases are needed? These two problems are discussed in detail below.

9.2.1 Definition of the correct alignment

The goal of a multiple sequence alignment is to identify equivalent residues in nucleic acid or protein molecules that have evolved from a common ancestor. However, the true evolutionary history cannot usually be reconstructed. Therefore, reference sequence alignments are generally constructed based on comparisons of the corresponding 3D structures. For proteins, the 3D structure is generally more conserved than the sequence and a reliable structural superposition is normally possible between very divergent proteins sharing little sequence identity (Koehl, 2001). Structural superposition is carried out between two known structures, and is typically based on the Euclidean distance between corresponding residues, instead of the distance between amino acid "types" used in sequence alignment. Thus, the structure alignment can provide an objective reference that is built independently of the sequences. For the new version of BAliBASE, we chose the SAP structural alignment program (Taylor, 2000), based on a number of reliability/functionality criteria. Firstly, SAP is a reliable program, derived from the SSAP method, which produced the best alignments with a longer match length in a recent study (Kolodny *et al.*, 2005). Secondly, SAP is available for local installation and has a command-line interface, facilitating its integration in an automatic protocol. Thirdly, the SAP program provides a sequence alignment based on the structural superposition, together with a reliability score for each pair of aligned residues. Although SAP produces high quality alignments for many cases, errors can still occur when aligning very distantly related protein structures. Therefore, the SAP sequence alignment was verified manually and corrected to ensure that annotated functional residues were aligned correctly.

Another issue is the most suitable quantitative score to use to compare an alignment obtained with a multiple alignment program with the reference alignment. For BAliBASE, we have defined two scores, reflecting different properties (defined in Thompson *et al.*, 1999b). The sum-of-pairs score is the percentage of correctly aligned pairs of residues in the alignment produced by the program. It is used to determine the extent to which the programs succeed in aligning some, if not all, of the sequences in an alignment. The column score is the percentage of correctly aligned columns in the alignment, which tests the ability of the programs to align all of the sequences correctly.

9.2.2 Selection of alignment test cases

A multiple alignment benchmark does not need to include all possible alignments. It is sufficient to provide enough representative tests, in order to be able to differentiate between alignment methods. The benchmark should however include as many different types of proteins as possible. For BAliBASE version 3, the most complete source of protein 3D structures is the PDB database. However, this set contains a certain amount of bias due to over-represented structures (Brenner *et al.*, 1997). Therefore, we decided to use the SCOP protein classification database as a structure resource in order to include representative protein families from as many different structural fold types as possible. Protein domains in SCOP are hierarchically classified into families, superfamilies, folds and classes (see Table 9.1).

Class	Number of folds	Number of superfamilies	Number of families
All alpha proteins	218	376	608
All beta proteins	144	290	560
Alpha and beta proteins (a/b)	136	222	629
Alpha and beta proteins (a+b)	279	409	717
Multi-domain proteins	46	46	61
Membrane and cell surface proteins	47	88	99
Small proteins	75	108	171
Total	945	1539	2845

Table 9.1 SCOP classification statistics
Release 1.6, from http://scop.mrc-lmb.cam.ac.uk/scop/.

Example folds were selected for inclusion in BAliBASE from each of the five main classes (excluding membrane and small proteins), provided that at least four different protein structures were available. If too few sequences existed with known structures, the reference alignment was augmented with sequences from the Uniprot database, whose 3D structure is not yet known. In this way, we were able to construct larger alignments, representing different alignment problems, such as divergent sequences, orphans, large N/C terminal extensions or internal insertions, etc. For each test case, two different reference alignments were constructed. The alignment of homologous regions only is widely used in the construction of protein domain databases, while the alignment of full-length, complex sequences, such as those detected by the database searches, is routinely performed in automatic, high throughput genome analysis projects.

In the face of the increased number of protein families and the number of sequences included in each family alignment, manual construction of the reference alignments was no longer possible. We therefore decided to automate as many steps as possible in the development of BAliBASE version 3, including the search for homologous proteins, the 3D structure superposition, the definition of the core blocks and the integration of structure/functional information for alignment annotation and display. This semi-automatic protocol allowed us to increase both the number of alignments and the number of sequences in the latest version, compared to previous releases of the database. Table 9.2 shows the size of the new benchmark. Reference 1 contains alignments of equidistant sequences and is divided into six subsets, according to three different sequence lengths and two levels of sequence variability. Reference 2 contains families aligned with one or more highly divergent

"orphan" sequences, Reference 3 contains divergent subfamilies, Reference 4 contains sequences with large N/C-terminal extensions, and Reference 5 contains sequences with large internal insertions.

Reference 1	Small number of equi-distant sequences			
	short	medium	long	sub-total
• V1 (<20% identity)	14	12	12	38
• V2 (20-40% identity	14	16	15	45
Reference 2	Family with one or more 'orphan' sequences			41
Reference 3	Divergent subfamilies			30
Reference 4	Large N/C terminal extensions			48
Reference 5	Large internal insertions			16
Total				217

Table 9.2 Number of test cases in version 3 of the BAliBASE alignment benchmark

Reference Sets 6-8, containing transmembrane sequences, repeats and circular permutations, have been maintained in this version, although they have not been updated.

9.3 Comparison of the latest alignment methods with BAliBASE 3.0

We have used the new version of BAliBASE to evaluate and compare some of the most recent multiple sequence alignment programs together with a selection of the more traditional methods, namely ClustalW (a global algorithm) and Dialign (a local algorithm). The goal of the comparison was not to determine which program is the 'best' for all alignments, but to measure the improvement in quality obtained by the more recently developed programs and to identify their strong and weak points. The programs included in this study are shown in Table 9.2.

Program	**Reference**	**Version**	**Features**
ClustalW	Thompson *et al.*, 1994	1.83	Global, progressive alignment
Dialign	Morgenstein *et al.*, 1996	2.2.1	Local alignment of sequence segments
MAFFT	Katoh *et al.*, 2002	5.32	Local anchors and global, progressive alignment
MAFFTi			MAFFT with iterative refinement
MUSCLE	Edgar, 2004	3.51	k-mer counting and global, progressive alignment with iterative refinement
TCoffee	Notredame *et al.*, 2000	2.66	Local and global pairwise alignment consistency scores in global progressive alignment
PROBCONS	Do *et al.*, 2005	1.1	Global, iterative alignment with HMM-derived posterior probabilities and local alignment consistency information

Table 9.3 Multiple alignment programs compared using BAliBASE 3.0
All programs were run with default parameters.

Table 9.3 shows the scores obtained by the different methods for the alignments containing only the homologous regions.

	Reference 1: Equidistant sequences		Reference 2: Family with orphans	Reference 3: Divergent subfamilies	Reference 4: Large extensions	Reference 5: Large insertions	Time (sec)
	V1:<20%	V2:20-40%					
ClustalW	0.63/0.42	0.90/0.78	0.91/0.42	0.76/0.52	0.75/0.41	0.75/0.38	902
Dialign	0.50/0.31	0.86/0.71	0.89/0.37	0.70/0.39	0.79/0.45	0.78/0.43	6043
Mafft	0.64/0.44	0.89/0.78	0.93/0.49	0.79/0.53	0.83/0.47	0.83/0.48	96
Maffti	0.71/0.54	0.91/0.83	0.94/0.55	0.84/0.60	0.85/0.49	0.87/0.57	327
Muscle	0.71/0.52	0.91/0.82	0.94/0.50	0.85/0.58	0.84/0.46	0.86/0.54	523
TCoffee	0.67/0.47	0.93/0.84	0.94/0.50	0.84/0.64	**0.87/0.54**	0.88/0.58	46335
Probcons	**0.79/0.63**	**0.94/0.89**	**0.95/0.60**	**0.87/0.65**	0.86/0.54	**0.90/0.63**	19035

Table 9.4 Scores for BAliBASE reference sets containing alignments of homologous regions only
The scores shown in each column are sum of pairs/column scores. For each reference set, the highest scores obtained by the different programs is shown in bold.

In all the reference tests, there is a significant difference between the traditional methods (ClustalW and Dialign) and the most recent developments. Nevertheless, in Reference 1, for all the programs tested, a decrease in accuracy of the alignments with decreasing residue identity is clearly demonstrated, with a significant difference between V2 (20-40% identity) and V1 (<20% identity), which corresponds to the 'twilight zone' of evolutionary relatedness.

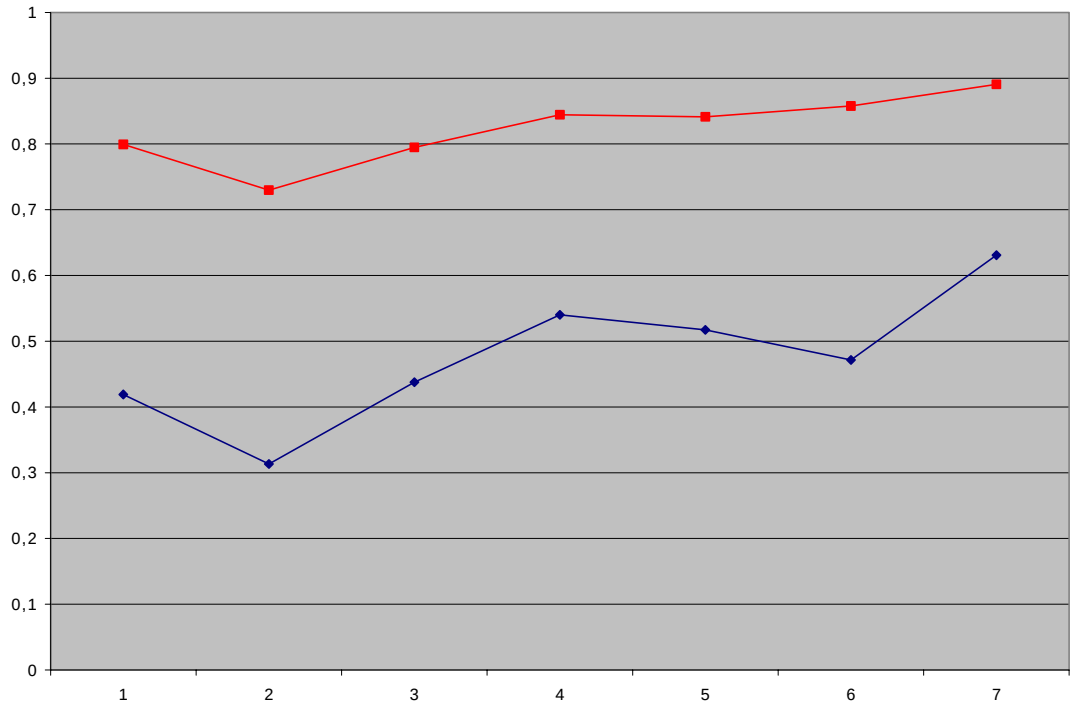


Figure 9.1 Mean column scores for the programs in Reference 1, V1 and V2

Of the two traditional methods, the global alignment program, ClustalW, performs better for the tests involving sequences of similar length in Reference Sets 1-3, while the local alignment program, Dialign, is more successful in Reference Sets 4-5, containing large N/C terminal extensions or internal insertions. This result confirms the observations made previously (e.g. Thompson *et al.*, 1999b; Blackshields *et al.*, 2005). By combining different complementary approaches, the more recent programs, TCoffee, Mafft, Muscle and Probcons

are more reliable in all the tests. The best alignments in all the reference tests were achieved by PROBCONS, although a significant time penalty was incurred.

	Reference 1: Equidistant sequences		Reference 2: Family with orphans	Reference 3: Divergent subfamilies	Reference 5: Large insertions	Time (sec)
	V1:<20%	V2:20-40%				
ClustalW	0.46/0.24	0.85/0.72	0.86/0.20	0.62/0.27	0.61/0.34	2227
Dialign	0.47/0.26	0.85/0.70	0.85/0.29	0.64/0.31	0.77/0.42	12595
Mafft	0.45/0.25	0.88/0.75	0.88/0.35	0.74/0.38	0.79/0.43	312
Maffti	0.57/0.35	0.90/0.80	0.88/0.40	0.78/0.50	0.84/0.53	1409
Muscle	0.56/0.34	0.90/0.79	0.88/0.36	0.76/0.39	0.83/0.46	3608
TCoffee	0.59/0.35	0.92/0.82	**0.91/0.40**	0.75/0.49	0.87/0.57	156373
Probcons	**0.65/0.43**	**0.93/0.86**	0.90/0.41	**0.79/0.54**	**0.88/0.57**	58488

Table 9.5 Scores for BAliBASE reference sets containing alignments of full length sequences
The scores shown in each column are sum of pairs/column scores. For each reference set, the highest scores obtained by the different programs is shown in bold. Reference 4 is not included in this test, because of the nature of the alignments.

Table 9.4 shows the scores obtained by the different programs when the full length sequences are aligned, instead of just the homologous regions. Reference 4 (large N/C terminal extensions), is excluded from this test because the full length alignment is the same as in the previous test. By comparing the scores obtained here with the scores in Table 9.3, it is clear that the inclusion of 'noise', in the form of non-homologous regions, represents a serious problem for all the programs tested, although the difference is less for the more related sequences in Reference 1, V2. Figure 9.2 shows the results for each reference set obtained by the best scoring program, PROBCONS, for the full length alignments compared to the homologous regions only.

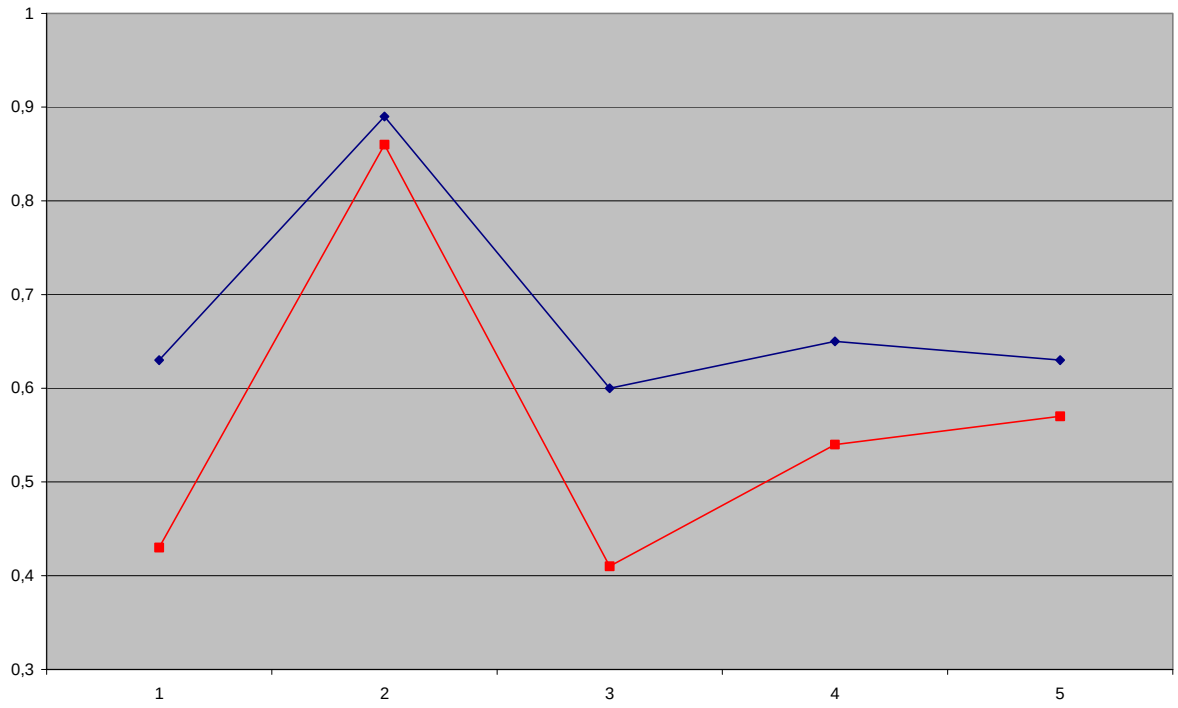


Figure 9.2 Comparison of alignment scores for full-length sequences versus homologous regions only

Mean column scores for PROBCONS alignments of full-length sequences (shown in red) compared to the scores obtained for the same alignments when only the homologous regions are included (shown in blue).

Again, PROBCONS produces the most accurate alignments, but requires significantly more CPU time than most of the other methods.

9.4 Conclusions

The latest version of BAliBASE includes a number of important developments that should significantly increase the utility of the benchmark:

- *Increased size.* The tests in the benchmark are designed to represent the tasks that multiple alignment tools or techniques are now expected to solve in the postgenomic era. The semi-automatic update protocol allows the construction of larger test cases, although manual refinement is still essential to maintain the high-quality of the reference alignments. The comparison of alignment programs has shown that the difference between the scores obtained by the different methods is statistically significant. We thus conclude that BAliBASE version 3 currently contains a sufficiently large number of tests, although further updates will undoubtedly be required in the future.

- *Increased complexity.* The field of multiple sequence alignment is evolving rapidly, with the development of new, more sophisticated algorithms designed to cope with the large amounts of complex information now available in the protein sequence and 3D structure databases. New multiple alignment benchmarks are now required to keep up with these developments and to avoid optimization of the tools on a particular set of tasks. Therefore, the complexity of the alignments in version 3 has been significantly increased with the addition of alignments containing full-length sequences for all the Reference Sets. These full-length alignments provide a large number of difficult tests for both global and local alignment algorithms.

- *Increased accessibility*. The benchmark needs to be easy to obtain and to use, otherwise few people will be likely to use it. Therefore, all BAliBASE alignments and associated annotations are freely available on the WWW or by ftp. Furthermore, the alignments and their associated annotations are now available in a standard data exchange XML format that should facilitate the development of automatic procedures for the evaluation and comparison of new multiple alignment methods.

The BAliBASE benchmark is designed to evaluate alignment quality and efficiency, but the criteria for selection of a program are numerous, including ease-of-use, stability, robustness, etc. Nevertheless, objective evaluation of new algorithms should lead to more robust multiple alignments, which in turn will lead to more reliable results for the many applications that rely on multiple alignments.

Publication n^{o}. 1

10 MAO: Multiple Alignment Ontology

This chapter concerns the creation of a new ontology for multiple alignments of nucleic acid and protein sequence/structure alignments. The design and development of MAO are described in Publication No. 2, included at the end of this chapter.

10.1 Introduction

High-throughput genome sequencing and assembly techniques, together with new information resources, such as structural proteomics, interactomics, transcriptome data from microarray analyses, or light microscopy images of living cells have lead to a rapid increase in the amount of data available. As a result, there now exists a vast array of heterogeneous data resources distributed over different Internet sites that cover genomic, cellular, structure, phenotype and other types of biologically relevant information. The complexity of the molecular biology domain makes the modelling, handling and exchange of data very difficult. This complexity is reflected in some of its common characteristics (Rojas *et al.*, 2003):

- *Ambiguous terms*. There are no strict definitions of the terms used to describe objects and their properties. Examples are the differences in the definitions of an operon or, most commonly cited, of a gene (Schulze-Kremer, 2002).

- *Multiple types of classifications*. It is common to find the same object being classified in multiple ways, e.g. sequence and structure protein domain classifications do not always have the same boundaries. The classification criteria, however, are not always explicitly presented or defined.

- *Multiple relationships*. Biological objects are frequently connected in various ways with each other, forming a highly interconnected graph of relationships. Consider the case of a biochemical reaction and the compounds that act as substrates in the reaction. This is clearly a many-to-many relationship between the reaction and the compound entity (a compound can act as a substrate for more than one reaction, and a reaction often has more than one substrate).

- *Missing information*. The improvement of experimental techniques is dramatically accelerating the accumulation of data in molecular biology. Nevertheless, there are many concepts which are important to the domain where little data is currently available. One example is the connection between the concepts of genome, chromosome, operon and gene. Genes could be identified within an organism, even though the genome has not been completely sequenced. Even if the corresponding genome is known, it might still be unknown whether this gene belongs to an operon.

- *Evolution of concepts and classifications*. As the amount of data in the domains of biochemistry and molecular biology is constantly growing, the concepts and classifications modelling the biological 'reality' has to be adapted. An example could be the concept of 'bird', which at a first glance could be defined as an 'animal that can fly'. But after a while one will notice that there are birds that cannot fly (like penguins) and animals that can fly but are no birds (like some insects). So one could redefine the concept of 'bird' as an 'animal that has wings, has feathers and lays eggs'.

- *Artificial relationships*. It is often the case that relationships between concepts are defined although they do not exist in reality. For example, there are relationships between DNA and RNA defined by the transcription process and between RNA and protein defined by the translation process, but there is no direct relationship between DNA and protein. However, direct relationships are needed between DNA and protein in computational experiments that predict protein sequences from DNA sequences.

- *Different levels of granularity*. Processes and their elements can commonly be described on different levels. An example is the description of a signal transduction pathway, which can refer to a general, organism-independent level, mentioning the protein families which participate; it may also describe this pathway at the level of certain members of a family; or at the level of particular proteins in particular organisms.

In the light of this complexity, controlled vocabularies and ontologies have become essential tools in modern molecular biology. They ensure compatibility between different data resources and software applications and increase the efficiency and accuracy of data queries by standardising the wide variations in terminology that exists in the biological sciences. Such a common framework can also facilitate the exchange, integration and validation of information. Also, the formalisation of a given sub-domain by means of an ontology allows knowledge to be expressed in a computer readable way. Furthermore, the fact that most of the formalisms used to describe ontologies have associated induction engines, will allow the induction of knowledge from the concepts, relationships and rules included in the ontology. The concepts and relationships of an ontology can be used as building blocks for the formulation of hypotheses that can be verified or rejected, either by experimentation or by using data integrated from different/multiple sources, including text based ones.

10.2 Design of the Multiple Alignment Ontology

The information included into an ontology strongly depends on the uses that are given to it. Although in principle an ontology should reflect facts of a given domain or sub-domain, pragmatically speaking the construction of the ontology is mainly guided by the intended need, meaning that the detail at which certain properties or relationships are specified are strongly influenced by the intended use or research interests. The MAO multiple alignment ontology is designed to improve interoperation and data sharing between different alignment protocols for the construction of a high quality, reliable multiple alignment and to facilitate the integration of structural and functional information in the context of the nucleic acid or protein family. The top-level concept is called the multiple_sequence_alignment, which may represent either nucleotide or protein sequences. Most of the basic features associated with multiple alignments are defined as MAO concepts, ranging from a single residue to sub-families of sequences. Attributes associated with the basic concepts allow the definition of more complex information, such as column conservation, residue or motif function, or 3D structural information.

An ontology should also contain agreed definitions, reflecting knowledge in the community. MAO was developed in collaboration with experts in RNA, protein alignment of sequences and structures. The multiple alignment ontology was established in close

collaboration with domain experts from both the DNA/RNA and protein communities, including specialists in the fields of both primary sequence and 2D/3D structure comparisons.

10.2.1 Ontology representation

There are three common approaches to representing heterogeneous biological data (reviewed in chapter 3). The first relies on hierarchical models, the second uses frames and the third description logic (DL). The most prominent example of a hierarchical model in the bio-sciences is undoubtedly the GeneOntology (GO). GO's controlled vocabulary has become a standard reference for databases and biological systems and is used both to extend the information about the related object as well as to order the related object under the GO ontology. Frames and DL describe more complex relationships and concepts in a formal framework, allowing automatic reasoning and inference. However, the development of such ontologies is a complex and time-consuming work and the associated reasoning and knowledge inference systems are not yet well established.

For MAO, a hierarchical model was considered sufficient for the intended purpose of the ontology. Specifically, MAO is organized as a complex hierarchy, known as a directed acyclic graph (DAG), where the nodes in the graph represent concepts and the branches joining the nodes represent relationships. DAGs can be considered to be a generalization of trees in which child nodes (more specialized terms) may have multiple parents (less specialized terms) and multiple relationships to their parents.

Another crucial factor in the design of MAO was the ability to integrate different information from a wide variety of different sources. Much as this information is now represented by ontologies that are available on the Open Biomedical Ontology (OBO) web site (http://obo.sourceforge.net/). OBO is a collaborative project for structured vocabularies and ontologies for use within the genomics and proteomics communities. For an ontology to be accepted as part of the OBO project, it must meet a number of requirements. The requirements for acceptance on the OBO site include:

- The ontology should be open and available to be used by all without constraint
- The ontology should be in a common formal language, either the OBO format or OWL
- The ontology should possess a unique identifier within OBO
- The ontologies should be orthogonal to other ontologies already lodged within OBO
- The ontology should include textual definitions for all terms

These acceptance criteria ensure coherence and facilitate the interoperation of the different ontologies using the same software tools. For example, the Ontology Lookup Service (Côté *et al.*, 2006) has been developed that integrates the OBO ontologies into a single database and provides a web service interface to obtain information about multiple ontologies. For the MAO ontology, we therefore decided to use the OBO ontology language, thus ensuring compatibility with a wide range of biomedical ontologies, including the Gene Ontology (GO), the Sequence Ontology (SO), protein-protein interaction data (PSI), and taxonomic information via the NCBI organismal classification.

10.2.2 Ontology construction

Numerous tools have been developed to aid the ontologist in the construction and maintenance of ontologies. These ontology editors have been compared previously in terms of availability, functionality, visualisation and input and output formats, among other criteria (Denny, 2002; Lambrix *et al.*, 2003). All the systems tested had particular strengths and weaknesses and no tool was found to be superior in all aspects.

The OBO-Edit tool (previously known as DAG-Edit) is the only editor capable of reading and writing ontologies in the OBO format, and. was therefore the clear choice for the construction of MAO. OBO-Edit is an open source, platform-independent application for viewing and editing OBO ontologies (http://sourceforge.net/projects/geneontology). Its emphasis on the overall graph structure of an ontology provides a friendly interface for biologists and makes OBO-Edit excellent for the rapid generation of large ontologies focusing on relationships between relatively simple classes.

10.3 Conclusions

The use of ontologies in all its forms; controlled vocabularies, taxonomies or more formal conceptual models, is contributing to the formal description of different aspects of biochemistry and molecular biology. Each ontology tends to focus on a certain aspect or sub-domain, thus the combination of ontologies can provide a more extended domain description. The knowledge representation encoded in MAO, together with the other OBO ontologies, can be used to facilitate data sharing between different programs. For example, the PipeAlign toolkit (Plewniak *et al.*, 2003) uses an XML format based on MAO to transfer information between the different steps in the alignment process. MAO also allows the annotation of multiple sequence alignments with structural and functional information, for example in BAliBASE version 3. It provides the basis for data integration, validation and analysis in a MACS-based information management system, that will be introduced in the next chapter.

Publication n^{o}. 2

11 MACS-based information management system

This chapter describes the development of a new system for the retrieval, organisation and analysis of all the information associated with gene families, based on Multiple Alignments of Complete Sequences (MACS). The MACS Information Management System (MACSIMS) is described in Publication No. 3, included at the end of the chapter.

11.1 Introduction

Recent experimental developments have made available new genome-wide sequence and functional datasets. The size and complexity of these sets have created substantial data management and analysis challenges. The datasets being produced are much larger than biologists have traditionally dealt with, and so the data must be stored in a manner that makes them amenable to computational analysis. Information management systems are now needed to successfully exploit this wealth of data. This data is heterogeneous, can be stored in various flat file formats, relational databases etc., and is geographically distributed. The data is complex and error-prone. For example, in the case of large-scale protein interaction screens, using yeast-two hybrid or affinity purification of complexes, a significant percentage of reported interactions may be false positives (von Mering *et al.*, 2002). Much of the data available are in fact computer predictions, with their inherent reliability. For example, most protein sequences are predicted from complete genome sequences by programs such as Glimmer etc. Many of the sequences contain errors (Bianchetti *et al.*, 2005). Intelligent systems are needed to store, validate and transform the data into useful information (see chapter 4).

In this context, multiple alignments of molecular sequences represent an ideal basis for the reliable integration of information, ranging from complete genomes to a gene and its related products (Woese *et al.*, 1993; Lecompte *et al.*, 2001). By placing the sequence in the framework of the overall family, multiple alignments can be used to identify important structural or functional motifs that have been conserved through evolution, and also to highlight particular non-conserved features resulting from specific events or perturbations. The goal of MACSIMS is to collect data and to generate information and knowledge about sequence/structure/function/evolution relationships.

11.2 Design of MACSIMS

11.2.1 Data storage and retrieval

There are two main approaches used in bioinformatics for the storage of heterogeneous data. The different databases can be installed locally on the user's own computer system, using a unified format. The advantage of this so-called 'data warehousing' approach is that the subsequent data retrieval is a relatively simple process. The disadvantage is that the operating costs can be very heavy in terms of the hardware required for database installation and maintenance. The alternative to local installation of the databases required for a particular application is to access the original data source remotely over the Internet. This approach reduces considerably the overheads required for local storage of the databases. However,

remote access requires complex systems to manage communication between the server and the client, particularly when errors occur because the remote systems are not available.

In the Laboratoire de Biologie et Genomique Structurales (LBGS), we have local access to many generalist sequence and structure databases, maintained by the Plate-forme Bio-informatique de Strasbourg (BIPS). In MACSIMS, we have used the Sequence Retrieval System (SRS) (Etzold and Argos, 1993) as a unified front end to independently access these databases. SRS is arguably the most widely used database query and navigation system for the life science community. It provides a single interface for most general biological databases, and allows a fast access via the creation of on-the-fly databases for large sets of sequences. Nevertheless, with the growing number of specialist databases, an alternative data retrieval system will be required to access new data resources that are not incorporated in SRS. For example, mutation data, protein-protein interactions and network/pathway information will be accessed remotely in future versions of MACSIMS. Another abundant data source that could be exploited is the scientific literature, thanks to the development of new methods and tools for literature-mining (Jensen *et al.*, 2006).

11.2.2 Data model

The efficient integration and exploitation of complex heterogeneous data requires a formal data model that describes the data types used and the relationships that exist between different data types, in a format that can be understood by the computer. MACSIMS is based on the data model embodied in the MAO ontology, described in chapter 10. MAO facilitates the integration of sequence, structure and function information by providing a unified representation of the concepts defined in the different domains. In the context of the multiple alignment, the data retrieved from different resources can be compared, validated and propagated from the known to unknown sequences. The MAO ontology also provides facilities for supplementing the verified information with the results of computer predictions. In this case, attributes are assigned to the predicted data concepts that describe the algorithm used to produce the prediction. This is important for automatic annotation and analysis systems, where the theoretical evidence for a given prediction provides an indication of its reliability.

11.2.3 Data visualisation

An important aspect of information management systems is the possibility for browsing and searching of the stored data and for interactive exploration and manipulation to facilitate analysis from multiple perspectives. In MACSIMS, the retrieved data and the results of the subsequent analyses are output in an XML format file that is used for high-throughput or automatic processing. In addition, a graphical, web-based user interface is provided via the JalView multiple alignment editor (Clamp *et al.*, 1994). JalView is a well-established, robust system that includes many features, such as a global multiple alignment overview, sequence annotation features, phylogenetic trees, etc.

In addition, the XML format file created by MACSIMS can be input to the OrdAlie, Ordered Alignment Information Explorer (L. Moulinier, manuscript in preparation) for more in-depth analyses of residue conservation. OrdAlie highlights residues that are conserved at the family or the sub-family levels, that are generally important for the structure or the

function of the protein. In addition, differentially conserved positions in the alignment are identified, i.e. positions that are conserved in more than 90% of one functional group and *strictly* absent in the other. The differentially conserved residues are often involved in sub-family specific functionalities. For example, OrdAlie was used to perform a sequence analysis of the ligand-binding domain of nuclear receptors (see section 5.2.5), revealing two sets of differentially conserved residues, which partitioned the entire nuclear receptor superfamily into two classes related to their oligomeric behaviour (Brelivet *et al.,* 2004).

11.2.4 *Ab initio* predictions

For sequences with known homologues, information can be propagated reliably. Annotation of orphan sequences by structure, prediction of transmembrane etc., core block regions.

11.3 MACSIMS applications

The potential applications of MACSIMS are numerous, but will include such fields as the automatic annotation of the ever-increasing number of hypothetical proteins being produced by the high-throughput genome sequencing projects or the definition of characteristic motifs for specific protein folds. In the LBGI, the integrative power of MACSIMS has been used in a number of projects where the detailed analysis of protein families represents a crucial component.

11.3.1 Validation of predicted protein sequences

Although high-throughput sequencing of genomes and HTC are producing an avalanche of data, the quality of the sequences in the public databases depends on the technique used to produce them. RNA sequences consist of high quality individually cloned cDNA, expressed sequence tags (EST) containing a 1% base error rate, or high-throughput cDNA (HTC) or more variable quality (Benson *et al*., 2006). Protein sequences are either translated from functionally cloned cDNA or HTC sequences, or are the result of computational predictions. Proteins derived from the conceptual translation (using the genetic code) of functionally cloned cDNA are generally of high quality, although some represent fragments of full-length proteins. Proteins translated from HTC sequences need to be verified. But the majority of the proteins in the public sequence databases are now predicted *in silico* from prokaryotic/eukaryotic genomes and it has become evident that such programs may produce invalid data. In prokaryotes, translation start site prediction is reported to be unsatisfactory (Hannenhalli *et al*., 1999) and in eukaryotes, the automatic determination of precise exon-intron boundaries remains an unsolved problem (Mathe *et al*., 2002).

We have developed a program, vALId (Bianchetti *et al*., 2005), that exploits the information content of MACSIMS to verify the quality of the sequences in a multiple alignment. The vALId method is described in detail in Publication No. 4 in Annex 1. The first step in the validation process is to determine the coding origin of the sequences, defined as either ‘complete cDNA’, HTC or ‘predicted’. In the next step, predicted proteins and translated HTC are analyzed according to their phylogenetic context. For each sequence, column conservation scores are compared at three different levels, namely the complete alignment, the sub-group or ‘sub-alignment’ and the closest neighbour (see figure 11.1).

Phylogenetic consistency implies that a given sequence should be increasing divergent when compared to its closest neighbour, to its sub-alignment and to the complete alignment. If a sequence segment satisfies these criteria it is considered to be reliable; otherwise, it is annotated as a suspicious insertion or divergent segment.

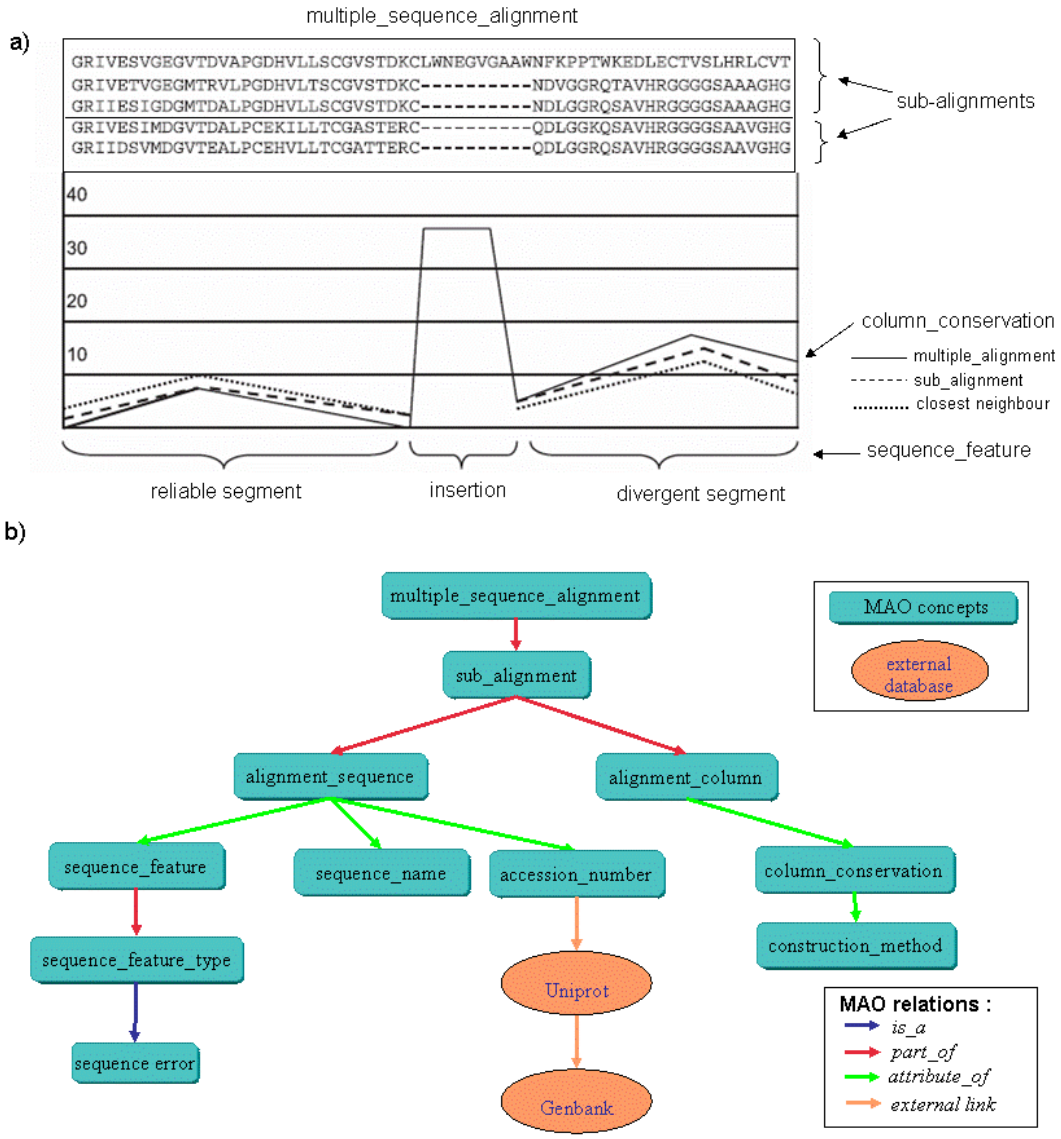


Figure 11.1 vALId determination of reliable sequence segments and detection of potential errors
a) Comparison of conservation scores for the complete multiple alignment, the sub-alignment and the nearest neighbour. b) Integration of the information by MACSIMS, based on the MAO ontology.

In the final "correction" step, vALId exploits the MACSIMS links to the external databases, Uniprot and Genbank, to extract transcriptional and genomic data that is then used to propose corrections of the delineated inconsistent regions to enhance protein quality. All the information generated by vALId can be integrated in a single system using MACSIMS (see figure 11.1b) and made available to the user via an interactive web-based interface. An example vALId analysis and the web server display are shown in section 5.2.3.

The accuracy of the vALId predictions was evaluated using a large scale test set consisting of 100 MACS automatically generated by PipeAlign (Plewniak *et al.,* 2003). Of the 6141 proteins in the test set, 65% were computational predictions and 3% were translated HTC. Although most predictions contained at least one reliable segment, 44% of the eukaryote predicted sequences contained at least one potential error. The evaluation also revealed an unexpectedly high number of inconsistent regions in HTC, with 31% of them containing suspicious regions. By identifying the reliable and unreliable segments in the sequences, vALId can improve the accuracy of subsequent analyses for the detection of conserved structural/functional motifs or the identification of non-conserved features resulting from specific events or perturbations.

11.3.2 Protein function annotation using the Gene Ontology

MACSIMS is also used in the GoAnno web server (Chalmel *et al.*, 2005) to reliably predict protein function based on the hierarchical and standardized vocabulary provided by the Gene Ontology (GO) (Gene Ontology Consortium, 2000). As described in Publication No. 5 in Annex 1, GOAnno takes a query protein as input and constructs a MACS using the PipeAlign system (Plewniak *et al.*, 2003). MACSIMS is then used to retrieve the GO annotations for each sequence in the query subfamily, which are then propagated according to a certain number of criteria (defined in detail in the publication in Annex 1). The GOAnno system was used to study the mechanisms leading to retinal degeneration, on microarray experiments to analyze 1046 proteins (Abou-Sleymane *et al*., 2006). Given the large number of transcripts showing altered level of expression in R7E and R6/2 retina, we used a systematic approach to interpret the biological significance of these gene deregulations. First, we re-annotated each transcript by analyzing the corresponding Affymetrix probe set sequences, based on BLAST homology searches in the public databases. GOAnno was then used to predict Gene Ontology annotations for each identified protein. Of the 727 deregulated transcripts, 541 were assigned to GO biological process terms. Disease onset in R7E was associated with under-expressed genes significantly enriched in signal transduction, cell communication and most importantly visual perception. Enrichment of down-regulated genes involved in visual perception makes perfect sense with the early and progressive ERG defect in R7E retinopathy and by itself validates the method that we used to identify deregulated pathways.

11.3.3 Target characterisation for structural proteomics

MACSIMS has been used to select and characterise potential targets in the Structural Proteomics in Europe (SPINE) project (Albeck *et al.*, 2006). SPINE is an integrated research project to develop new methods and technologies for high-throughput structural biology. Bioinformatics plays an important role in SPINE, for target selection and analysis, in the development of laboratory information management systems and in the dissemination of the results of SPINE activities. The bioinformatics developments for the SPINE project are described in detail in Publication No. 6 in Annex 1.

The SPINE target list included a wide range of proteins from all domains of life, ranging from virus and bacteria to eukaryotic targets of potential pharmaceutical interest. In order to fully understand the potential biomedical role of a target protein, such diverse data as the type of organism, domain organisation, splicing variants, 2D/3D structures and mutations and their associated illnesses, must be organised into an information network for presentation to the experimentalist. The Gscope platform was therefore used to perform PipeAlign and

MACSIMS analyses of all targets in the SPINE target database and an "identity card" was created for each potential target. These identity cards include lists of full-length and partial protein homologues, similar 3D structures that provide templates for homology modelling, and domain boundaries used for defining protein constructs. The identity cards for each target are available for all SPINE members, as well as the general public, via the Project web site at http://www.spineurope.org/.

11.3.4 Prediction of structural/functional effects of mutations

The MS2PH (Structural Mutation to Human Pathologies Phenotype) project uses the information provided by MACSIMS to facilitate the analysis of proteins involved in human genetic disease and the identification of mutations that cause structural or functional perturbations. In the context of this project, a web server has been developed (described in Publication No. 7 in Annex 1) that combines automated protein modelling with the creation of a hierarchical and annotated MACS (Garnier *et al.*, 2006). The MAGOS server is designed to allow in-depth structural, functional and evolutionary analyses of protein families. The structure-to-function relationship can be directly addressed through three-dimensional (3D) structure determination, while the sequence-to-function relationship can be understood through the analysis of conserved patterns and evolution of protein organization mainly based on amino acid sequence comparisons in the context of the multiple alignments. MAGOS accepts a single protein sequence as input and incorporates four main steps: (i) a high quality MACS is first computed using the PipeAlign system (Plewniak *et al.*, 2003), (ii) the validated MACS is annotated using MACSIMS and at the same time, the aligned proteins are characterized according to their homology with proteins implicated in human genetic diseases, (iii) the query protein is modelled using Geno3D (Combet *et al.*, 2002), whose main advantage is the ability to generate homology 3D structure models at a low rate of identity, (iv) the final step is the retrieval of all computed results and their interconnection via a user-friendly web interface based on the Jmol applet (http://jmol.sourceforge.net).

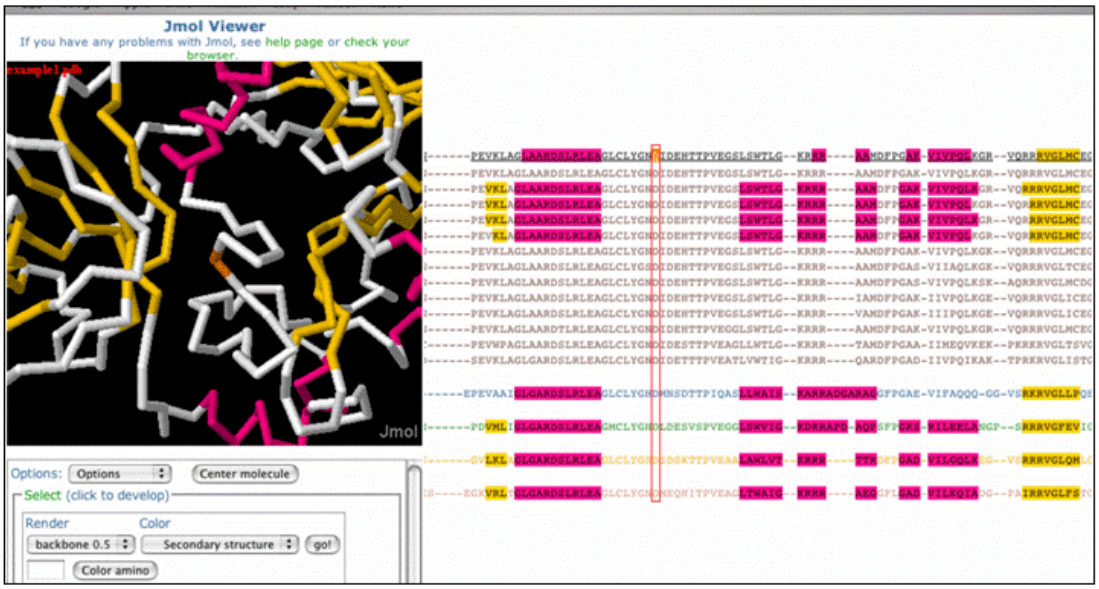


Figure 11.2 MAGOS web server display
Secondary structure elements are coloured red (helix) and yellow (beta strand) in both the MACS and the 3D structure representation. The residue highlighted in orange corresponds to a mutation (D276H) involved in non-ketotic hyperglycinemia (NKH)..

For example, figure 11.2 shows a MAGOS display of sequences similar to the human T-protein of the glycine cleavage system (Uniprot:P48728). Defects in this protein are a cause of non-ketotic hyperglycinemia (NKH) (OMIM:605899), also known as glycine encephalopathy (GCE). The MAGOS web server thus illustrates the data integration potential of MACSIMS, by characterizing mutations in terms of their evolutionary conservation, their position in the 3D structure and their role in functional sites.

11.4 Conclusions

Providing access to information and simulations to large communities of biologists should accelerate the process of biological discovery itself. MACSIMS has been designed to facilitate studies concerning the sequence/structure/function/evolution relationships of RNA and proteins. MACSIMS has already been expoited in a number of different projects, but the potential applications of MACSIMS are much broader, including such fields as the automatic annotation of the ever-increasing number of hypothetical proteins being produced by the high-throughput genome sequencing projects, or the definition of characteristic motifs for specific protein folds.

Publication n^{o}. 3

"It is perhaps hard to make firm statements on such questions without having examined them many times"
Aristotle, *Categories,* 8b21

12 MACSIMS : systematic testing of research hypotheses

12.1 Introduction

Information management systems (IMS) such as MACSIMS are designed to efficiently retrieve and organise the vast amount of biological data that is now available, including genomic sequences, 3D structures, cellular, phenotype and other types of biologically relevant information. Such systems are helping biologists systematically gather and exploit all the data crucial for their research, by automating many aspects, from data acquisition to knowledge discovery. But the number and variety of new data resources are increasing at an exponential rate, thanks mainly to the new, high-throughput technologies. For example, the last update of the NAR Molecular Biology Database Collection included 858 databases, which is 139 more than the previous one (Galperin, 2006). It is clearly impossible to include all the information now available, and a critical factor for the success of future IMS will be their ability to select specific, targeted information that will reduce the time, effort and resources required to sift through the Web's massive data storehouses. For a specific research problem, removing irrelevant information also allows one to focus on the key areas. For example, in a microarray discriminant analysis, the purpose of filtering out low-density and non-differentially-expressed genes is to remove genes that are unlikely to contribute to the phenotype difference. For automatic knowledge extraction and inference systems, we need to determine the appropiate information that will allow us to accurately model the biological data. In a recent editorial (Li, 2006), Wentian Li proposed that the optimal solution would be to "keep the appropriate level of model complexity that matches that of the data and at least throw away the irrelevant information". However, the question of what information is relevant for a given research question is not always evident.

For example, one area that has been the subject of much study recently is the prediction of functional residues, such as those involved in catalytic sites, protein modifications, protein interactions or ligand binding. Such predictions have important implications in many areas, including protein engineering, metabolic modelling, genetic studies of human disease susceptibility, and the development of new drug discovery strategies. Many methods for the prediction of functional sites have been developed that use amino acid conservation as the primary indicator of potential sites, based on the assumption that functional sites are more conserved during evolution (e.g. Lichtarge *et al*., 1996; Valdar, 2002). Other prediction methods have exploited structural information in order to identify functional sites (e.g. Laskowski *et al.*, 1996; Ondrechen *et al.*, 2001). More recently, it has become clear that neither sequence nor structure alone is sufficient for accurate predictions and efforts are now being concentrated on the combined use of both sequence conservation and structural information (e.g. Armon *et al.*, 2001; Madabushi *et al.,* 2002; Chelliah *et al.,* 2004; Cheng *et al.,* 2005). In the search for an accurate prediction of functional residues, a large number of different sequence/structure descriptors have been proposed. For example, the accuracy of prediction of functional effects of nsSNPs was investigated, using a 32-descriptor set including physiochemical properties of amino acids, protein electrostatics, amino-acid residue flexibility, and binding interactions (Karchin *et al*., 2005). It was shown that two descriptors, one describing the solvent accessibility of "wild-type" and "mutant" amino-acid

residues and one residue conservation score, achieved similar overall accuracy and produced less false positives than the complete 32-descriptor set.

Clearly, identifying the most informative descriptors remains critical to the success of any computational prediction method. This chapter introduces the methodology we have developed to select the most suitable information for integration in MACSIMS. We make use of the integrative power of MACSIMS, together with the high quality, large scale tests in the BAliBASE benchmark, to estimate the relevance of different sequence and 3D structure descriptors for the accurate prediction of functional residues. We showed in chapter 9 that BAliBASE represents a useful benchmark for the objective evaluation and comparison of multiple sequence alignment algorithms. Now, thanks to MACSIMS and the MAO ontology, it is possible to automatically integrate new structural and functional information in the BAliBASE reference alignments. In the resulting annotated alignments, the potential sequence and structure based criteria can be easily assessed by comparing the predictions to the known functional residues. The increased size of the latest release of the BAliBASE benchmark means that the predictive power of the descriptors can be reliably evaluated. Furthermore, because BAliBASE provides high quality, manually refined multiple alignments, the effect of the noise associated with sequence misalignments is signficantly reduced in the tests we performed.

To illustrate the potential of this methodology, we have selected a number of different descriptors that have been used recently in functional residue prediction methods. We chose two scores that can be calculated based only on the multiple sequence alignment, namely residue conservation and residue hydrophobicity scores, and two scores that are determined from the 3D structure i.e. surface accessibility, and the number of inter-residue contacts.

Residue conservation: Evolutionary conservation of residues is probably the most widely-used descriptor for the identification of functionally important residues (e.g. Lichtarge *et al*., 1996; Reddy *et al*., 2001). However, there has been some debate in the literature as to whether certain functional residues are in fact more conserved than other residues at the surface of the protein (e.g. Grishin *et al.*, 1994; Caffrey *et al.*, 2004).

Residue type: A previous study of 178 enzymes with 615 catalytic residues (Bartlett *et al*., 2002) showed that catalytic residue types are limited, with just six residue types (H, C, E, D, R, K) accounting for 70% of all catalytic residues. Residue hydrophobicity has been proposed as a feature of protein-protein interaction sites (e.g. Young *et al*., 1994; Glaser *et al*., 2001), although hydrophobicity at the interfaces of certain, transient complexes is not as distinguishable from the remainder of the surface as hydrophobicity at the interfaces of obligate complexes (Jones and Thornton, 1996).

Residue accessibility: It is generally assumed that functional residues should be exposed on the surface of the protein. Solvent accessibility or accessible surface area has been proven to be a useful factor in the prediction of the functional effects of amino acid substitutions (Karchin *et al*., 2005).

Inter-residue contacts: It has been proposed that functional sites might be spatially organised, with physically connected networks linking distant functional sites in the structure through packing interactions (Socolich *et al*., 2005).

The goal of this study is to determine which of these descriptors are correlated with functional residues and can consequently be exploited in future versions of MACSIMS.

12.2 Material and Methods

Large scale tests

Version 3.0 of the BAliBASE benchmark contains 217 high quality multiple alignments. The alignments were constructed based on 3D structural superpositions, followed by manual verification and refinement to ensure the correct alignment of functional sites. All the reference alignments in BAliBASE contain at least one sequence of a protein whose 3D structure is known and available in the PDB database. Information concerning functional residues was integrated in the multiple alignments using MACSIMS. Functional sites were extracted automatically from two manually verified sources:

(i) Functional sites in the PDB 'SITE' entries are annotated by the authors. They include a variety of different functional residues, such as catalytic sites, binding sites, or even 'residues around catalytic site'.
(ii) The Catalytic Site Atlas (CSA) contains reliable information about enzyme catalytic sites. However, the CSA is manually annotated from publications and not all PDB entries are currently included.

Calculation of conserved columns

We use two scores for the estimation of the conservation of a column in the BAliBASE multiple alignments:

(i) The Mean Distance (MD) column score is used in the calculation of the norMD objective function (Thompson *et al.*, 2001). For each column, a 20 dimensional sequence space is defined and the amino acids present in the column are assigned a position in the space, depending on a set of residue similarity scores. By default, the scores are based on the Gonnet 250 matrix (Benner *et al.*, 1993). The MD score is then defined as the weighted pairwise sum of the distances between all amino acids in the column. The MD column scores are normalized in the range of 0 to 100, allowing the direct comparison of the conservation scores for different alignments containing different numbers of sequences.
(ii) The PC measure is based on the conservation of physico-chemical residue groups. The PC score is calculated using the same algorithm as MD, but the residue comparison scores are based amino acids groups of physico-chemical properties (see table 12.1).

Amino acids	Physico-chemical property
KRH	Hydrophilic, basic
DEQN	Hydrophilic, non-basic
ACILMVFYW	Hydrophobic
FYW	Aromatic
PGST	Small, neutral

Table 12.1 Amino acid groups based on physico-chemical properties
Amino acid groups were determined based on selected physico-chemical properties (French and Robson, 1983).

Calculation of residue hydrophilicity

Residue hydrophilicity scores are calculated using the method developed by Kyte and. Doolittle in 1982. A *hydropathy scale* was composed where the hydrophilic and hydrophobic properties of each of the 20 amino acid side-chains was taken into consideration. The scale was based on experimental observations derived from the literature.

Calculation of solvent accessibility

The residue accessibility score is based on the classical definition of residue accessibility (Richards, 1977). The accessible surface of a protein is defined as the surface spanned by the centre of a spherical solvent probe as it rolls over the molecule. Here we compute the area of the accessible surface using a numerical integration (Koehl and Delarue, 1994). Calculations were performed on a single protein chain, in the absence of ligands. The total accessible surface area of the protein is broken down into accessible surface area for each of its residues.

Calculation of interresidue contacts

Inter-residue contacts are defined acording to Miyazawa and Jernigan (1993). For each residue, the centroid of its sidechain is computed. Two residues are considered to be in contact if their centroids are distant by less than 6 angstroms. Two descriptors based on interresidue contacts are defined for each residue:

(i) the number of interresidue contacts
(ii) the number of interresidue contacts with conserved residues

Integration in MACSIMS

The six sequence/structure descriptors described above were calculated by external programs and integrated in the context of the BAliBASE alignments using MACSIMS, based on the MAO ontology (see figure 12.1).

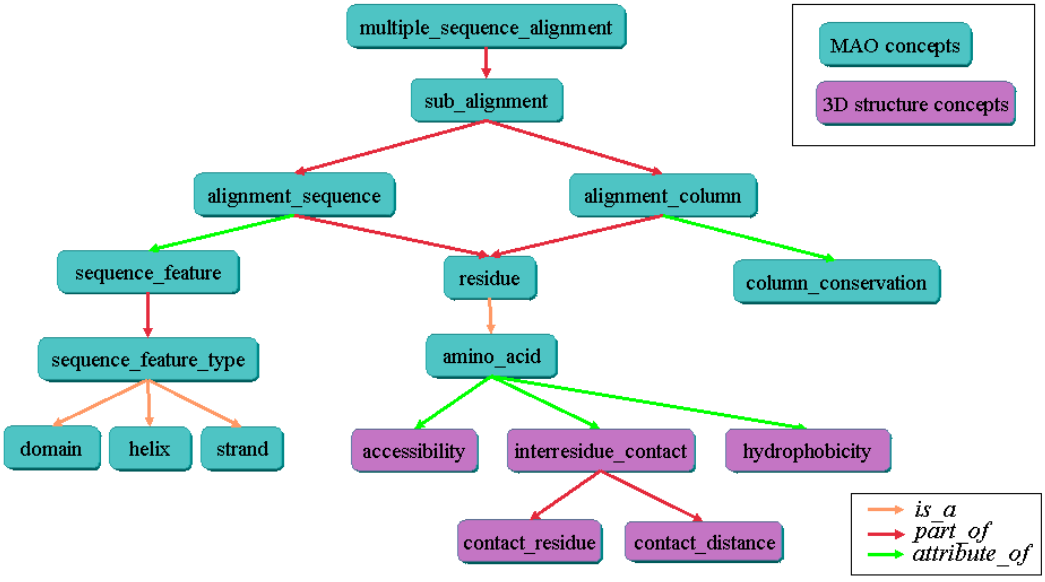


Figure 12.1Integration of 3D structural information in MAO

Descriptor evaluation

For each of the six sequence/structure descriptors, the same protocol was used to evaluate the pertinence of the information for the prediction of functional residues.

1. For each column in the alignment, we calculate the mean value of the descriptor for the PDB sequences present in the alignment. Using the mean values over sequences has been used previously to improve the prediction of 3D structure of a protein from its sequence (e.g. Finkelstein, 1998; Cui and Wong, 2000).

2. A histogram of the mean descriptor values allows a manual selection of the most appropriate threshold value for prediction of functional residues.

3. The accuracy of the functional residue predictions based on the descriptor is then evaluated by calculating the sensitivity and specificity:

Sensitivity=(100*TP)/(TP+FN)
Specificity=(100*TN)/(TN+FP)

where TP=number of true positives, TN=number of true negatives, FP=number of false positives, FN=number of false negatives.

12.3 Results and discussion

We have designed a protocol to study the pertinence and expoitability of different types of information in the context of MACSIMS. In particular, we have focused on the prediction of residues that are important for the function of the protein, based on six sequence and structure based descriptors. The predicted residues are compared to known functional sites in a large scale test, using the 217 multiple sequence alignments in the BAliBASE benchmark (version 3). The known functional sites were extracte from the CSA and PDB databases and integrated in the BAliBASE alignments using MACSIMS. Table 12.2 summarises the functional annotation of the BAliBASE alignments. The 217 full-length alignments contained a total of 207069 columns, of which 2623 columns were annotated with at least one functional site.

	Ref1		Ref2	Ref3	Ref4	Ref5	Total
	V1	V2					
Total columns	19018	28471	32843	25358	81777	19602	207069
Total columns with >=1 site	564	288	532	390	606	243	2623
Total core block columns	3581	9071	4314	3051	6123	2250	28390
Total core block columns with >=1 site	168	164	199	143	226	101	1001

Table 12.2 Known functional sites in BAliBASE alignments

Some of the functional sites are specific to sub-families in BAliBASE and are not conserved in the complete alignment at the structural similarity level. Therefore, in the subsequent tests, we consider only the functional sites that are located in conserved core

block regions. Thus, the test set used here consists of a total of 28390 alignment columns, of which 1001 have been identified as being important functional positions.

This test set was used to evaluate the six descriptors identified as being potentially useful for prediction of functional residues, namely MD residue conservation, PC residue conservation, hydrophilicity, surface accessibility, number of interresidue contacts and number of interresidue contacts with conserved residues. These descriptors are clearly not all independent. Table 12.3 shows the degree of correlation between the different descriptors.

	MD conservation	PC conservation	hydrophilicity	accessibility	contacts	conserved contacts
MD	1.0	0.84	-0.05	-0.19	0.17	0.28
PC		1.0	0.22	-0.35	0.35	0.47
hydrophilicity			1.0	-0.47	0.49	0.47
accessibility				1.0	-0.87	-0.75
contacts					1.0	0.87
conserved contacts						1.0

Table 12.3 Correlation coefficients between potential descriptors for prediction of functional residues
Correlation scores around 0 indicate non-correlated values.

The high correlation between the sequence-based descriptors, MD and PC, is to be expected as these two descriptors both measure the degree of residue conservation observed at each position in the alignment. However, the residue hydrophilicity score is clearly unrelated to the two conservation scores and may provide additional, complementary information for the functional residue information. Some correlation is also to be expected between the structure-based descriptors, surface accessibility and the number of residue contacts, since buried residues should generally make more interresidue contacts.

The sequence-based hydrophilicity scores and the structure-based surface accessibility values are inversely correlated. This result is in agreement with the previous observation of a significant correlation between hydrophobicity and surface exposure (Moelbert *et al*., 2004). Nevertheless the correlation in this study was not optimal. A number of factors were proposed by the authors to explain this. First, the poor correlation seen at the single sequence level may have been due to naturally occurring proteins with significant mutational stability or designability. Second, there are amino acids for which hydrophobicity is not the prime factor in determining exposure. For example, amino acids such as glycine can appear either on the surface or in the core, and charged amino acids can form salt bridges.

12.3.1 Residue conservation

Functional residues are often assumed to be more conserved during evolution. To test this hypothesis, we compared the two different residue conservation scores, mean distance (MD) and physico-chemical groups (PC), obtained for the functional and non-functional residues in BAliBASE (figure 12.2).

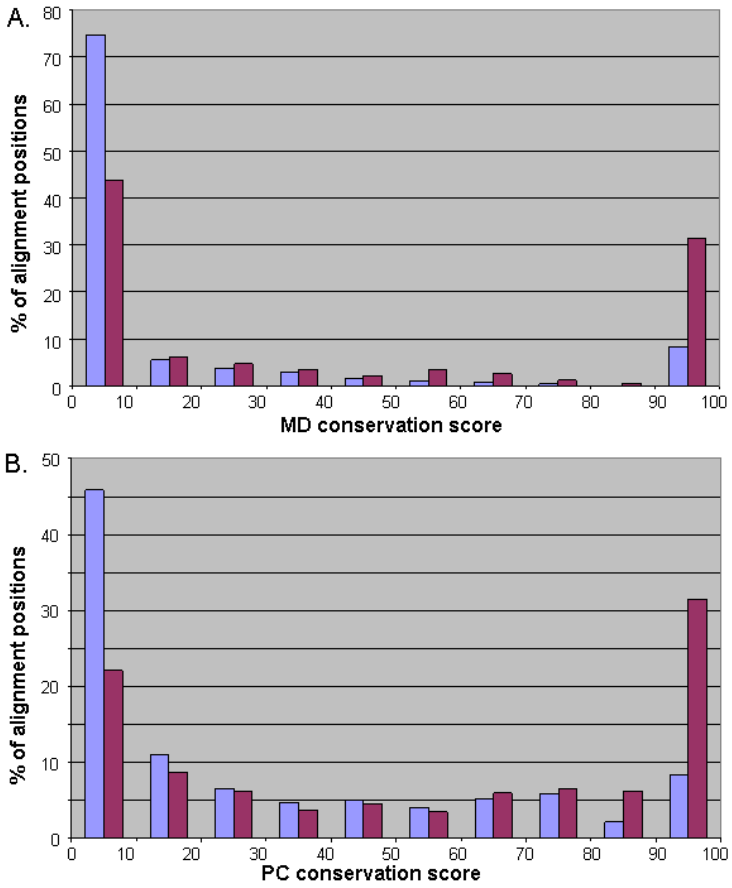


Figure 12.2 Frequency distribution of conservation scores for functional versus non-functional residues
A. MD conservations scores, B. PC conservation scores for functional (red) versus non-functional (blue) residues. Vertical axis represents proportions of residues for each conservation score.

We then selected various threshold scores for both conservation scores based on the histograms in figure 12.2. Alignment positions scoring higher than the threshold value were predicted to correspond to functional residues. Table 12.4 shows the results of the predictions based on the two conservation scores at different thresholds..

Prediction descriptors	Ref1		Ref2	Ref3	Ref4	Ref5	Total	
	V1	V2						
MD>20								
TP	69	112	121	73	131	54	560	SE=56%
FP	487	2505	1083	479	1402	558	6514	SP=76%
MD>10								
TP	76	122	141	80	150	65	634	SE=63%
FP	687	3140	1511	715	1892	756	8701	SP=68%
PC>20								
TP	98	140	170	106	178	77	769	SE=69%
FP	1356	4594	2461	1431	3103	1202	14147	SP=48%
PC>10								
TP	132	150	190	136	214	91	913	SE=91%
FP	2164	6461	3346	2212	4454	1670	20307	SP=26%

Table 12.4 Prediction of functional residues based on column conservation only
TP=true positive, FP=false positive. SE=sensitivity, SP=specificity. Here, a true positive indicates a position defined as conserved that contains at least one known functional site. False positives are conserved positions for which no function is currently known

A stricter definition of conservation leads to better specificity (less false positives), but lower sensitivity (more false negatives). Even a very loose definition of conservation (PC>10) results in some false negatives, that correspond to functional residues that are specific to sub-families in the BAliBASE alignment. Figure 12.3 shows an example of a false negative prediction in the BAliBASE alignment, BB11004. Column 153 in this alignment contains a conserved arginine and corresponds to a true positive prediction, regardless of the definition of conservation used. However, columns 170,172,174 and 176 contain functional residues, but are not conserved according to either the MD or the PC scores.

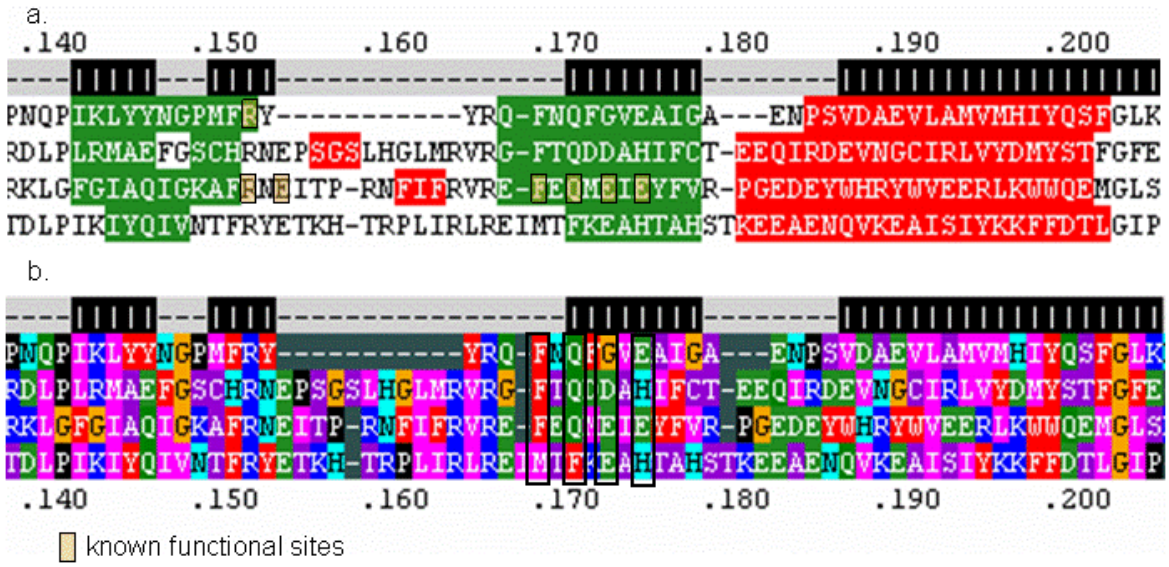


Figure 12.3 Part of BAliBASE alignment BB11004. Black boxes above the alignment indicate core blocks
The alignment contains 4 PDB sequences: 1qe0A=Histidyl-tRNA synthetase, 1evkA=threonyl-tRNA synthetase, 1atiA=glycyl-tRNA synthetase, 1nj8=prolyl-tRNA synthetase. a) Alignment coloured by secondary structure (red=helix, green=strand). Black boxes above the alignment indicate reliable core block regions. Yellow boxes indicate residues annotated as functional sites in PDB or CSA databases. b) Alignment coloured by residue type. Black boxes indicate columns containing at least one known functional site.

We conclude from these tests that residue conservation is a pertinent descriptor for the prediction of functional sites, but is not sufficient. We therefore need to include other information.

12.3.2 Residue type

In a previous study of 178 enzymes with 615 catalytic residues (Bartlett *et al.*, 2002), it was shown that catalytic residue types are limited, with just six residue types (H, C, E, D, R, K) accounting for 70% of all catalytic residues. We obtained similar results for the CSA catalytic sites in this data set (Figure 12.4A).

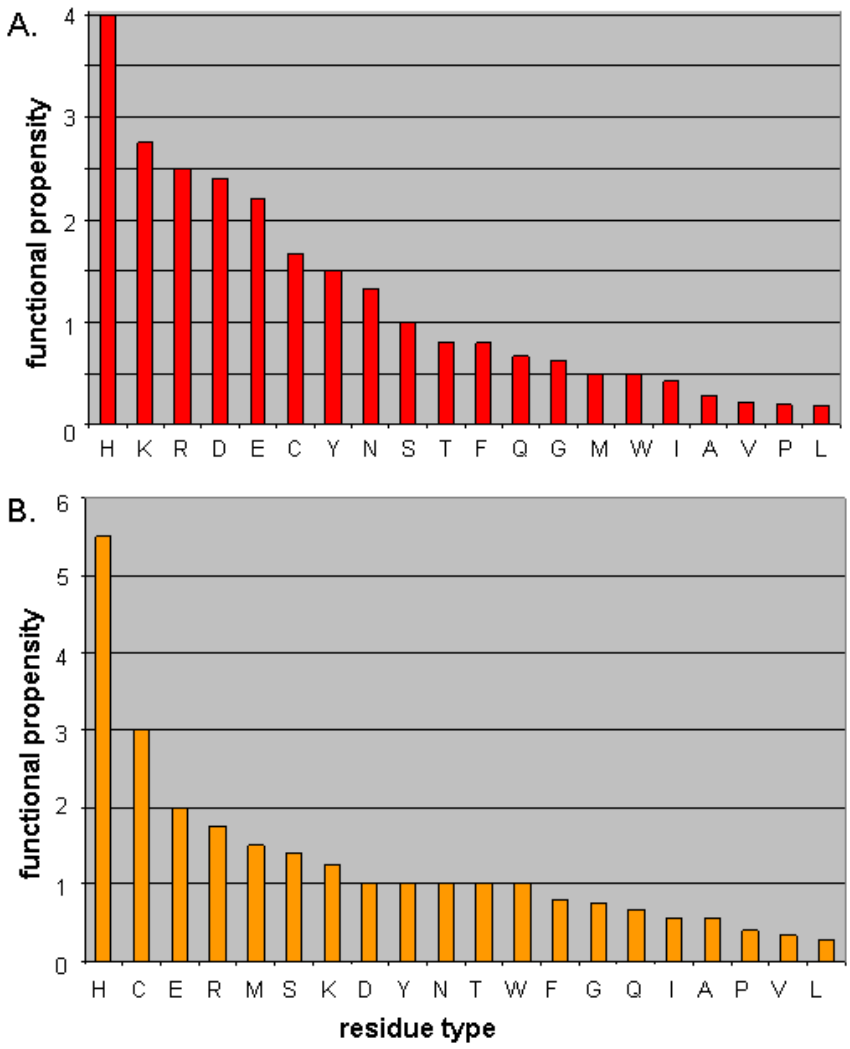


Figure 12.4 Functional propensities of the 20 amino acid types
A. Percentage of CSA catalytic residues, divided by the percentage of non-functional residues for each residue type. B. Percentage of PDB site residues, divided by the percentage of non-functional residues for each residue type.

The ranking of the observed functional propensities was different for the sites extracted from the PDB entries, although some similarities were observed. The residues H,C,E,R had high functional propensities for both test sets, while the residues I,A,V,L and P had very low propensities. Note that the PDB sites include not only catalytic residues, but also a wider variety of functional residues, such as those involved in protein-protein interactions, ligand binding etc.

Residue hydrophobicity has also been proposed as a feature of certain protein interaction sites (e.g. Young et al., 1994; Glaser et al., 2001). Figure 12.5 shows the Kyte-Doolittle hydrophilicity scores for functional versus non-functional residues.

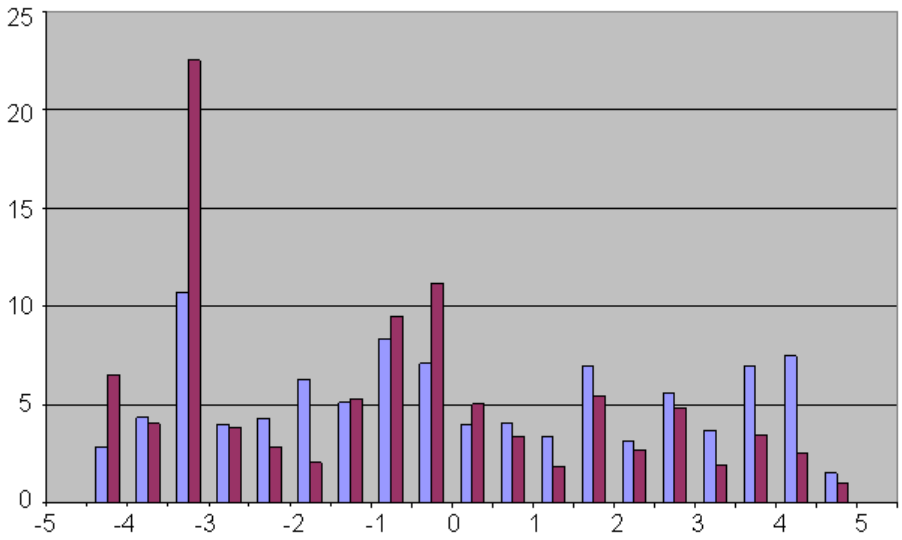


Figure 12.5 Frequency distribution of hydrophilicity scores for functional versus non-functional residues
Hydrophibilicity score for functional (red) versus non-functional (blue) residues. Vertical axis represents proportion of residues.

The peak at -3.5 to -3.0 is due to the high frequency of histidine residues in the functional residue dataset (11% compared to 2% for non-functional residues). In this histogram, there is no obvious threshold hydrophilicity score that differentiates functional and non-functional residues and the hydrophilicity score was therefore excluded from the prediction tests.

12.3.3 Solvent accessiblity

In a study of potential descriptors for prediction of functional effects of amino acid substitutions (Karchin *et al*., 2005), it was shown that solvent accessibility was one of the most pertinent parameters. One of the most widely used solvent accessibility scores follows the definition of residue accessibility defined by Richards in 1997. The score estimates the percentage of the residue surface that is accessible to a solvent, i.e. an accesssiblity score of zero means that the residue is buried.

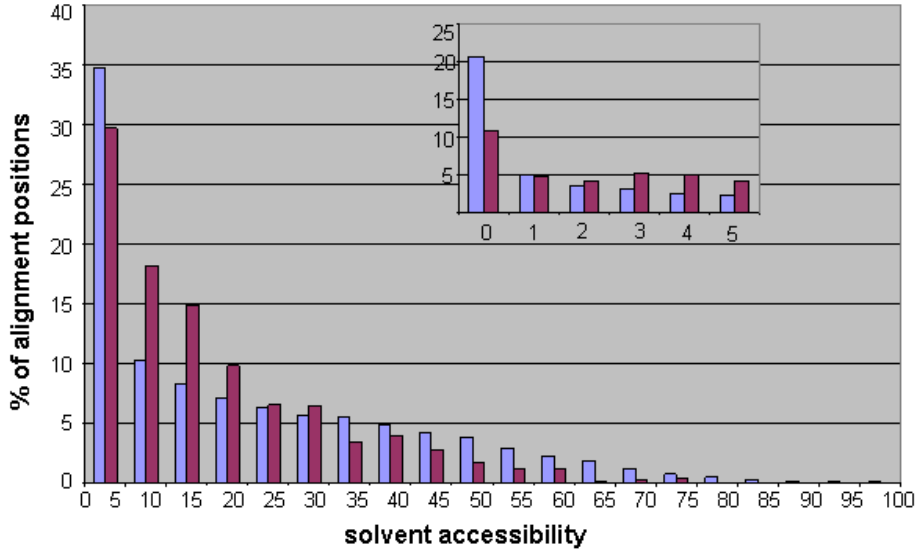


Figure 12.6 Frequency distribution of accessibility scores for functional versus non-functional residues
Accessibility scores for functional (red) versus non-functional (blue) residues. Vertical axis represents proportions of residues for each accessiblity score.

As shown in figure 12.6, 29% of the known functional residues have an accessibility score of less than 5% and 12% have accessibility=0%. This could be considered to be a surprising result, as one might expect the majority of functional residues to be on the surface of the protein. However, a similar result has been observed previously for catalytic residues (Bartlett *et al.*, 2002). The authors showed that 5% of all catalytic residues in the study had 0% relative solvent accessibility and were totally buried. There are a number of possible reasons for this result. First, the definition of solvent accessibility may not distinguish residues in restricted regions of the surface, such as pockets or clefts. Second, the crystallographic structures in the PDB database represent one conformation of the protein, while functional sites may only be exposed under certain conditions, such as in the presence of cofactors leading to changes in the structure of the protein (allostery).

For the prediction of functional sites based on the multiple alignment, we calculated the mean accessibility score for each column. We chose a relatively low threshold for the prediction of functional sites, since a higher threshold would lead a large number of false negatives and thus, to less sensitivity. Table 12.5 shows the results of the predictions for different mean accessibility thresholds.

Prediction descriptors	Ref1		Ref2	Ref3	Ref4	Ref5	Total	
	V1	V2						
Conserved+ accessibility>0%								
TP	131	140	190	136	206	90	893	SE=89%
FP	1956	5450	3145	2081	3728	1494	17854	SP=35%
Conserved+ accessibility>1%								
TP	117	116	177	125	189	81	805	SE=80%
FP	1558	4435	2563	1687	2992	1193	14428	SP=47%
Conserved+ accessibility>2%								
TP	112	111	171	114	178	75	761	SE=76%
FP	1428	4070	2370	1556	2739	1081	13244	SP=52%
Conserved+ accessibility>5%								
TP	96	87	143	95	141	64	626	SE=63%
FP	1153	3453	2031	1316	2291	893	11137	SP=59%

Table 12.5 Prediction of functional residues based on residue conservation and mean accessibility
Conserved columns were defined ashaving PC>10. TP=true positive, FP=false positive. SE=sensitivity, SP=specificity.

The inclusion of the accessibility factor reduces the number of false positives, compared to that obtained when using only residue conservation. The criteria of >0% accessibility leads to a high recall of conserved sites (sensitivity= 89%), but has a low specificity (=35%). Using a 5% threshold for accessibility, decreases the sensitivity to 63%, but increases the specificity to 59%.

12.3.4 Interresidue contacts

It has been proposed that functional sites might be spatially organised, with physically connected networks linking distant functional sites in the structure through packing interactions (Socolich *et al.*, 2005). Furthermore, it was hypothesised that the amino acid

interactions specifying the atomic structure should be conserved throughout the members of a protein family. To investigate the possibility that the extent of residue interactions may differ between functional and non-functional residues, we calculated the number of interresidue contacts for each residue in the BAliBASE alignments. Figure 12.7 shows the number of interresidue contacts for functional and non-functional residues, averaged over the PDB sequences present in the alignment.

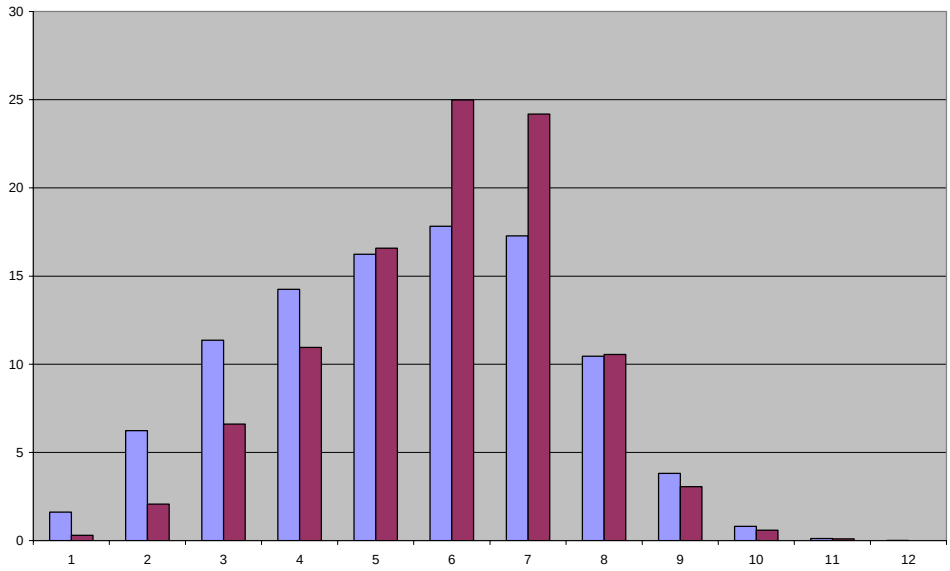


Figure 12.7 Frequency distribution of interresidue contacts for functional versus non-functional residues
Number of interresidue contacts for functional (red) versus non-functional (blue) residues. Vertical axis represents proportion of residues.

In general, functional residues have more interresidue contacts, but the difference is not sufficient for this score to be useful for distinguishing functional and non-functional residues. We also calculated the number of interresidue contacts with conserved residues. On average, functional residues have more contacts with conserved residues than non-functional ones (figure 12.8).

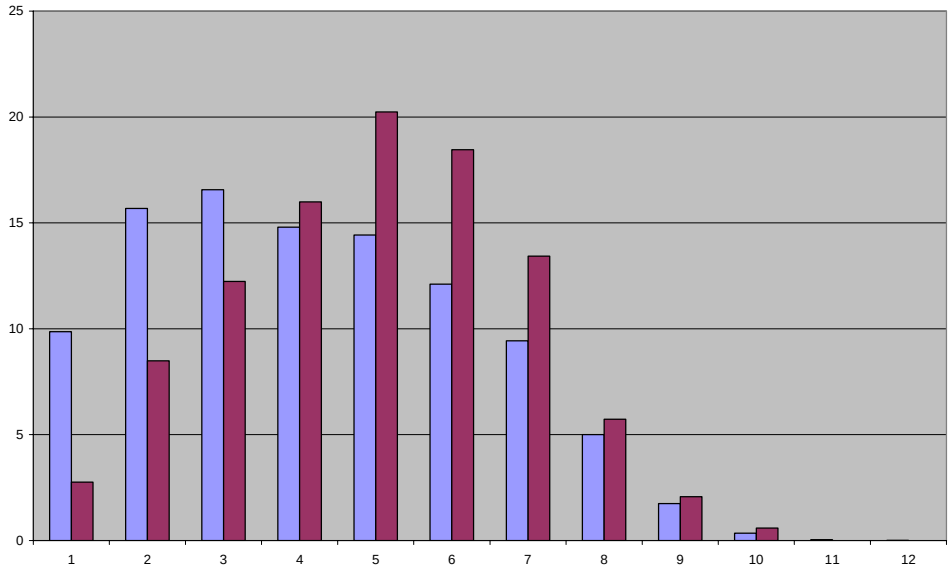


Figure 12.8 Frequency distribution of interresidue contacts with conserved residues for functional versus non-functional residues
Number of interresidue contacts with conserved residues for functional (red) versus non-functional (blue) residues. Vertical axis represents proportion of residues.

For the prediction of functional sites based on the multiple alignments, we calculated the mean number of conserved contacts for each column. Table 12.6 shows the results of the predictions for different thresholds of accessibility and number of contacts with conserved residues.

Prediction descriptors	Ref1		Ref2	Ref3	Ref4	Ref5	total	
	V1	V2						
Conserved+ accessibility>0%+contacts>1								
TP	126	139	185	131	202	86	869	SE=87%
FP	1677	4609	2926	1906	3265	1377	15760	SP=42%
Conserved+ accessibility>0%+contacts>2								
TP	102	131	175	118	189	83	798	SE=80%
FP	1398	3845	2567	1611	2772	1218	13411	SP=51%
Conserved+ accessibility>1%+contacts>1								
TP	112	115	172	120	185	77	781	SE=78%
FP	1281	3601	2344	1513	2531	1077	12347	SP=55%
Conserved+ accessibility>1%+contacts>2								
TP	88	107	162	107	172	74	710	SE=71%
FP	1009	2854	1985	1220	2042	921	10031	SP=63%

Table 12.6 Prediction of functional residues, based on conservation, mean accessibility and mean conserved contacts.
Conserved columns were defined ashaving PC>10. TP=true positive, FP=false positive. SE=sensitivity, SP=specificity.

The inclusion of the structure based descriptors (accessibility and number of interresidue contacts with conserved residues) leads to an improvement in the sensitivity of the predictions. For example, using the combined sequence and structure descriptors, we achieved a sensitivity of 71%, with a specificity of 63%. Using only the residue conservation descriptor, for a similar level of sensitivity (=69%), we obtained a specificity of only 48%.

12.4 Conclusions and Perspectives

We have shown that MACSIMS can be used to extend the range of applications for the BAliBASE benchmark. When structural/functional information is integrated in the BAliBASE alignments, the benchmark can be used as a test environment for the numerous bioinformatics applications that exploit evolutionary information. Here, we have developed a methodology for testing the pertinence of new information in the context of the MACS. We have demonstrated the efficiency of the method for a number of different sequence/structure predictors that might represent potential predictors of functional residues.

Another reason for the low specificity observed in these tests, may be that not all the functional sites in the alignments are annotated in the PDB and CSA databases. It is therefore likely that some of the false negative predictions may actually be functional sites. This problem will hopefully be alleviated by the recent efforts towards standardisation of bioinformatics data resources and a more systematic annotation of biological sequences and

structures, which should lead to more precise definitions of protein function and functional sites.

Using two different residue conservation scores, MD and PC, we detected most of the known functional sites, but we also obtained a large number of false positives. In the future, we will investigate alternative definitions of residue conservation. One problem is the threshold used to specify whether the alignment position is conserved or not. Alignments of closely related sequences will obviously contain more positions with high column conservation scores, which may not necessarily correspond to functionally important sites. Other measures have been developed recently that take into account the overall similarity of the sequences in the alignment (for a review, see Valdar, 2002). An informative conservation measure should also be able to take into account residues that are conserved in only a certain number of sub-families in the alignment.

Although, the inclusion of structure-based descriptors increased the specificity of the functional residue predictions, some improvement is still needed. It is possible that other structure based measures, such as atom depth (Pintar *et al.*, 2003) or designed sequence profiles (Koehl and Levitt, 2002b), may provide more pertinent information. In the future, these measures will also be integrated in MACSIMS and systematic tests will be performed to find an optimal combination of descriptors for functional predictions.

We have focused here on the prediction of residues that are important for the function of a protein. In the future, the 3D structure information in MACSIMS will also be exploited for other applications. For example, we have identified a large number of conserved residues that are not at the surface of the protein, but are buried in the core of the protein. It would be interesting to investigate in more detail the nature of these residues and their role in the protein. Are these residues important for the structural stability of the protein? Do they correspond to the "topohydrophobic" positions identified by Poupon and Mornon (Poupon and Mornon, 1998)? Or do they form communication pathways that link distant functional sites at the surface of the protein (Socolich *et al.*, 2005)?

The structural and functional genomics projects are now providing the raw data needed in order to address these issues. MACSIMS represents an ideal environment for the integration and analysis of this data and should hopefully contribute to future studies aimed at providing the answers to such fundamental biological questions.

13 Conclusions and Perspectives

The discovery of the DNA structure in 1953 opened a new field of biological research. Fifty years later, in 2003, the human genome sequence was completed. During this time, huge amounts of biological data have been collected in databases that are now publically available on the Internet. However the *data* is not to be confused with *information* (which is the data that we understand) and with the *knowledge* (which is a larger structure of the different information that make sense for humans). The large-scale accumulation of data is only the beginning of the path to the ultimate goal of understanding the basic principles underlying the complexity of living cells and organisms. Recently, the field of systems biology has emerged with the goal of understanding of existing data, via integration and knowledge extraction, combined with mathematical modelling in order to predict behaviour the system under different conditions (Kanehisa and Bork, 2003).

In this context, Multiple Alignments of Complete Sequences (MACS) represent an ideal tool for the study of the relationships between sequence and structure, function and evolution. The work presented here represents the first steps in the evolution of the traditional multiple sequence alignment from a simple stacking of letters to become an interactive tool, incorporating not only the sequence itself, but also structural/functional information in the context of the complete protein family.

Quality evaluation of multiple alignment programs

The first part of this thesis addressed the problem of the accuracy and reliability of multiple alignment algorithms. Multiple sequence alignment 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. The quality of the multiple alignment is critical for all these applications because errors introduced at the alignment stage will lead to further errors in the subsequent analyses.

We developed a new version of the BAliBASE benchmark database, which has become a reference for the evaluation and comparison of alignment programs. BAliBASE provides high quality, manually refined, reference alignments based on 3D structural superpositions. A semi-automatic protocol has been introduced to allow the creation of larger reference alignments that are more representative of the problems that are now encountered in the post-genomic era. The alignments are organised into different reference sets, containing test cases that cover the most common multiple alignment problems, from alignment of single domains e.g. in the construction of protein domain databases to the alignment of full-length, complex sequences, such as those detected by the database searches routinely performed in automatic, high throughput genome analysis projects.

In the search for more accurate alignments, most state-of-the-art methods now use a combination of complementary techniques, or integrate information other than the sequence itself. A comparison of the most recent alignment programs using BAliBASE version 3 has shown that significant improvements have been achieved, in particular by the use of information from both local and global alignment algorithms, in programs such as MAFFT, MUSCLE or ProbCons. Nevertheless, a number of problems remain and more progress will be needed for the reliable alignment of complex, multi-domain proteins.

Multiple alignment ontology

The second part of this thesis involved the development of the Multiple Alignment Ontology (MAO), a task-oriented ontology for nucleic acid and protein sequence and structure alignments. MAO has been designed for two main purposes. Firstly, the ontology facilitates the interoperation of different methods for multiple alignment and analysis. Secondly, MAO serves as a data model for information management, in order to facilitate data integration and knowledge extraction. Most of the basic features associated with multiple alignments are defined as MAO concepts, ranging from a single residue to sub-families of sequences. Attributes associated with the basic concepts allow the definition of more complex information, such as column conservation, residue or motif function, or 3D structural information. One of the most powerful features of the MAO ontology is that it provides a natural, intuitive link between a number of different ontologies in the domains of genomics and proteomics. Using the cross-references defined in MAO, diverse functional information from external data resources, such as active sites, mutation data and their associated phenotypes, etc. can be integrated, either for a single sequence or for a family of sequences.

The ontology has been developed in collaboration with domain experts from both the DNA/RNA and protein communities, who intend to offer compatible multiple alignment tools and analysis results that commit to the MAO ontology.

MACS-based information management system

MACSIMS is a MACS-based information management program that allows the integration of diverse structural and functional information in the context of the multiple alignment. The goal is not simply to provide convenient links between the different data resources, but to provide an interactive workbench for data validation and analysis, and presentation of the pertinent information to the biologist. In MACSIMS, the data retrieved from the public databases is cross-validated and the reliable information is propagated from the known to the unknown sequences. New algorithms have been developed that identify the well-aligned regions of the multiple alignment, in order to ensure that information is only transferred between sequences that are homologous. In addition to these knowledge-based annotations, *ab initio* sequence analysis methods have been incorporated, such as the prediction of transmembrane regions, coiled coil or low complexity sequence segments. These methods provide valuable information in the case of ‘orphan’ proteins, for which no known homologues detected in the sequence databases.

The informational content of MACSIMS has been exploited in a number of projects in the LBGS, such as the validation of predicted protein sequences, the characterisation of targets for the Structural Proteomics IN Europe (SPINE) project or the definition of genotype/phenotype correlations for the Structural Mutation to Human Pathologies Phenotype (MS2PH) project. The integrative power of MACSIMS has also been used as a research tool to investigate the importance of different types of information for the prediction of protein functional sites. By comparing the conserved residues in multiple sequence alignments with 3D structural information, such as solvent accessibility and inter-residue contacts, we were able to improve the accuracy of functional residue predictions. However, the efficient exploitation of structural information remains a challenging problem that needs to be addressed.

Future perspectives

The comparison of the most recent multiple alignment programs using BAliBASE version 3 has shown that, despite significant progress, none of the available methods is capable of producing reliable alignments for the complex, divergent proteins that are detected by today's advanced database search algorithms. Therefore, we plan to develop a new multiple alignment method that will exploit all the structural/functional information integrated in MACSIMS to construct a high quality multiple sequence alignment, even in the difficult case of complex, multi-domain proteins. An important aspect of the new method will be the definition of a novel knowledge-based objective function to estimate the biological significance of the alignment.

In the future, the MAO ontology will be extended to incorporate other data resources, such as gene structure, mutation and phenotype information and residue interaction data. This will require more formal links between MAO and the other biological ontologies. The integration of this information in MACSIMS will increase its potential applications, to include such fields as the automatic annotation of the ever-increasing number of hypothetical proteins being produced by the high-throughput genome sequencing projects or the definition of characteristic motifs for protein folds. To achieve this, we will combine the knowledge processing power of MACSIMS with the versatility of empirical learning systems, such as artificial neural networks (ANNs). Such Hybrid Learning (HL) systems that exploit simultaneously theoretical and empirical data should be more efficient than either of the approaches working separately (Towell and Shavlik, 1993). Another critical factor in the potential utility of MACSIMS will be the development of a new, more user-friendly interface for the presentation of the discovered knowledge to the biologist. Hopefully, these developments will also have significant consequences for more wide-reaching areas, such as protein engineering, metabolic modelling, genetic studies of human disease susceptibility, and the development of new drug discovery strategies.

Another growing area of research is the application of multiple alignment methods for the comparison of other kinds of data, such as 3D structure fragment libraries or structural alphabets (e.g. Kolodny *et al*., 2002; Camproux *et al.*, 2004; de Brevern, 2005), molecular networks (Sharan and Ideker, 2006), or even time use data and activity patterns in the social sciences (Thompson *et al*., 1999c; Wilson, 2006). Here, data that is fundamentally a sequence of events is represented by an alphabet defined by experts in the field, who also define the similarity scores between the different events. These emerging fields are exploiting the power of the multiple alignment methodologies developed over the years for the comparison of molecular sequences, but will also contribute new concepts and formulations that will undoubtedly prove beneficial in the future.

References

1. Abou-Sleymane G, Chalmel F, Helmlinger D, Lardenois A, Thibault C, Weber C, Merienne K, Mandel JL, Poch O, Devys D, Trottier Y. Polyglutamine expansion causes neurodegeneration by altering the neuronal differentiation program. *Hum Mol Genet*. **2006** 15:691-703.
2. Aderem A. Systems biology: its practice and challenges. *Cell*. **2005** 121:511-3.
3. Aebersold, R Molecular Systems Biology: a new journal for a new biology? *Mol Syst Biol* **2005** 1:5.
4. Ahlquist P. Parallels among positive-strand RNA viruses, reverse-transcribing viruses and double-stranded RNA viruses. *Nat Rev Microbiol*. **2006** 4:371-82.
5. Aggarwal G, Ramaswamy R. Ab initio gene identification: prokaryote genome annotation with GeneScan and GLIMMER. *J Biosci.* **2002** 27:7-14.
6. Albeck S, Alzari P, Andreini C, Banci L, Berry IM, Bertini I, Cambillau C, Canard B, Carter L, Cohen SX, Diprose JM, Dym O, Esnouf RM, Felder C, Ferron F, Guillemot F, Hamer R, Ben Jelloul M, Laskowski RA, Laurent T, Longhi S, Lopez R, Luchinat C, Malet H, Mochel T, Morris RJ, Moulinier L, Oinn T, Pajon A, Peleg Y, Perrakis A, Poch O, Prilusky J, Rachedi A, Ripp R, Rosato A, Silman I, Stuart DI, Sussman JL, Thierry J-C, Thompson JD, Thornton JM, Unger T, Vaughan B, Vranken W, Watson JD, Whamond G, Henrick K. SPINE bioinformatics and data-management aspects of high-throughput structural biology *Acta Cryst.* **2006** D62, 1184-1195.
7. Al-Lazikani, B., Jung, J., Xiang, Z. and Honig, B. Protein structure prediction. *Curr Opin Chem Biol.* **2001** 5:51-56.
8. Altman R, Bada M, Chai XJ, Whirl Carillo M, Chen RO, and Abernethy NF. RiboWeb: An Ontology-Based System for Collaborative Molecular Biology. *IEEE Intelligent Systems*, **1999** 14:68-76.
9. Altschul SF, Gish W. Local alignment statistics. *Methods Enzymol.* **1996** 266:460-80.
10. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* **1990** 215:403–410.
11. 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** 25:3389-3402.
12. Andrade MA, Bork P. Automated extraction of information in molecular biology. *FEBS Lett*. **2000** 476:12-7.
13. Apic G, Ignjatovic T, Boyer S, Russell RB. Illuminating drug discovery with biological pathways. *FEBS Lett*. **2005** 579: 1872-7.
14. Armon A, Graur D, Ben-Tal N. ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J Mol Biol.* **2001** 307:447-63.
15. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*. **2000** 25:25-9.
16. Atteson, K. The performance of neighbor-joining algorithms of phylogeny reconstruction, **1997** pp. 101-110. In Jiang, T., and Lee, D., eds., Lecture Notes in Computer Science, 1276, Springer-Verlag, Berlin.
17. Attwood TK. The PRINTS database: a resource for identification of protein families. *Brief Bioinform*. **2002** 3:252-63.
18. Bader GD, Betel D, Hogue CW. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res*. **2003** 31:248-50.
19. Bahr A, Thompson JD, Thierry JC, Poch O. BAliBASE (Benchmark Alignment dataBASE): enhancements for repeats, transmembrane sequences and circular permutations. *Nucleic Acids Res*. **2001** 29:323-6.
20. Bairoch A, Boeckmann B. The SWISS-PROT protein sequence data bank. *Nucleic Acids Res*. **1991** 19:2247-9.
21. Bairoch A. The ENZYME database in 2000. *Nucleic Acids Res*. **2000** 28:304-5.
22. Baker PG, Brass A, Bechhofer S, Goble C, Paton N, and Stevens R.. TAMBIS: Transparent Access to Multiple Bioinformatics Information Sources. An Overview. In *Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology (ISMB'98)*, **1998** pages 25-34.
23. Baker PG, Goble CA, Bechhofer S, Paton NW, Stevens R, and Brass A. An Ontology for Bioinformatics Applications. *Bioinformatics.* **1999** 15:510-520.
24. Bao L, Cui Y. Prediction of the phenotypic effects of nonsynonymous single nucleotide polymorphisms using structural and evolutionary information. *Bioinformatics*. **2005** 21:2185-2190.

25. Bard JB, Rhee SY. Ontologies in biology: design, applications and future challenges. *Nat Rev Genet.*, **2004** 5:213-222.
26. Barkai N, Leibler S. Robustness in simple biochemical networks. *Nature*. **1997** 387:913-7.
27. Bartlett GJ, Porter CT, Borkakoti N, Thornton JM. Analysis of catalytic residues in enzyme active sites. *J Mol Biol.* **2002** 324:105-21.
28. Barton GJ, Sternberg JE. A strategy for the rapid multiple alignment of protein sequences. Confidence levels from tertiary structure comparisons *J Mol Biol* **1987** 198:327-337.
29. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, Eddy SR. The Pfam protein families database. *Nucleic Acids Res*. **2004** 32:D138-41.
30. Bejerano, G., Seldin, Y., Margali,t H. And Tishby, N. Markovian domain fingerprinting: statistical segmentation of protein sequences. *Bioinformatics* **2001** 17, 927-934.
31. Benner SA, Cohen MA, Gonnet GH. Empirical and structural models for insertions and deletions in the divergent evolution of proteins. *J Mol Biol*. **1993** 229:1065-82.
32. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. *Nucleic Acids Res.* **2006** 34:D16-20.
33. Berners–Lee T, Hendler J, Lassila O. "The Semantic Web," *Scientific American*, May **2001**.
34. Bernstein FC, Koetzle TF, Williams GJ, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. The Protein Data Bank: a computer-based archival file for macromolecular structures. *J Mol Biol*. **1977** 112:535-42.
35. Bertone P, Kluger Y, Lan N, Zheng D, Christendat D, Yee A, Edwards AM, Arrowsmith CH, Montelione GT, Gerstein M. SPINE: an integrated tracking database and data mining approach for identifying feasible targets in high-throughput structural proteomics. *Nucleic Acids Res.* **2001** 29:2884-2898.
36. Bianchetti L, Thompson JD, Lecompte O, Plewniak F, Poch O. vALId: validation of protein sequence quality based on multiple alignment data. *J Bioinform Comput Biol*. **2005** 3:929-47.
37. Bienkowska J. Computational characterization of proteins. *Expert Rev Proteomics*. **2005** 2:129-38.
38. Blackshields G, Wallace IM, Larkin M, Higgins DG. Analysis and comparison of benchmarks for multiple sequence alignment. *In Silico Biology* **2006** 6, 0030.
39. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*. **2003** 31:365-70.
40. Bonneau R, Strauss CE, Baker D. Improving the performance of Rosetta using multiple sequence alignment information and global measures of hydrophobic core formation. *Proteins*. **2001** 43:1-11.
41. Borgida A. Description Logics in Data Management. *IEEE Trans Knowledge and Data Engineering*, **1995** 7:671-782.
42. Bork P, Serrano L. Towards cellular systems in 4D. *Cell*. **2005** 121:507-9.
43. Breast Cancer Linkage Consortium. Pathology of familial breast cancer: differences between breast cancers in carriers of BRCA1 and BRCA2 mutations and sporadic cases. *Lancet* **1997** 349:1505-1510.
44. Brelivet Y, Kammerer S, Rochel N, Poch O, Moras D. Signature of the oligomeric behaviour of nuclear receptors at the sequence and structural level. *EMBO Rep*. **2004** 5:423-9.
45. Brenner SE, Chothia C, Hubbard TJ. Population statistics of protein structures: lessons from structural classifications. *Curr Opin Struct Biol*. **1997** 7:369-76.
46. Brion P, Westhof E. Hierarchy and dynamics of RNA folding. *Annu Rev Biophys Biomol Struct.* **1997** 26:113-37.
47. Bru C, Courcelle E, Carrere S, Beausse Y, Dalmar S, Kahn D. The ProDom database of protein domain families: more emphasis on 3D. *Nucleic Acids Res*. **2005** 33:D212-5.
48. Bussard AE. A scientific revolution? The prion anomaly may challenge the central dogma of molecular biology. *EMBO reports* **2005** 6, 8, 691–694
49. Butler, BA. Sequence analysis using GCG. *Methods Biochem Anal*. **1998** 39:74-97.
50. Caffrey, D.R., Somaroo, S., Hughes, J.D., Mintseris, J., Huang, E.S. Are protein–protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci* **2004** 13:190–202
51. Camproux AC, Gautier R, Tuffery P. A hidden markov model derived structural alphabet for proteins. *J Mol Biol.* **2004** 339:591-605.
52. Carrillo H, Lipman D. The Multiple Sequence Alignment Problem in Biology. *SIAM J Appl Math* **1988** 48:1073-1082
53. Chakrabarti R, Klibanov AM, Friesner RA. Computational prediction of native protein ligand-binding and enzyme active site sequences. *Proc Natl Acad Sci U S A*. **2005** 102:10153-8.
54. Chalmel F, Lardenois A, Thompson JD, Muller J, Sahel JA, Leveillard T, Poch O. GOAnno: GO annotation based on multiple alignment. *Bioinformatics*. **2005** 21:2095-6.
55. Chelliah V, Chen L, Blundell TL, Lovell SC. Distinguishing structural and functional restraints in

evolution in order to identify interaction sites. J Mol Biol. 2004 342:1487-504.
56. Chen CP, Kernytsky A, Rost B. Transmembrane helix predictions revisited. *Protein Sci*. **2002** 11:2774-91.
57. Cheng G, Qian B, Samudrala R, Baker D. Improvement in protein functional site prediction by distinguishing structural and functional constraints on protein family evolution using computational design. *Nucleic Acids Res.* **2005** 33:5861-7.
58. Cho S, Park SG, Lee do H, Park BC. Protein-protein interaction networks: from interactions to networks. *J Biochem Mol Biol*. **2004** 37: 45-52.
59. Chong L and Ray LB. Whole-istic biology. *Science*, **2002** 295:1661.
60. Chothia,C. Principles that determine the structures of proteins. *Ann. Rev. Biochem*. **1984** 53, 537–572.
61. Chothia C and Lesk A. The relation between the divergence of sequence and structure in proteins. *EMBO J.* **1986** 5**:**823–826.
62. Chung S. and Subbiah S. A structural explanation for the twilight zone of protein sequence homology. *Structure* **1996** 4**:**1123–1127.
63. Clamp M, Cuff J, Searle SM, Barton GJ. The Jalview Java alignment editor. *Bioinformatics*. **2004** 20:426-7.
64. Cline,M., Hughey,R. and Karplus,K. Predicting reliable regions in protein sequence alignments. *Bioinformatics*, **2002** 18:306–314.
65. Cochrane G, Aldebert P, Althorpe N, Andersson M, Baker W, Baldwin A, Bates K, Bhattacharyya S, Browne P, van den Broek A, Castro M, Duggan K, Eberhardt R, Faruque N, Gamble J, Kanz C, Kulikova T, Lee C, Leinonen R, Lin Q, Lombard V, Lopez R, McHale M, McWilliam H, Mukherjee G, Nardone F, Pastor MP, Sobhany S, Stoehr P, Tzouvara K, Vaughan R, Wu D, Zhu W, Apweiler R. EMBL Nucleotide Sequence Database: developments in 2005. *Nucleic Acids Res*. **2006** 34:D10-5.
66. Combet C, Jambon M, Deleage G, Geourjon C. Geno3D: automatic comparative molecular modelling of protein. Bioinformatics. 2002 18:213-4.
67. Cooke GE. Pharmacogenetics of multigenic disease: heart disease as an example. *Vascul Pharmacol*. **2006** 44:66-74.
68. Cornell M, Paton NW, Hedeler C, Kirby P, Delneri D, Hayes A, Oliver SG. GIMS: an integrated data storage and analysis environment for genomic and functional data. *Yeast*. **2003** 20:1291-306.
69. Costa FF. Non-coding RNAs: new players in eukaryotic biology. *Gene*. **2005** 357:83-94.
70. Côté RG, Jones P, Apweiler R, Hermjakob H. The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries. *BMC Bioinformatics*, **2006** 7, 97.
71. Creighton C, Hanash S. Mining gene expression databases for association rules. *Bioinformatics* **2003** 19:79-86.
72. Crick, FHC. On protein synthesis. *Symp. Soc. Exp. Biol*. **1958** 12:138-183.
73. Crick, FHC. The origin of the genetic code, *J Mol Biol* **1968** 38:367–379.
74. Crooks GE, Hon G, Chandonia JM, Brenner SE WebLogo: A sequence logo generator. *Genome Research*, **2004** 14:1188-1190.
75. Cui Y, Wong WH. Multiple-sequence information provides protection against mis-specified potential energy functions in the lattice model of proteins. *Phys. Rev. Lett.* **2000** 85:5242–5.
76. Curwen V, Eyras E, Andrews TD, Clarke L, Mongin E, Searle SM, Clamp M. The Ensembl automatic gene annotation system. *Genome Res.* **2004** 14:942-50.
77. Dandekar T, Huynen M, Regula JT, Ueberle B, Zimmermann CU, Andrade MA, Doerks T, Sanchez-Pulido L, Snel B, Suyama M, Yuan YP, Herrmann R, Bork P. Re-annotating the mycoplasma pneumoniae genome sequence: adding value, function and reading frames. *Nucleic Acids Res*. **2000** 28:3278-3288.
78. Darling AE, Mau B, Blattner FR, Perna NT. GRIL: genome rearrangement and inversion locator. *Bioinformatics* **2004** 20: 1224.
79. Davidson SB, Overton C, Buneman P. Challenges in integrating biological data sources. *J Comput Biol*. **1995** 2:557-72.
80. Davidson EH, Rast JP, Oliveri P, Ransick A, Calestani C, Yuh C-H, Minokawa T, Amore G, Hinman V, Arenas-Mena C, Otim O, Brown CT, Livi CB, Lee PY, Revilla R., Rust AG, Pan Z, Schilstra MJ, Clarke PJ, Arnone MI, Rowen L, Cameron RA, McClay DR, Hood L, Bolouri H. A genomic regulatory network for development. *Science* **2002** 295:1669-1678.
81. Davidov E, Holland J, Marple E, Naylor S. Advancing drug discovery through systems biology. *Drug Discov Today.* **2003** 8: 175-83.
82. Dayhoff MO, Eck RV, Chang MA, Sochard MR. In "*Atlas of Protein Sequence and Structure*", National Biomedical Research Foundation. **1965**.
83. Dayhoff MO, Schwartz RM, Orcutt BC. A model of evolutionary change in proteins. In "*Atlas of Protein Sequence and Structure*." National Biomedical Research Foundation. **1978** 345-352.

84. de Brevern AG. New assessment of a structural alphabet. *In Silico Biol.* **2005** 5:283-9.
85. de Jong H. Modeling and simulation of genetic regulatory systems: a literature review. *J Comput Biol.* **2002** 9:67-103.
86. Deleage, G., Combet, C., Blanchet, C. and Geourjon, C. ANTHEPROT: an integrated protein sequence analysis software with client/server capabilities. *Comput Biol Med.* **2001** 31:259-267.
87. Delsuc F, Brinkmann H, Philippe H. Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet.* **2005** 6:361-75.
88. Denny M. Ontology Building: A Survey of Editing Tools. www.xml.com/pub/a/2002/11/06/ontologies.html.
89. de Parseval N, Heidmann T. Human endogenous retroviruses: from infectious elements to human genes. *Cytogenet Genome Res*. **2005** 110:318-32.
90. Devos D, Valencia A. Intrinsic errors in genome annotation. *Trends Genet* **2001** 17, 429-431.
91. Dickerson, R. E. and Geis, I. Hemoglobin: Structure, Function, Evolution, and Pathology. **1983** The Benjamin/Cummings Publishing Co., Inc., Menlo Park, California.
92. Do CB, Mahabhashyam MS, Brudno M, Batzoglou S. ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res.* **2005** 15:330-40.
93. Doolittle RF. Of URFs and ORFs: a primer on how to analyze derived amino acid sequences. University Science Books, Mill Valley California. **1986.**
94. Eddy SR. Profile hidden Markov models. *Bioinformatics* **1998** 14:755-763.
95. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **2004** 32:1792-7.
96. Edwards JS, Palsson BO. Systems properties of the Haemophilus influenzae Rd metabolic genotype. *J. Biol. Chem.* **1999** 274:17410−17416.
97. Eichler J, Adams MW. Posttranslational protein modification in Archaea. *Microbiol Mol Biol Rev*. **2005** 69:393-425.
98. Eilbeck K, Lewis SE, Mungall CJ, Yandell M, Stein L, Durbin R, Ashburner M. The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol*., **2005** 6, R44.
99. Eisen, J. and Fraser, C.M. Phylogenomics: intersection of evolution and genomics. *Science*, **2003** 300, 1706–1707.
100. Elnitski L, Giardine B, Shah P, et al. Improvements to GALA and dbERGE II: databases featuring genomic sequence alignment, annotation and experimental results. *Nucleic Acids Res* **2005** 33:D466-70.
101. Errami M, Geourjon C, Deleage G. Detection of unrelated proteins in sequences multiple alignments by using predicted secondary structures. *Bioinformatics* **2003** 19:506-512.
102. Etzold T, Argos P. SRS--an indexing and retrieval tool for flat file data libraries. *Comput Appl Biosci.* **1993** 9:49-57.
103. Farquhar A., Fikes R., and Rice J.P. The ontolingua server: A tool for collaborative ontology construction. *Journal of Human-Computer Studies*, **1997** 46:707-728.
104. Feng DF, Doolittle RF. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J Mol Evol.* **1987** 25:351-60.
105. Finkelstein AV. 3D protein folds: Homologs against errors—A simple estimate based on the random energy model. *Phys. Rev. Lett.* **1998** 80:4823–5.
106. Fitch WM. An improved method of testing for evolutionary homology. *J Mol Biol*. **1966** 16:9-16.
107. Forterre P., Philippe H. "Where is the root of the universal tree of life?" *BioEssays* **1999** 21:871-879.
108. Foster, I. The grid: computing without bounds. *Sci Am.* **2003** 288, 78-85.
109. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. VISTA: computational tools for comparative genomics. *Nucleic Acids Res*. **2004** 32:W273-9.
110. French S, Robson B. What is a conservative substitution? *J. Mol. Evol.* **1983** 19:171-175.
111. Gaasterland T, Sensen CW. Fully automated genome analysis that reflects user needs and preferences. A detailed introduction to the MAGPIE system architecture. *Biochimie* **1996** 78:302-310.
112. Galperin MY. The Molecular Biology Database Collection: 2006 update. *Nucleic Acids Res.* **2006** 34:D3-5.
113. Gan HH, Perlow RA, Roy S, Ko J, Wu M, Huang J, Yan S, Nicoletta A, Vafai J, Sun D, Wang L, Noah JE, Pasquali S, Schlick T. Analysis of protein sequence/structure similarity relationships. *Biophys J*. **2002** 83:2781-91.
114. Gardner PP, Wilm A, Washietl S. A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Res.* **2005** 33:2433-9.
115. Gardezi SA, Nguyen C, Malloy PJ, Posner GH, Feldman D, Peleg S. A rationale for treatment of hereditary vitamin D-resistant rickets with analogs of 1 alpha,25-dihydroxyvitamin D(3). *J Biol Chem*. **2001** 31: 29148-56.
116. Garfinkel D, Garfinkel L, Pring M, Green SB, Chance B. Computer applications to biochemical kinetics.

Annu. Rev. Biochem. **1970** 39:473−498.
117. Garnier N, Friedrich A, Bolze R, Bettler E, Moulinier L, Geourjon C, Thompson JD, Deleage G, Poch O. MAGOS: multiple alignment and modelling server. *Bioinformatics*. **2006** 22:2164-5.
118. Ge H, Walhout AJ, Vidal M. Integrating 'omic' information: a bridge between genomics and systems biology. *Trends Genet*. **2003** 19: 551-60.
119. George, R.A. and Heringa, J. SnapDRAGON: a method to delineate protein structural domains from sequence data. *J Mol Biol.* **2002** 316:839-851.
120. Gilbert W. The RNA World. *Nature* **1986** 319:618.
121. Gilbert DR, Schroeder M, van Helden J. Interactive Visualization and Exploration of Relationships between Biological Objects. *Trends in Biotechnology* **2000** 18:487-494.
122. Glaser F, Steinberg DM, Vakser IA, Ben-Tal N. Residue frequencies and pairing preferences at protein-protein interfaces. *Proteins.* **2001** 43:89-102.
123. Gouet P, Courcelle E. ENDscript: a workflow to display sequence and structure information. *Bioinformatics*. **2002** 18:767-768.
124. Gotoh O. Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments. *J Mol Biol* **1996** 264:823-838.
125. Gribskov M, McLachlan AD, Eisenberg D. Profile analysis: detection of distantly related proteins. *Proc Natl Acad Sci U S A.* **1987** 84:4355-8.
126. Grishin, N.V. and Phillips, M.A. The subunit interfaces of oligomeric enzymes are conserved to a similar extent to the overall protein sequences. *Protein Sci* **1994** 3:2455–2458
127. Gruber,T.R. Toward Principles for the design of ontologies used for knowledge sharing. In "*Formal Ontology in Conceptual Analysis and Knowledge Representation".* **1993** Kluwer Academic Publishers, Deventer, The Netherlands.
128. Hannenhalli SS, Hayes WS, Hatzigeorgiou AG, Fickett JW. Bacterial start site prediction. *Nucleic Acids Res.* **1999** 27:3577-82**.**
129. Hardison RC. Comparative genomics. *PLoS Biology* **2003** 1:E58.
130. Hass L, Schwartz P, Kodali P, Kotlar E, Rice J, Swope W. DiscoveryLink: a system for integrating life sciences data. *IBM Syst. J.* **2001** 40:489-511.
131. Hein J. Unified approach to alignment and phylogenies. *Methods in Enzymology* **1990** 183:626-645.
132. Heinrich R, Rapoport SM. Metabolic regulation and mathematical models. In *"Progress in Biophysics and Molecular Biology,* Vol. 32" **1977** Butler, J. A. V., Noble, D., Ed., Pergamon Press: Oxford, UK, pp 1-82.
133. Henikoff JG, Greene EA, Pietrokovski S, Henikoff S. Increased coverage of protein families with the blocks database servers. *Nucleic Acids Res*. **2000** 28:228-230.
134. Henikoff S. Beyond the Central Dogma. *Bioinformatics*, **2002** 18, 223-225.
135. Hermjakob H, Montecchi-Palazzi L, Bader G, Wojcik J, Salwinski L, Ceol A, Moore S, Orchard S, Sarkans U, von Mering C, Roechert B, Poux S, Jung E, Mersch H, Kersey P, Lappe M, Li Y, Zeng R, Rana D, Nikolski M, Husi H, Brun C, Shanker K, Grant SG, Sander C, Bork P, Zhu W, Pandey A, Brazma A, Jacq B, Vidal M, Sherman D, Legrain P, Cesareni G, Xenarios I, Eisenberg D, Steipe B, Hogue C, Apweiler R. The HUPO PSI's molecular interaction format--a community standard for the representation of protein interaction data. Nat Biotechnol., **2004** 22, 177-183.
136. Hertz GZ, Stormo GD. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*. **1999** 15:563-577.
137. Higgins DG, Sharp PM. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene*. **1988** 73:237-44.
138. Hoersch S, Leroy C, Brown NP, Andrade MA, Sander C. The GeneQuiz web server: protein functional analysis through the Web. *Trends Biochem Sci.* **2000** 25:33-35.
139. Hofacker IL, Bernhart SH, Stadler PF. Alignment of RNA base pairing probability matrices. *Bioinformatics*. **2004** 20:2222-7.
140. Holbrook SR. RNA structure: The long and the short of it. *Curr. Opin. Struct. Biol*. **2005** 15:302–308.
141. Hood L., Ideker T., and Galitski T. A new approach to decoding life: Systems biology. *In Annual Review of Genomics and Human Genetics* **2001** volume 2, pages 343-372.
142. Hornberg JJ, Binder B, Bruggeman FJ, Schoeberl B, Heinrich R, Westerhoff HV. Control of MAPK signalling: from complexity to what really matters. *Oncogene*. **2005** 24:5533-42.
143. Hsu F, Pringle TH, Kuhn RM, Karolchik D, Diekhans M, Haussler D, Kent WJ. The UCSC Proteome Browser. *Nucleic Acids Res.* **2005** 33:D454-8.
144. Hubbard, T.J., Ailey, B., Brenner, S.E., Murzin, A.G., and Chothia, C. SCOP: A structural classification of proteins database. *Nucleic Acids Res.* **1999** 27:254–256.

145. Hucka M, Finney A, Sauro HM, Bolouri H, Doyle J, Kitano H. The ERATO Systems Biology Workbench: enabling interaction and exchange between software tools for computational biology. *Pac Symp Biocomput.* **2002** 450-61.
146. Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, Langendijk-Genevaux PS, Pagni M, Sigrist CJ. The PROSITE database. *Nucleic Acids Res.* **2006** 34:D227-30.
147. Hunter A, Kaufman MH, McKay A, Baldock R, Simmen MW, Bard JB. An ontology of human developmental anatomy. *J Anat*. **2003** 203:347-55.
148. Hunter PJ. The IUPS Physiome Project: a framework for computational physiology. *Prog Biophys Mol Biol*. **2004** 85:551-69.
149. Ideker T, Galitski T, Hood L. A new approach to decoding life: systems biology. *Annu Rev Genomics Hum Genet*. **2001a** 2: 343-72.
150. Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R, Hood L. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*. **2001b** 292:929-34.
151. Ilyin VA, Pieper U, Stuart AC, Marti-Renom MA, McMahan L, Sali A. ModView, visualization of multiple protein sequences and structures. *Bioinformatics*. **2003** 19:165-6.
152. International Mouse Genome Sequencing Consortium Initial sequencing and comparative analysis of the mouse genome. *Nature* **2002** 420: 520–562.
153. IwabeN, Kuma K, Hasegawa M, Osawa S, Miyata T. Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc. Natl Acad. Sci. USA*, **1989** 86, 9355–9359.
154. Jareborg N, Durbin R: Alfresco-A workbench for comparative genomic sequence analysis. *Genome Res.* **2000** 10:1148-1157.
155. Jensen LJ, Saric J, Bork P. Literature mining for the biologist: from information retrieval to biological discovery. *Nat Rev Genet*. **2006** 7:119-29.
156. Johnson JM, Mason K, Moallemi C, Xi H, Somaroo S, Huang ES. Protein family annotation in a multiple alignment viewer. *Bioinformatics*. **2003** 19:544-5.
157. Jones S, Thornton JM. Principles of protein-protein interactions. *Proc Natl Acad Sci U S A*. **1996** 93:13-20.
158. Jossinet F, Westhof E. Sequence to Structure (S2S): display, manipulate and interconnect RNA data from sequence to structure. *Bioinformatics*. **2005** 21:3320-1.
159. Kabsch W, Sander C. On the use of sequence homologies to predict protein structure: identical pentapeptides can have completely different conformations. *Proc Natl Acad Sci USA*. **1984** 81:1075-8.
160. Kanehisa M. The KEGG database. *Novartis Found Symp*. **2002** 247:91-101.
161. Kanehisa M, Bork P. Bioinformatics in the post-sequence era. *Nat Genet*. **2003**:33 Suppl:305-10.
162. Karchin R, Kelly L, Sali A. Improving functional annotation of non-synonomous SNPs with information theory. *Pac Symp Biocomput*. **2005**:397-408
163. Karp P, Riley M, Paley S, Pellegrini-Toole A, Krummenacker M. EcoCyc: Electronic Encyclopedia of E. coli Genes and Metabolism. *Nucl. Acids Res.* **1999a** 27:55-58.
164. Karp PD, Chaudhri VK, and Paley SM. A Collaborative Environment for Authoring Large Knowledge Bases. *Journal of Intelligent Information Systems*, **1999** 13:155-194.
165. Karplus K, Barrett C, Hughey R. Hidden Markov models for detecting remote protein homologies. *Bioinformatics* **1998** 10:846-856.
166. Kasprzyk A, Keefe D, Smedley D, London D, Spooner W, Melsopp C, Hammond M, Rocca-Serra P, Cox T, Birney E. EnsMart: A Generic System for Fast and Flexible Access to Biological Data. *Genome Research* **2004** 14:160-9.
167. Katoh, K., Misawa, K., Kuma, K. and Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **2002** 30:3059-3066.
168. Kidd KK, Pakstis AJ, Speed WC, Kidd JR. Understanding human DNA sequence variation. *J Hered.* **2004** 95:406-20.
169. Kim Y, Subramaniam S. Locally defined protein phylogenetic profiles reveal previously missed protein interactions and functional relationships. *Proteins*. **2006** 62:1115-24.
170. Kitano H. Computational systems biology. *Nature*. **2002** 420:206-10.
171. Koehl P, Delarue M. Polar and nonpolar atomic environments in the protein core: implications for folding and binding. *Proteins*. **1994** 20:264-78.
172. Koehl P, Levitt M. Structure-based conformational preferences of amino acids. *Proc Natl Acad Sci U S A.* **1999** 96:12524-9.
173. Koehl P. Protein structure similarities. *Curr Opin Struct Biol.* **2001** 11:348-53.
174. Koehl P, Levitt M. Sequence variations within protein families are linearly related to structural variations. *J Mol Biol.* **2002a** 323:551-62.

175. Koehl P, Levitt M. Protein topology and stability define the space of allowed sequences. *Proc Natl Acad Sci U S A.* **2002b** 99:1280-5.
176. Kohler J, Schulze-Kremer S. The semantic metadatabase (SEMEDA): ontology based integration of federated molecular biological data sources. *In Silico Biol*. **2002** 2:219-31.
177. Kolodny R, Koehl P, Guibas L, Levitt M. Small libraries of protein fragments model native protein structures accurately. *J Mol Biol.* **2002** 323:297-307.
178. Kolodny R, Koehl P, Levitt M. Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. *J Mol Biol*. **2005** 346:1173-88.
179. Korenberg JR, Rimoin DL. *Medical genetics. JAMA*. **1995** 273:1692-3.
180. Kouranov A, Xie L, de la Cruz J, Chen L, Westbrook J, Bourne PE, Berman HM. The RCSB PDB information portal for structural genomics. *Nucleic Acids Res*. **2006** 34:D302-5.
181. Kumar SP, Feidler JC. BioSPICE: A computational infrastructure for integrative biology. *OMICS: J Integrative Biol* **2003** 7: 225–225
182. Kyte and Doolittle. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **1982** 157:105–132.
183. Lambrix P, Habbouche M, Perez M. Evaluation of ontology development tools for bioinformatics. *Bioinformatics*. **2003** 19:1564-71.
184. Laskowski RA, Luscombe NM, Swindells MB, Thornton JM. Protein clefts in molecular recognition and function. *Protein Sci*. **1996** 5:2438-52.
185. Lassmann T, Sonnhammer EL. Automatic assessment of alignment quality. *Nucleic Acids Res*. **2005** 33:7120-8
186. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **1993** 262:208-214.
187. Lecompte O, Thompson JD, Plewniak F, Thierry J, Poch O. Multiple alignment of complete sequences (MACS) in the post-genomic era. *Gene*. **2001** 270:17-30.
188. Lecompte, O., Ripp, R., Thierry, J.C., Moras, D. and Poch, O. Comparative analysis of ribosomal proteins in complete genomes: an example of reductive evolution at the domain scale. *Nucleic Acids Res*. **2002** 30:5382-5390.
189. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne JB, Volkert TL, Fraenkel E, Gifford DK, Young RA.Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*. **2002a** 298:799-804.
190. Lee C, Grasso C, Sharlow MF. Multiple sequence alignment using partial order graphs. *Bioinformatics*. **2002b** 18:452-64.
191. Lee S, Lee BC, Kim D. Prediction of protein secondary structure content using amino acid composition and evolutionary information. *Proteins*. **2006** 62:1107-14.
192. Lenz GR, Nash HM, Jindal S. Chemical ligands, genomics and drug discovery. *Drug Discov Today*. **2000** 5:145-156.
193. Leontis NB, Westhof E. Analysis of RNA motifs. *Curr Opin Struct Biol* **2003** 13:300–308.
194. Leontis NB, Altman RB, Berman HM, Brenner SE, Brown JW, Engelke DR, Harvey SC, Holbrook SR, Jossinet F, Lewis SE, Major F, Mathews DH, Richardson JS, Williamson JR, Westhof E. The RNA Ontology Consortium: an open invitation to the RNA community. *RNA*. **2006** 12:533-41.
195. Lesk AM. Computational molecular biology. In *A. Kent and J.G. Williams, editors, Encyclopedia of Computer Science and Technology,* Vol. 31, pages 101-165. Marcel Dekker, **1994**.
196. Letunic I, Copley RR, Pils B, Pinkert S, Schultz J, Bork P. SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res*. **2006** 34:D257-60.
197. Levitt,M. Protein folding by restrained energy minimization and molecular dynamics. *J. Mol. Biol*. **1983** 170:723–764.
198. Li W. The-more-the-better and the-less-the-better. *Bioinformatics*. **2006** 22:2187-8.
199. Li W, Godzik A. VISSA: a program to visualize structural features from structure sequence alignment. *Bioinformatics*. **2006** 22:887-8.
200. Lichtarge O, Bourne HR, Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol* **1996** 257:342-58.
201. Lipman DJ, Pearson WR. Rapid and sensitive protein similarity searches. *Science*. **1985** 227:1435-41.
202. Lipman DJ, Altschul SF, Kececioglu JD. A tool for multiple sequence alignment. *Proc Natl Acad Sci U S A.* **1989** 86:4412-5.
203. Liu ET. Systems biology, integrative biology, predictive biology. *Cell* **2005** 121:505-6.
204. Livingstone CD, Barton GJ. Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Comput Appl Biosci* **1993** 9:745-756.
205. Loomis W, Thomas S. Kinetic analysis of biochemical differentiation in Dictyostelium discoideum. *J.*

Biol. Chem. **1976** 251:6252−6258.
206. Lopez-Garcia P, Moreira D. Metabolic symbiosis at the origin of eukaryotes. *Trends Biochem Sci*. **1999** 24:88-93.
207. Luthy R, McLachlan AD, Eisenberg D. Secondary structure-based profiles: use of structure-conserving scoring tables in searching protein sequence databases for structural similarities. *Proteins*. **1991** 10:229-39.
208. Madabushi S, Yao H, Marsh M, Kristensen DM, Philippi A, Sowa ME, Lichtarge O. Structural clusters of evolutionary trace residues are statistically significant and common in proteins. *J Mol Biol*. **2002** 316:139-54.
209. Makarova KS, Koonin EV. Comparative genomics of Archaea: how much have we learned in six years, and what's next? *Genome Biol* **2003** 4: 115.
210. Malloy PJ, Pike JW. Feldman D. The Vitamin D Receptor and the Syndrome of Hereditary 1,25-Dihydroxyvitamin D-Resistant Rickets. *Endocrine Reviews* **1999** 20: 156-88.
211. Margulies EH, Chen CW, Green ED. Differences between pair-wise and multi-sequence alignment methods affect vertebrate genome comparisons. *Trends Genet.* **2006** 22:187-93.
212. Mathe C, Sagot MF, Schiex T, Rouze P. Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res*. **2002** 30: 4103-17.
213. Mattick, JS. Non-coding RNAs: the architects of eukaryotic complexity. *EMBO Rep*. **2001** 2:986-991.
214. McClelland M, Florea L, Sanderson K, et al. Comparison of the Escherichia coli K-12 genome with sampled genomes of a Klebsiella pneumoniae and three salmonella enterica serovars, Typhimurium, Typhi and Paratyphi. *Nucleic Acids Res* **2000** 28: 4974-86.
215. McClure MA, Vasi TK and Fitch WM Comparative analysis of multiple protein sequence alignment methods. *Molecular Biology and Evolution* **1994** 11:571-592.
216. Medigue C, Rechenmann F, Danchin A, Viari A. Imagene: an integrated computer environment for sequence annotation and analysis. *Bioinformatics*. **1999** 15:2-15.
217. Miyazawa S, Jernigan RL. A new substitution matrix for protein sequence searches based on contact frequencies in protein structures. Protein Eng. 1993 6:267-78.
218. Mizuguchi K, Deane CM, Blundell TL, Overington JP. HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Sci.* **1998** 7 2469-71.
219. Moelbert S, Emberly E, Tang C. Correlation between sequence hydrophobicity and surface-exposure pattern of database proteins. *Protein Sci*. **2004** 13:752-62.
220. Morgenstein B, Dress A, Werner T. Multiple DNA and protein sequence alignment based on segment-to-segment comparison. *Proc Natl Acad Sci USA*. **1996** 93:12098-103.
221. Morrison DA, Ellis JT. Effects of nucleotide sequence alignment on phylogeny estimation: a case study of 18S rDNAs of apicomplexa. *Mol Biol Evol*. **1997** 14:428-41.
222. Moult J. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr Opin Struct Biol.* **2005** 15:285-9.
223. Mourier T. Reverse transcription in genome evolution. *Cytogenet Genome Res.* **2005** 110:56-62.
224. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bradley P, Bork P, Bucher P, Cerutti L, Copley R, Courcelle E, Das U, Durbin R, Fleischmann W, Gough J, Haft D, Harte N, Hulo N, Kahn D, Kanapin A, Krestyaninova M, Lonsdale D, Lopez R, Letunic I, Madera M, Maslen J, McDowall J, Mitchell A, Nikolskaya AN, Orchard S, Pagni M, Ponting CP, Quevillon E, Selengut J, Sigrist CJ, Silventoinen V, Studholme DJ, Vaughan R, Wu CH. InterPro, progress and status in 2005. *Nucleic Acids Res*. **2005** 33:D201-5.
225. Muller, A., MacCallum, R.M., and Sternberg, M. Benchmarking PSI-BLAST in genome annotation. *J. Mol. Biol*. **1999** 293:1257–1271.
226. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **1970** 48: 443-453.
227. Neuwald AF, Liu JS, Lipman DJ, Lawrence CE. Extracting protein alignment models from the sequence database. *Nucleic Acids Res*. **1997** 25:1665-1677.
228. Niehaus F, Bertoldo C, Kahler M, Antranikian G. Extremophiles as a source of novel enzymes for industrial application. *Appl Microbiol Biotechnol*. **1999** 51:711-729.
229. Ng PC, Henikoff JG, Henikoff S. PHAT: a transmembrane-specific substitution matrix. *Bioinformatics* **2000** 16:760–766.
230. Noller HF. RNA structure: reading the ribosome. *Science*. **2005** 309:1508-14.
231. Notredame C, Higgins DG. SAGA: sequence alignment by genetic algorithm. *Nucleic Acids Res* **1996** 24:1515-1524.
232. Notredame, C., Holm, L. and Higgins, D.G. COFFEE: an objective function for multiple sequence alignments. *Bioinformatics.* **1998** 14:407-422.

233. Notredame, C., Higgins, D.G., Heringa, J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* **2000** 302:205-217.
234. Novak B, Tyson JJ. Quantitative analysis of a molecular model of mitotic control in fission yeast. *J. Theor. Biol.* **1995** 173:283−305.
235. O'Donoghue SI, Meyer JE, Schafferhans A, Fries K: The SRS 3D module: integrating structures, sequences and features. *Bioinformatics*. **2004** 20:2476-2478.
236. Oldfield TJ. Data mining the Protein Data Bank: residue interactions. *Proteins* **2002** 49**:**510-529.
237. Ondrechen MJ, Clifton JG, Ringe D. THEMATICS: a simple computational predictor of enzyme function from structure. *Proc Natl Acad Sci U S A.* **2001** 98:12473-8.
238. Owen MJ, Craddock N, O'Donovan MC. Schizophrenia: genes at last? *Trends Genet*. **2005** 21:518-25.
239. Oyama T, Kitano K, Satou K, Ito T. Extraction of knowledge on protein-protein interaction by association rule discovery. *Bioinformatics* **2002** 18:705-714.
240. Page RDM, Holmes EC. Molecular Evolution: a Phylogenetic Approach, Blackwell Science. **1998**.
241. Pagni M, Ioannidis V, Cerutti L, Zahn-Zabal M, Jongeneel CV, Falquet L: MyHits: a new interactive resource for protein annotation and domain identification. *Nucleic Acids Res* **2004** 32:W332-325.
242. Pazos F, Helmer-Citterich M, Ausiello G, Valencia A. Correlated mutations contain information about protein-protein interaction. J Mol Biol. 1997 271:511-23.
243. Pazos F, Valencia A. Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng*. **2001** 14:609-14.
244. Pazos F, Valencia A. In silico two-hybrid system for the selection of physically interacting protein pairs. *Proteins* **2002** 47:219-27.
245. Pearson WR. Empirical statistical estimates for sequence similarity searches. *J Mol Biol.* **1998** 276:71-84.
246. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A.* **1988** 85:2444-8.
247. Pei J, Grishin NV. AL2CO: calculation of positional conservation in a protein sequence alignment. *Bioinformatics*. **2001** 17:700-712.
248. Pennisi E. Genome data shake tree of life. *Science*. **1998** 280:672-674.
249. Phillips A, Janies D, Wheeler W. Multiple sequence alignment in phylogenetic analysis. *Mol Phylogenet Evol*. **2000** 16:317-330.
250. Pieper U, Eswar N, Davis FP, Braberg H, Madhusudhan MS, Rossi A, Marti-Renom M, Karchin R, Webb BM, Eramian D, Shen MY, Kelly L, Melo F, Sali A. MODBASE: a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res.* **2006** 34:D291-5.
251. Pintar A, Carugo O, Pongor S. DPX: for the analysis of the protein core. *Bioinformatics*. **2003** 19:313-4.
252. Plewniak F, Thompson JD, Poch O. Ballast: blast post-processing based on locally conserved segments. *Bioinformatics*. **2000** 16:750-9.
253. Plewniak F, Bianchetti L, Brelivet Y, Carles A, Chalmel F, Lecompte O, Mochel T, Moulinier L, Muller A, Muller J, Prigent V, Ripp R, Thierry JC, Thompson JD, Wicker N, Poch O. PipeAlign: A new toolkit for protein family analysis. *Nucleic Acids Res*. **2003** 31:3829-32.
254. Poole A, Jeffares D, Penny D. Early evolution: prokaryotes, the new kids on the block. *Bioessays*. **1999** 21:880-9.
255. Poupon A, Mornon JP. Populations of hydrophobic amino acids within protein globular domains: identification of conserved "topohydrophobic" positions. *Proteins*. **1998** 33:329-42.
256. Raghava GP, Searle SM, Audley PC, Barber JD, Barton GJ. OXBench: a benchmark for evaluation of protein multiple sequence alignment accuracy. *BMC Bioinformatics*. **2003** 4:47.
257. Ramensky V, Bork P, Sunyaev S. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.* **2002** 30: 3894-900.
258. Reddy BV, Li WW, Shindyalov IN, Bourne PE. Conserved key amino acid positions (CKAAPs) derived from the analysis of common substructures in proteins. *Proteins*. **2001** 42:148-63.
259. Regnstrom K, Burgess DJ. Pharmacogenomics and its potential impact on drug and formulation development. *Crit Rev Ther Drug Carrier Syst*. **2005** 22:465-92.
260. Richards, F.M. Areas, volumes, packing and protein structure. *Ann. Rev. Biophys. Bioeng.,* **1977** 6:151-176
261. Rigoutsos I, Huynh T, Floratos A, Parida L, Platt D: Dictionary-driven protein annotation. *Nucleic Acids Res* **2002** 30:3901-3916.
262. Roberts RJ. Identifying protein function--a call for community action. *PLoS Biol*. **2004** 2:E42.
263. Robinson AJ, Flores TP: Novel Techniques for Visualising Biological Information. *In Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology.* **1996** 241-249.
264. Rochel N, Wurtz JM, Mitschler A, Klaholz B, Moras D. The crystal structure of the nuclear receptor for vitamin D bound to its natural ligand. *Mol Cell*. **2000** 5: 173-9.

265. Rodi DJ, Mandava S, Makowski L. DIVAA: analysis of amino acid diversity in multiple aligned protein sequences. *Bioinformatics*. **2004** 20:3481-9.
266. Rogers J.E., Solomon W.D., Rector A.L., Pole P.M., Zanstra P., and van der Haring E. Rubrics to Dissections to GRAIL to Classifications. *In Medical Informatics, Europe* **1997** 241-245.
267. Rojas I, Ratsch E, Saric J, Wittig U. Notes on the use of ontologies in the biochemical domain. *In Silico Biology* **2003** 4:0009.
268. Roos DS. Computational biology. Bioinformatics--trying to swim in a sea of data. *Science*. **2001** 291:1260-1.
269. Rossmann,M.G. and Argos,P. Exploring structural homology of proteins. J. Mol. Biol. 1976 105:75–95.
270. Rubin GM, Yandell MD, Wortman JR, Gabor Miklos GL, Nelson CR, Hariharan IK, Fortini ME, Li PW, Apweiler R, Fleischmann W, Cherry JM, Henikoff S, Skupski MP, Misra S, Ashburner M, Birney E, Boguski MS, Brody T, Brokstein P, Celniker SE, Chervitz SA, Coates D, Cravchik A, Gabrielian A, Galle RF, Gelbart WM, George RA, Goldstein LS, Gong F, Guan P, Harris NL, Hay BA, Hoskins RA, Li J, Li Z, Hynes RO, Jones SJ, Kuehl PM, Lemaitre B, Littleton JT, Morrison DK, Mungall C, O'Farrell PH, Pickeral OK, Shue C, Vosshall LB, Zhang J, Zhao Q, Zheng XH, Lewis S. Comparative genomics of the eukaryotes. *Science* **2000** 287: 2204-15.
271. Ruvolo, M. Comparative primate genomics: the year of the chimpanzee. *CUrr Opin Genet Dev*. **2004** 14:650-656.
272. Said MR, Begley TJ, Oppenheim AV, Lauffenburger DA, Samson LD. Global network analysis of phenotypic effects: protein networks and toxicity modulation in Saccharomyces cerevisiae. *Proc Natl Acad Sci U S A.* **2004** 101: 18006-11.
273. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Evol. Biol* **1987** 4:406-425
274. Sammeth M, Heringa J. Global multiple-sequence alignment with repeats. *Proteins*. **2006** 64:263-74.
275. Sankoff, D. Minimal mutation trees of sequences. *SIAM J. Appl. Math.,* **1975** 28:35-42.
276. Sayle RA, Milner-White EJ. RASMOL: biomolecular graphics for all. *Trends Biochem Sci*. **1995** 20:374.
277. Schaff J., Loew L.M. "The Virtual Cell", *Pacific Symposium on Biocomputing*, **1999** 4:228-239.
278. Schneider TD, Stephens RM. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* **1990** 18:6097-100.
279. Schoeberl B, Eichler-Jonsson C, Gilles ED, Muller G. Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors. *Nat Biotechnol*. **2002** 20:370-5.
280. Schuler GD, Epstein JA, Ohkawa H, Kans JA. Entrez: molecular biology database and retrieval system. *Methods Enzymol*. **1996** 266:141-62.
281. Schulze-Kremer, S. Adding Semantics to Genome Databases: Towards an Ontology for Molecular Biology. *Proc. Int. Conf. Intell. Syst. Mol. Biol*. **1997** 5, 272-275.
282. Schulze-Kremer S. Ontologies for molecular biology and bioinformatics. *In Silico Biol*. **2002** 2:179-93.
283. Senger M, Flores T, Glatting K, Ernst P, Hotz-Wagenblatt A, Suhai S. W2H: WWW interface to the GCG sequence analysis package. *Bioinformatics*. **1998** 14:452-7.
284. Shah SP, He DY, Sawkins JN, Druce JC, Quon G, Lett D, Zheng GX, Xu T, Ouellette BF. Pegasys: software for executing and integrating analyses of biological sequences. *BMC Bioinformatics*. **2004** 5:40.
285. Shah SP, Huang Y, Xu T, Yuen MMS, Ling J, Ouellette BFF: Atlas: a data warehouse for integrative bioinformatics. *BMC Bioinformatics* **2005** 6:34.
286. Shannon W, Culverhouse R, Duncan J. Analyzing microarray data using cluster analysis. *Pharmacogenomics*. **2003** 4:41-52.
287. Shapiro BA, Kasprzak W. STRUCTURELAB: a heterogeneous bioinformatics system for RNA structure analysis. *J Mol Graph*. **1996** 14:194-205.
288. Shapiro JA. A 21st century view of evolution: genome system architecture, repetitive DNA, and natural genetic engineering. *Gene* **2005** 345:91-100.
289. Sharan R, Ideker T. Modeling cellular machinery through biological network comparison. *Nat Biotechnol.* **2006** 24:427-33.
290. Shindyalov, I.N. and Bourne, P.E. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.* **1998** 11**:** 739–747.
291. Sim SE, Easterbrook S, Holt RC. Using benchmarking to advance research: a challenge to software engineering. *25th International Conference on Software Engineering* **2003** :74-83.
292. Simossis VA, Heringa J. PRALINE: a multiple sequence alignment toolbox that integrates homology-extended and secondary structure information. *Nucleic Acids Res.* **2005** 33:W289-94.
293. Smagala JA, Dawson ED, Mehlmann M, Townsend MB, Kuchta RD, Rowlen KL. ConFind: a robust tool for conserved sequence identification. *Bioinformatics*. **2005** 21:4420-2.

294. Smith B, Ceusters W, Klagges B, Kohler J, Kumar A, Lomax J, Mungall CJ, Neuhaus F, Rector A, Rosse C. Relations in Biomedical Ontologies.*Genome Biology*. **2005** 6:R46.
295. Smith, R.F. and Smith, T.F. Pattern-induced multi-sequence alignment (PIMA) algorithm employing secondary structure-dependent gap penalties for use in comparative protein modelling. *Protein Eng* **1992** 5, 35-41.
296. Smith TF, Waterman MS. Identification of common molecular subsequences *J. Mol. Biol*. **1981** 215:403-10.
297. Sneath PH, Sokal RR. In "*Numerical taxonomy*". **1973** W. H. Freeman. San Francisco.
298. Snyder M, Gerstein M. Genomics. Defining genes in the genomics era. *Science*. **2003** 300:258-60.
299. Socolich M, Lockless SW, Russ WP, Lee H, Gardner KH, Ranganathan R. Evolutionary information for specifying a protein fold. *Nature*. **2005** 437:512-8.
300. States D J, Boguski MS. Similarity and homology. In "Sequence Analysis Primer" New York: Stockton Press **1991**.
301. Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, Abeysinghe S, Krawczak M, Cooper DN. Human gene mutation database (HGMD): 2003 update. *Hum. Mutat*., **2003** 21:577–581.
302. Stoetzel C, Laurier V, Davis EE, Muller J, Rix S, Badano JL, Leitch CC, Salem N, Chouery E, Corbani S, Jalk N, Vicaire S, Sarda P, Hamel C, Lacombe D, Holder M, Odent S, Holder S, Brooks AS, Elcioglu NH, Da Silva E, Rossillion B, Sigaudy S, de Ravel TJ, Alan Lewis R, Leheup B, Verloes A, Amati-Bonneau P, Megarbane A, Poch O, Bonneau D, Beales PL, Mandel JL, Katsanis N, Dollfus H. BBS10 encodes a vertebrate-specific chaperonin-like protein and is a major BBS locus. *Nat Genet*. **2006** 38:521-4.
303. Taylor WR. Multiple sequence alignment by a pairwise algorithm. *J Mol Evol* **1988** 28: 161-169.
304. Taylor WR, Orengo CA. Protein structure alignment. *J Mol Biol*. **1989** 208:1-22.
305. Taylor WR. Protein structure comparison using SAP. *Methods Mol Biol*. **2000** 143:19-32.
306. Teng CM. Applying Noise Handling Techniques to Genomic Data: A Case Study. *International Conference on Data Mining* **2003** 743-746.
307. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **1994** 22:4673-80.
308. Thompson JD. Introducing variable gap penalties to sequence alignment in linear space. *Comput Appl Biosci*. **1995** 11:181-6.
309. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res*. **1997** 25:4876-82.
310. Thompson JD, Plewniak F, Poch O. BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics.* **1999a** 15:87-8.
311. Thompson JD, Plewniak F, Poch O. A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res*. **1999b** 27:2682-90.
312. Thompson JD, Harvey A, Wilson C. ClustalG: Software for analysis of activities and sequential events. *Workshop on Longitudinal Research in Social Science: A Canadian Focus, London, Ontario, Canada*, October 25-27, **1999c**.
313. Thompson, J.D., Plewniak, F., Thierry, J.C. and Poch, O. DbClustal: rapid and reliable global multiple alignments of protein sequences detected by database searches. *Nucleic Acids Res* **2000 28**:2919-2926.
314. Thompson JD, Plewniak F, Ripp R, Thierry JC, Poch O. Towards a reliable objective function for multiple sequence alignments. *J Mol Biol.* **2001** 314:937-51.
315. Thompson JD, Thierry JC, Poch O. RASCAL: rapid scanning and correction of multiple sequence alignments. *Bioinformatics*. **2003** 19:1155-61.
316. Thompson JD, Prigent V, Poch O. LEON: multiple aLignment Evaluation Of Neighbours. *Nucleic Acids Res*. **2004** 32:1298-307.
317. Thompson JD, Koehl P, Ripp R, Poch O. BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. *Proteins.* **2005a** 61:127-36.
318. Thompson JD, Holbrook SR, Katoh K, Koehl P, Moras D, Westhof E, Poch O. MAO: a Multiple Alignment Ontology for nucleic acid and protein sequences. *Nucleic Acids Res*. **2005b** 33:4164-71.
319. Thompson,JD, Poch,O. Multiple sequence alignment as a workbench for molecular systems biology. Current *Bioinformatics*. **2006a** 1:95-104.
320. 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*. **2006b** 7:318.
321. Thornton JM, Todd AE, Milburn D, Borkakoti N, Orengo C. From structure to function: approaches and limitations. *Nat Struct Biol* **2000** 7:991–994.

322. Tomita M., Whole-cell simulation: a grand challenge of the 21st century. *Trends Biotechnol*. **2001** 19:205-10.
323. Topaloglou T, Kosky A, Markowitz V. Seamless integration of biological applications within a database framework. *Proc Int Conf Intell Syst Mol Biol.* **1999** 272-81.
324. Towell GG, Shavlik JW. The extraction of refined rules from knowledge based neural networks. *Machine learning*. **1993** 131:71-101.
325. Tress,M.L., Jones,D. and Valencia,A. Predicting reliable regions in protein alignments from sequence profiles. *J Mol Biol* **2003** 330:705-718.
326. Uetz P, Finley RL Jr. From protein networks to biological systems. *FEBS Lett.* **2005** 579: 1821-7.
327. Valdar WS. Scoring residue conservation. *Proteins*. **2002** 48:227-41.
328. Van Walle I, Lasters I, Wyns L. Align-m--a new algorithm for multiple alignment of highly divergent sequences. *Bioinformatics*. **2004** 20:1428-35.
329. Van Walle I, Lasters I, Wyns L. SABmark--a benchmark for sequence alignment that covers the entire known fold space. *Bioinformatics*. **2005** 21:1267-8.
330. Vingron M, Sibbald PR. Weighting in sequence space: a comparison of methods in terms of generalized sequences. *Proc Natl Acad Sci U S A.* **1993** 90:8777-81.
331. von Dassow G, Meir E, Munro EM, Odel GM. The segment polarity network is a robust developmental module. *Nature* **2000** 406:188-192.
332. von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*. **2002** 417:399-403.
333. Wallace IM, O'Sullivan O, Higgins DG, Notredame C. M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res*. **2006** 34:1692-9.
334. Wang ZX. A re-estimation of the total numbers of protein folds and superfamilies. *Protein Eng*. **1998** 11:621–626.
335. Wang,L. and Xu,Y. SEGID: identifying interesting segments in (multiple) sequence alignments. *Bioinformatics* **2003** 19:297-298.
336. Watson JD, Crick FH. The structure of DNA. *Cold Spring Harb Symp Quant Biol*. **1953** 18:123-31.
337. Watson JD, Laskowski RA, Thornton JM. Predicting protein function from sequence and structural data. *Curr Opin Struct Biol*. **2005** 15:275-84.
338. Wei L, Liu Y, Dubchak I, Shon J, Park J. Comparative genomics approaches to study organism similarities and differences. *J Biomed Inform* **2002** 35: 142-50.
339. Wicker, N., Perrin, G.R., Thierry, J.C. and Poch, O. Secator: a program for inferring protein subfamilies from phylogenetic trees. *Mol Biol Evol.* **2001** 18:1435-1441.
340. Wicker, N., Dembele, D., Raffelsberger, W. and Poch, O. Density of points clustering, application to transcriptomic data analysis. *Nucleic Acids Res*. **2002** 30:3992-4000.
341. Wilkinson MD, Gessler DD, Farmer A, Stein L. The BioMOBY project explores open-source, simple, extensible protocols for enabling biological database interoperability. *Proc Virt Conf Genom and Bioinf* **2003** 3:16-26.
342. Wilson CA, Kreychman J, Gerstein M. Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J. Mol. Biol*. **2000** 17:233-249.
343. Wilson C. Reliability of sequence-alignment analysis of social processes: Monte Carlo tests of ClustalG software. *Environment and Planning* **2006** 38:187–204.
344. Woese CR. The evolution of the genetic code. In *"The Genetic Code"*. New York: Harper & Row; **1967**:179-195.
345. Woese CR, Pace NR. Probing RNA structure, function and history by comparative analysis. In "*The RNA World*". Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY. **1993**.
346. Wong L. Technologies for integrating biological data. *Brief Bioinform*. **2002** 3:389-404.
347. Wright BE. The use of kinetic models to analyze differentiation. *Behavioral Sci*. **1970** 15:37−45.
348. Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Mazumder R, O'Donovan C, Redaschi N, Suzek B. The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res*. **2006** 34:D187-91.
349. Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM, Eisenberg D. DIP: the database of interacting proteins. *Nucleic Acids Res.* **2000** 28:289-91.
350. Yang AS, Honig B. An integrated approach to the analysis and modeling of protein sequences and structures. II. On the relationship between sequence and structural similarity for proteins that are not obviously related in sequence. *J Mol Biol.* **2000** 301:679-89.
351. Ye L, Huang X. MAP2: multiple alignment of syntenic genomic sequences. *Nucleic Acids Res* **2005** 33: 162-70.

352. Yona G, Levitt M. Within the twilight zone: a sensitive profile-profile comparison tool based on information theory. *J Mol Biol*. **2002** 315:1257-1275.
353. Young L, Jernigan RL, Covell DG. A role for surface hydrophobicity in protein-protein recognition. *Protein Sci.* **1994** 3:717-29.
354. Yusupov MM, Yusupova GZ, Baucom A, Lieberman K, Earnest TN, Cate JH, Noller HF. Crystal structure of the ribosome at 5.5 A resolution. *Science*. **2001** 292:883-96.
355. Zeleny M. Management support systems: towards integrated knowledge management. *Human systems management.* **1987** 1:59-70.
356. Zhang MQ. Computational prediction of eukaryotic protein-coding genes. *Nat Rev Genet*. **2002** 3:698-709.
357. Zuker M. On finding all suboptimal foldings of an RNA molecule. *Science* **1989** 244:48–52.