Thèse présentée pour obtenir le titre de

Docteur de l'Université Louis Pasteur Strasbourg 1

Discipline : Sciences du Vivant

Spécialité : Bioinformatique

par

Jean MULLER

Analyse du cytosquelette par des approches bioinformatiques à haut débit de génomique comparative et de transcriptomique.

Soutenue publiquement le 29 novembre 2006 devant le jury :

Directeur de thèse	Olivier POCH, Directeur de recherche, Strasbourg
Co-Directeur de thèse	Evelyne FRIEDERICH, Directeur de recherche, Luxembourg
Rapporteur interne	Jean CAVARELLI, Professeur, Strasbourg
Rapporteur externe	Christophe AMPE, Professeur, Gant
Rapporteur externe	André STEINMETZ, Professeur, Luxembourg
Examinateur	Serge POTIER, Professeur, Strasbourg

Remerciements

Cette thèse a été une aventure magnifique. Elle m'a permis de rencontrer et de travailler avec un nombre important de personnes. Je ne sais pas comment toutes les remercier tant les moments que j'ai passé avec elles ont été formateur et si précieux à mes yeux.

« Si mon expérience est gravée dans mes mains, si mes souvenirs eux le sont dans ma tête, alors vous, vous êtes bien au chaud dans mon cœur »

Jean

Je tiens particulièrement à remercier les Professeurs Christophe Ampe, Jean Cavarelli, Serge Potier et André Steinmetz pour avoir accepter de juger mon travail de thèse.

« Ce qui a échappé aux spectateurs pourra être remarqué par les lecteurs. »

Jean Racine

Je tiens à remercier Dino Moras et Jean-Claude Thierry pour m'avoir permis de travailler au sein d'un groupe et d'un département riche par ses hommes et ses travaux.

« On fait la science avec des faits, comme on fait une maison avec des pierres : mais une accumulation de faits n'est pas plus une science qu'un tas de pierres n'est une maison. »

Henri Poincaré

A mes deux directeurs de thèse qui m'ont permis de m'amuser pendant 4 ans. Merci pour tout ce que vous avez fait pour rendre ce parcours initiatique parfois évident mais souvent compliqué, parfois long et souvent si court, mais toujours au combien passionnant.

A Olivier :

« On rencontre sa destinée souvent par des chemins qu'on prend pour l'éviter. »

Jean de la Fontaine

A Evelyne :

« Il faut toujours un coup de folie pour bâtir un destin. »

Marguerite Yourcenar

Merci à Barbara pour le temps si précieux que nous avons passés ensemble lors de ces longs mois d'écriture.

Un grand merci à la team BBS et bien entendu ses « BBS girls » ; Hélène, Corinne, Virginie et Cécile, vous m'avez tant donné. Merci à Jean-Louis Mandel dont la science est un puit sans fin.

Merci à toi le **Laboratoire de Biologie et Génomique Intégratives**, voici mes pensées :

Raymond, les heures passées depuis 5 ans devant ton écran ont été des moments parfois irréalistes mais toujours joyeux ! Merci de m'avoir ouvert à ton univers. La programmation m'a sauté à la figure et si elle ne m'a pas défiguré, c'est certainement grâce à toi.

Julie, que dis-je « Mme Julie » je peux enfin te l'avouer tes programmes ne sont pas des programmes de m...e ! Merci pour nos discussions matinales et pour tout le reste bien sûr…

Luc, que dire ? sinon la question ultime qui me taraude depuis si longtemps : « Mon petit Luc, puis-je te tutoyer ? ». Merci pour ton humour et ta joie de vivre.

Odile, on partageait déjà la passion des analyses de familles de protéines, je comprends désormais ta plus grande passion.

Fred, merci pour toutes nos discussions passionnées sur bien souvent des sujets peu scientifiques mais aussi pour tes réponses à tout mes S[R/O]S.

Annaïck, ma très chère voisine de palier scientifique. Un vrai bonheur d'avoir pu partager cet espace vital avec toi pendant si longtemps. Bonne chance à toi dans toutes tes entreprises.

Yann, mon second voisin de confinement… courage la thèse est au final, si belle. J'attends avec impatience ta soutenance ! Merci pour tous les moments débiles partagés.

Véro, pas loin d'être les 2 plus vieux jeunes du labo… et bientôt seule enfin il te reste Anne ☺

Anne, la star montante seulement en 2ème année ! Bonne route !

Laurent, merci également d'avoir partagé notre espace vital.

Manu, un mot et tu sais lequel (F__N). Merci pour ta confiance sans failles pour ICDS.

Nicolas, merci à toi d'avoir toujours participer à l'organisation de mes données et surtout à ton groupe poubelle.

Serge, combien de fois suis-je passé te voir pour des problèmes ? Merci pour ton abnégation et nos discussions techniques qui m'ont toujours plus qu'intéressées !

Wolfgang, merci à toi pour nos discussions sur les puces. Merci également pour ta gentillesse en cette fin de thèse.

Guillaume, ahlala le java et l'opéra… un grand mélomane !

Laeticia, un grand courage il te faut, pour dans le couloir travailler ;-)

Non, il n'y aura pas de jaloux ! Merci à toi le **Laboratoire de Biologie Moléculaire, d'Analyse Génique et de Modélisation**, voici ton tour :

Laurent, il est loin le temps où nous avons commencé toute cette aventure ! Merci à toi d'avoir dès le début suggérer l'importance de la bioinformatique et qui finalement m'a conduit à Luxembourg ! Merci aussi pour tous les moments joyeux ou difficiles que nous avons partagés ensemble (notamment nos longues conversations à Thionville devant l'église☺).

Arno, mon premier voisin à l'IGBMC ! Que de chemin parcouru ! Et ensuite à Luxembourg ! Merci pour ton acharnement et ta gentillesse lors de mes venues, toujours courtes mais intenses. J'espère que nos élucubrations communes auront été pour toi de bons moments ! Pour moi, cela a toujours été un plaisir de travailler avec toi.

Guillaume, le monde est petit, trop petit ! Merci à toi, la boule de nerfs, pour tes moments de folies et surtout ne lache pas… tu es sur la bonne voie ! Courage !

André, je ne dirai rien d'autre que prost. Un vrai bonheur.

Bassam, merci à toi mon petit terroriste syrien ! ;-)

François, merci pour ta bonne humeur.

Mikalai, thanks a lot and good luck in all your research.

Sandrine, merci pour ta gentillesse et courage pour ta thèse, tu verras ca en vaut la peine.

Delphine, merci pour les moments passés ensemble depuis le tout début.

Marie, merci pour ta gentillesse et ta folie.

A Fred et Aurélie simplement merci pour tout :

« On ne fait jamais attention à ce qui a été fait ; on ne voit que ce qui reste à faire. »

Marie Curie

« Un grand pouvoir implique de grandes responsabilités. »

Ben Parker

A tous mes amis, Tom, Nicole, Leïa, Fred, Caro, Juliette, Louise, Jibé, Caro, Antoine, Rachel, Anne merci pour les moments passés ensemble, ils sont si importants.

« L'amitié double les joies et réduit de moitié les peines. »

Francis Bacon

A Antoine, merci de m'avoir accueilli dans ton humble demeure à Thionville puis à Luxembourg et surtout pour tout le reste.

« Un cousin, c'est à mi-chemin entre un ami et un frère. »

Franck Oudit

A Ben, David, Johanna, Olivier, Juliette, Domi, Titi, Caro, Clara et Pierre-Louis mes premiers remparts contre l'ennui :

« Si tu diffères de moi, mon frère, loin de me léser, tu m'enrichis. »

Antoine de Saint-Exupéry

A mes « jolis » parents, Denise et Jeep :

« Le plaisir des grands est de pouvoir faire des heureux. »

Blaise Pascal

A mes grands-parents dont la logique implacable et leur amour intarissable a toujours été source de réconfort :

« Mais pourquoi former des chercheurs alors qu'il suffit de former des trouveurs ! »

Joseph Fritz

A mes parents, qui ont toujours eu l'intelligence de laisser à leur grand fils la liberté de faire les choses qu'il aime.

« Si nous faisions tout ce que nous sommes capable de faire, nous en serions abasourdis. »

Thomas Edison

A mon amour, celle qui m'a poussé, soutenu, supporté et qui me donne encore tant, ces quelques mots :

« Naît-on deux fois ? Oui. La première fois, le jour où l'on naît à la vie ; la seconde fois, le jour où l'on naît à l'amour. »

Victor Hugo

A Anna, parce que tu es sans aucun doute ma plus belle découverte.

« Les enfants sont des énigmes lumineuses. »

Daniel Pennac

Enfin, mes plus belles pensées vont à papy qui aurait été si fier…

PS : Enfin, je tiens à faire un dernier hommage à mon défunt PC (siemens Celsius Mobile !) et son clavier sans fil (si pratique !) que j'ai sauvagement cassé la veille de mon mariage lors d'une tentative non voulue de salto arrière…

Table des matières

Table des Figures

Tables des Tableaux

Avant propos

Ces travaux ont été effectués sous la co-direction de Olivier Poch responsable du « Laboratoire de BioInformatique et Génomique Intégratives » au sein de l'IGBMC en France et de Evelyne Friederich responsable du « Laboratoire de Biologie Moléculaire, d'Analyse Génique et de Modélisation » au Centre de Recherche Public-Santé à Luxembourg.

Le LBGI développe et applique des programmes dédiées à l'analyse de familles de protéines centrés autour de l'alignement multiple de séquences protéiques complètes et à l'intégration des données provenant des approches de génomique comparative et fonctionnelle. Le LBMAGM concentre ses efforts sur l'étude du cytosquelette d'actine et son implication dans les processus biologiques majeurs, tels que son rôle dans l'établissement du cancer. Sous l'impulsion du « Fonds National de la Recherche » à Luxembourg, il a implémentée en 2002 une plateforme nationale de puces à ADN.

L'importance de la bioinformatique dans les études à haut débit ainsi que la connaissance d'un système biologique dédié ont naturellement rapproché les 2 laboratoires autour d'un projet commun d'analyse du cytosquelette. Le projet commun orienté autour de la biologie à haut débit vise une étude intégrée du cytosquelette par les techniques de bioinformatiques récentes. Ces techniques sont génératrices de données à la fois nombreuses et hétérogènes. La bioinformatique apparaît ainsi incontournable pour extraire, stocker et analyser ses données. Ainsi, au cours de ma thèse, je me suis attaché à mettre en place l'ensemble des procédures et des programmes bioinformatiques indispensables au calcul, à l'acquisition, à la validation et à l'analyse des données et mesures issues des expériences entreprises, notamment l'étude des profils phylogénétiques et des profils d'expression.

Ce manuscrit se compose en trois parties. La première partie est une introduction consacrée au cytosquelette et à son importance au sein des cellules eucaryotes, à l'avènement de la bioinformatique devenue incontournable dans les études de biologie à haut débit, et notamment, aux développements récents liés à l'émergence de la génomique fonctionnelle et des méthodes de génomique comparative et des technologies des puces à ADN. La seconde partie décrit l'environnement dans lequel j'ai pu évoluer pour développer les outils bioinformatiques associés aux différentes recherches menées. Le dernier chapitre présente les résultats obtenus à la fois en génomique comparative et d'autre part en transcriptomique. Ce travail a donné lieu à 3 publications en premier auteur, ainsi que deux publications supplémentaires.

Introduction

Chapitre 1 - Le cytosquelette

« On ne connaît pas complètement une science tant qu'on n'en sait pas l'histoire. »

Auguste Comte

Le cytosquelette est un réseau dense et dynamique de polymères protéiques présent dans la totalité du cytoplasme de la cellule (Figure 1) (Machesky *et al.* 2000). Il existe 3 types de cytosquelettes, les filaments d'actine, les microtubules et les filaments intermédiaires.

Le cytosquelette fait partie des larges structures cellulaires importantes tant au niveau fonctionnel que structural (Yates *et al.* 2005). Présent, à la fois dans le cytoplasme et dans le noyau de la cellule eucaryote, il interagit avec quasiment tous les organelles (Boldogh *et al.* 2006; Egea *et al.* 2006; Shibata *et al.* 2006). Les protéines qui le composent sont en conséquence régulièrement identifiées comme des composants des protéomes de ces organelles.

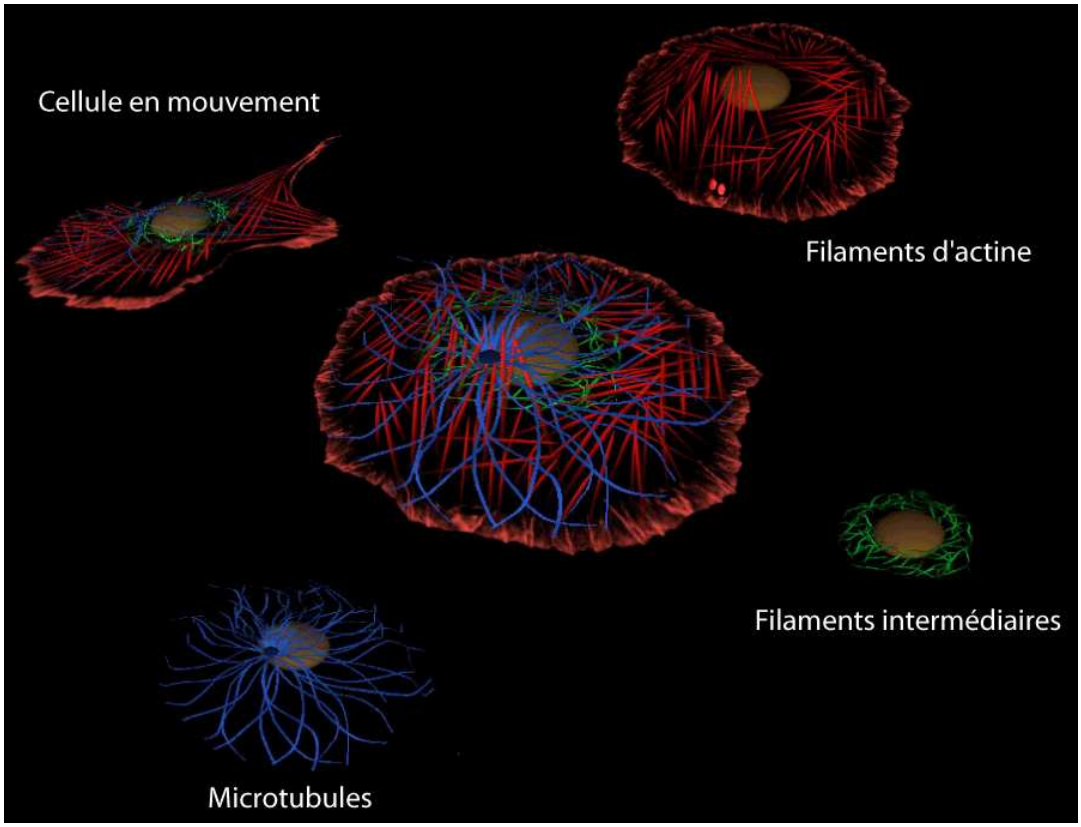


Figure 1 **Positionnement des 3 cytosquelettes dans la cellule.**
Les filaments d'actine sont indiqués en rouge, les microtubules en bleu et les filaments intermédiaires en vert (adapté de http://cellix.imolbio.oeaw.ac.at/Videotour/video_tour.html).

Ses fonctions sont nombreuses et leurs implications sont majeures pour la vie de la cellule. En effet, il représente l'échafaudage sur lequel la cellule se repose et participe largement à sa morphologie et à son développement. Le cytosquelette est impliqué dans bon nombre de processus biologiques importants pour la vie de la cellule. Ainsi, il joue un rôle prépondérant dans la division cellulaire puisque les microtubules constituent le faisceau mitotique responsable de la ségrégation des chromosomes et que les microfilaments d'actine génèrent l'anneau contractile responsable de la cytodiérèse (revue dans (Scholey *et al.* 2003)). Il joue également un rôle prépondérant dans la motilité cellulaire avec la création de structures cellulaires comme le flagelle (microtubule) et les lamellipodes (actine) (Small *et al.* 2002). Enfin, il est également impliqué dans la méiose, la communication des cellules, la régulation de la transcription et le transport intracellulaire de leurs constituants, autant de grandes fonctions cellulaires qui font de ses constituants des éléments d'étude majeurs (revue dans (Howard *et al.* 2003; Nelson 2003; Pollard 2003; Schliwa *et al.* 2003; Scholey *et al.* 2003)).

De plus, ses constituants sont souvent impliqués dans plusieurs pathologies (Janmey *et al.* 1995; Fuchs *et al.* 1998; Moir *et al.* 2001; Andrieux *et al.* 2003; Lambrechts *et al.* 2004; Gerdes *et al.* 2005) comme par exemple les cancers, les pathologies musculaires, les pathologies du système immunitaire, de la vision, de l'ouïe et de la peau. Ils constituent donc des cibles potentielles pour l'établissement de nouveaux traitements thérapeutiques (Giganti 2003).

Dans ce chapitre je commencerai par présenter des généralités sur le cytosquelette, puis je décrirai chacun des 3 cytosquelettes plus en détail, enfin dans un dernier point j'aborderai les structures cellulaires liées aux cytosquelettes.

1.1 Historique et généralités

La découverte des cytosquelettes (revue dans (Frixione 2000)) remonte à plus de 300 ans, allant de la découverte de fibres ou de câbles dans le muscle à l'analyse fine des éléments régulateurs de leur dynamique et à la caractérisation des structures 3D des éléments majeurs de chacun des 3 cytosquelettes. L'actine a été identifiée expérimentalement en 1887 par W.D. Halliburton (Halliburton 1887) à partir de cellules de muscle comme un élément capable de coaguler des préparations de myosine et fut désignée comme telle par F.B. Straub en 1942. Les microtubules également observés dès le 19ème siècle n'ont été nommés ainsi qu'en 1963 (Slautterback 1963). Enfin, les filaments intermédiaires identifiés plus tardivement (Ishikawa *et al.* 1968) justifient leur nom par un diamètre intermédiaire à celui des filaments d'actine et des microtubules.

Identifié initialement dans les eucaryotes supérieurs, le cytosquelette est un élément beaucoup plus ancien. Relativement bien caractérisé tant au niveau fonctionnel qu'au niveau du répertoire de gènes dans les organismes eucaryotes, l'hypothèse de la possible existence du cytosquelette et donc de sa probable origine dans les organismes procaryotiques a longtemps été supposée. Bien qu'aucune protéine homologue n'avait pu clairement être identifiée, quelques éléments de séquence suggéraient déjà quelques candidats (Bork *et al.* 1992; Mukherjee *et al.* 1993). C'est récemment que la preuve a été apportée par la mise en évidence des homologues structuraux et fonctionnels des éléments principaux des 3 cytosquelettes dans la bactérie (pour revue voir (Lowe *et al.* 2004) et (Shih *et al.* 2006)). Ce fut d'abord la tubuline (microtubules) qui fut identifiée dans la bactérie sous la forme de la protéine apparentée FtsZ (Nogales *et al.* 1998), puis MreB apparentée à l'actine (Jones *et al.* 2001; van den Ent *et al.* 2001) chez *Thermotoga maritima* et ParM (van den Ent *et al.* 2002) et enfin pour les filaments intermédiaires la crescentine chez les spirochètes (Ausmees *et al.* 2003). Ces 3 repliements structuraux distincts, conservés de la bactérie aux eucaryotes, sont la preuve individuelle de l'existence des cytosquelettes chez les procaryotes. Cependant, il faut noter que la séquence de ces protéines a largement évoluée. La question de savoir si ces protéines ont divergées à partir d'un ancêtre commun ou si cette homologie structurale est due à une convergence d'ancêtres différents reste l'objet d'âpres débats (Hartman *et al.* 2002).

Un autre élément intéressant est la distribution intracellulaire du cytosquelette. Il est omniprésent dans la cellule puisqu'on le retrouve à la fois dans le cytoplasme, associé à certaines structures membranaires (voir 1.3 Structures cellulaires du cytosquelette p 33) et de façon plus surprenante dans le noyau. En effet, concernant les filaments intermédiaires, les lamines sont connues depuis longtemps pour apporter un soutien à l'enveloppe nucléaire (pour revue voir (Goldman *et al.* 2002)). Bien que longtemps mise en doute, la présence de l'actine dans le noyau a bel et bien été démontrée (Pederson *et al.* 2002). La présence d'autres composants du cytosquelette d'actine dans le noyau suggère des rôles divers au sein du noyau qui peuvent être directement liés à la transcription et à sa régulation (Blessing *et al.* 2004; Pederson *et al.* 2005; Miralles *et al.* 2006). Selon le type cellulaire, chacun des cytosquelettes adoptera une organisation spatiale spécifique. Par exemple, dans les cellules épithéliales, le cytosquelette d'actine est ancré à la membrane plasmique dans les microvillosités et les jonctions intercellulaires, les microtubules assurent le transport cytoplasmique d'organelles ou de vésicules et les filaments intermédiaires sont situés autour du noyau et en connexion avec des points d'adhésion comme les desmosomes (intercellulaire et avec la lame basale).

1.2 Le 3 en 1

1.2.1 Les filaments d'Actine

Les filaments d'actine ou microfilaments sont des polymères protéiques formés de monomères d'actine. Ces filaments sont également appelés microfilaments et possèdent un diamètre compris entre 3 et 7 nm.

La forme monomérique est nommée, actine G, et les polymères sont appelés actine F. L'actine est une protéine d'environ 380 acides aminés qui possèdent une poche de fixation de l'ATP (Figure 2).

L'actine G ayant fixé l'ATP polymérise plus vite en microfilaments et se dépolymérise plus facilement sous leur forme ADP. Cette polymérisation qui consiste à l'ajout successif de monomères d'actine au filament naissant est précédée par une phase de nucléation. La phase de nucléation est initiée à partir de nucléateurs (voir ci-dessous). Ces nucléateurs vont permettre de créer des noyaux de nucléation (dimère ou trimère d'actine) à partir desquels les filaments d'actine vont pouvoir se construire. La dépolymérisation consiste en la perte de monomère à une extrémité. Les microfilaments sont des doubles hélices en apparence de polymères d'actine polarisés. Chacune des 2 hélices est orientée de la même manière proposant ainsi à chacune des extrémités une face différente de la molécule d'actine. Historiquement, la différence entre les 2 extrémités a été faite visuellement par le dessin particulier en forme de flèche des chaines de myosines attachées aux filaments d'actine. On distingue ainsi l'extrémité pointue (« *pointed end* ») ou « - » de l'extrémité barbée (« *barbed end* ») ou « + » (Figure 2) qui présentent des propriétés biochimiques distinctes en termes de vitesses d'association et de dissociation des monomères. A une concentration donnée de monomères d'actine et en présence d'ATP, *in vitro,* le microfilament s'allonge à l'extrémité barbée et se dépolymérise à l'extrémité pointue. Cette capacité de remplacement des monomères au sein d'un filament, connue sous le nom de « *treadmilling* » ou tapis roulant (Wegner 1976), est un des moteurs de la dynamique du cytosquelette puisqu'on le rencontre également pour les microtubules.

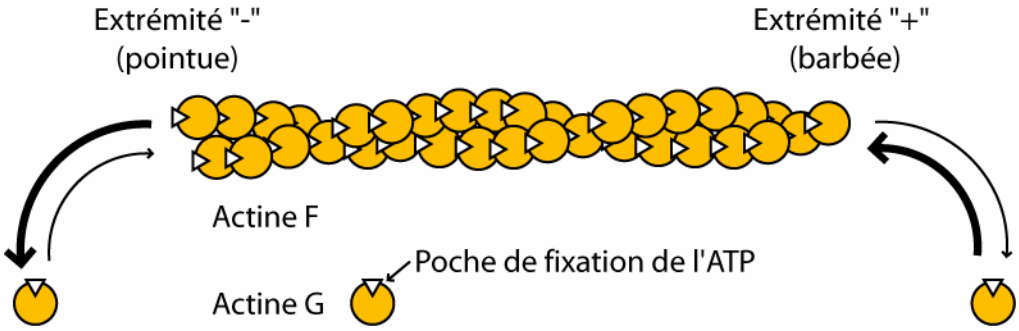


Figure 2 Vue schématique d'un filament d'actine (actine F).
Le filament d'actine est constitué d'une double hélice d'actine G empilée de la même manière et présentant une même face de la molécule d'actine à chacune des 2 extrémités.

L'extraordinaire capacité de l'actine, qu'elle soit sous sa forme G ou F, à s'associer à de nombreuses protéines constitue dans sa diversité une adaptation évolutive unique pour contrôler l'assemblage ou le désassemblage des filaments, leur organisation spatiale en faisceaux ou en réseaux ainsi que leur ancrage à des structures membranaires, en réponse à des signaux externes. Un grand nombre de ces protéines ont été identifiées depuis 1986 ((Pollard *et al.* 1986) pour revue voir (dos Remedios *et al.* 2003)). Le répertoire des protéines liant l'actine ou ABP (*Actin Binding Proteins*) est divisé en classes fonctionnelles regroupant des modules structuraux différents (Van Troys *et al.* 1999). Il convient de noter que certaines de ces protéines peuvent être assignées à plusieurs classes.

Ainsi, il existe des protéines dites de coiffe qui se fixent spécifiquement à l'une des 2 extrémités du filament. Elles ont un rôle dans la nucléation ou l'arrêt de la polymérisation (CapZ, formine, ARP2/3). Des protéines stabilisatrices (tropomyosine) ou de découpage des filaments (gelsoline, villine). Des protéines de séquestration des monomères (thymosine béta 4, profiline). Il existe également des protéines de pontage (filamine, fimbrine, fascine) qui organisent les microfilaments en faisceaux ou en réseau. On retrouve également des protéines motrices qui permettent de générer des forces ou des déplacements le long des microfilaments (myosine de type II ou non conventionnelle). Elles ont par exemple, un rôle majeur lors de la contraction des cellules des muscles ou du cortex dans les cellules non musculaires et dans le transport d'organelles et de vésicules le long des microfilaments (Mermall *et al.* 1998).

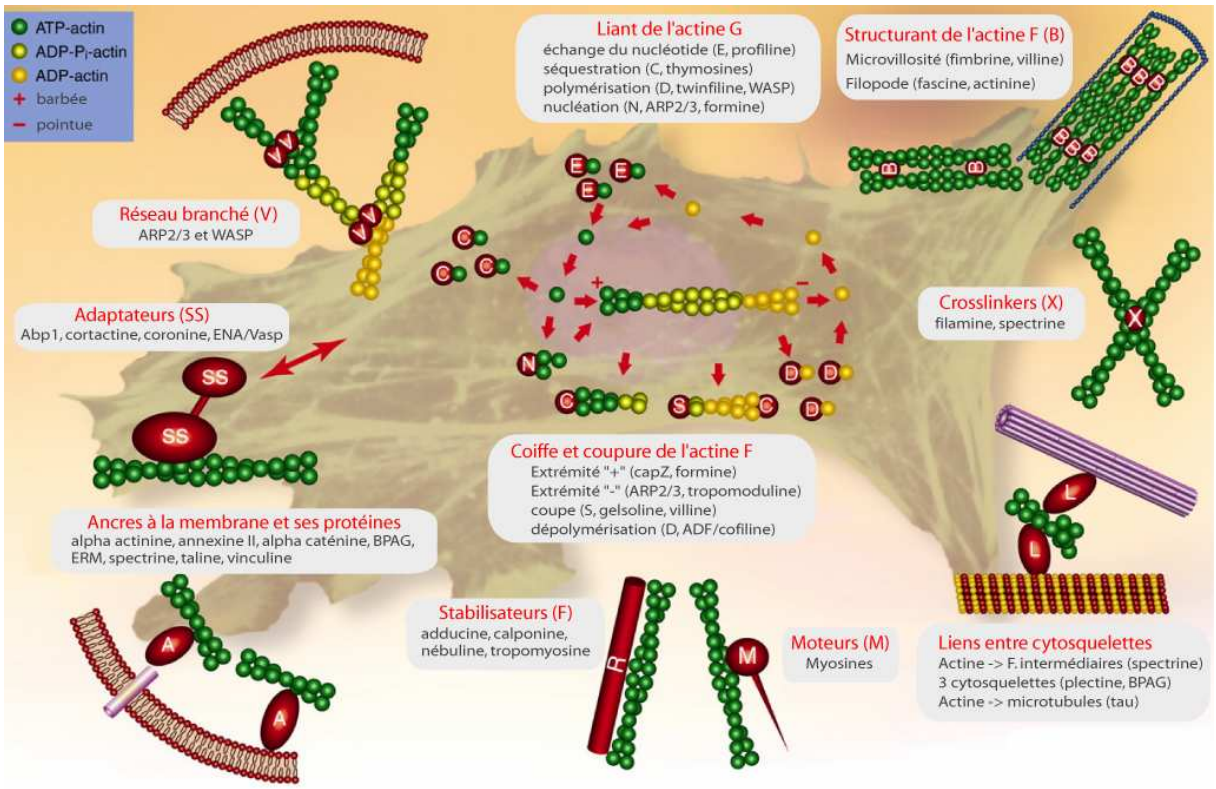


Figure 3 **Schéma général de différents types de protéines liant l'actine.**
La figure montre les différentes catégories ainsi que quelques exemples de protéines (adapté de (Winder *et al.* 2005)).

Parmi les ABP, les nucléateurs, une classe contrôlant l'organisation spatiale du cytosquelette au sein de la cellule, sont capables d'initier l'assemblage d'un nouveau filament d'actine (nucléation) (Figure 4). Il existe pour l'instant 3 types de nucléateurs différents. Le plus connu est le complexe multiprotéique ARP2/3 à 7 sous-unités distinctes. En outre, il est constitué de 2 ARPs, ARP2 et ARP3, initialement décrites comme des actines non-conventionnelles qui sont désormais nommées ARP pour « *Actin-Related Proteins* » (Machesky *et al.* 1994). Ces 2 ARPs utilisent leur proximité structurale avec l'actine pour mimer le début d'un filament permettant ainsi l'initiation d'un nouveau filament (Robinson *et al.* 2001; Nolen *et al.* 2004). La capacité des autres protéines du complexe ARP2/3 à ancrer le complexe à un filament d'actine déjà formé leur confère la capacité de créer des réseaux branchés de filaments (Figure 4).

La deuxième classe de nucléateurs comprend les formines, une famille de protéines présentes dans la plupart des organismes vivants (Faix *et al.* 2006). Les formines se fixent à l'extrémité barbée tout en y permettant l'insertion de monomères d'actine et ainsi l'élongation du filament par cette extrémité. Elles se distinguent des autres nucléateurs par leur capacité à rester fixées sur l'extrémité croissante du filament et non au point de départ.

Plus récemment, la protéine Spire a été identifiée comme le troisième nucléateur des filaments d'actine (Quinlan *et al.* 2005).

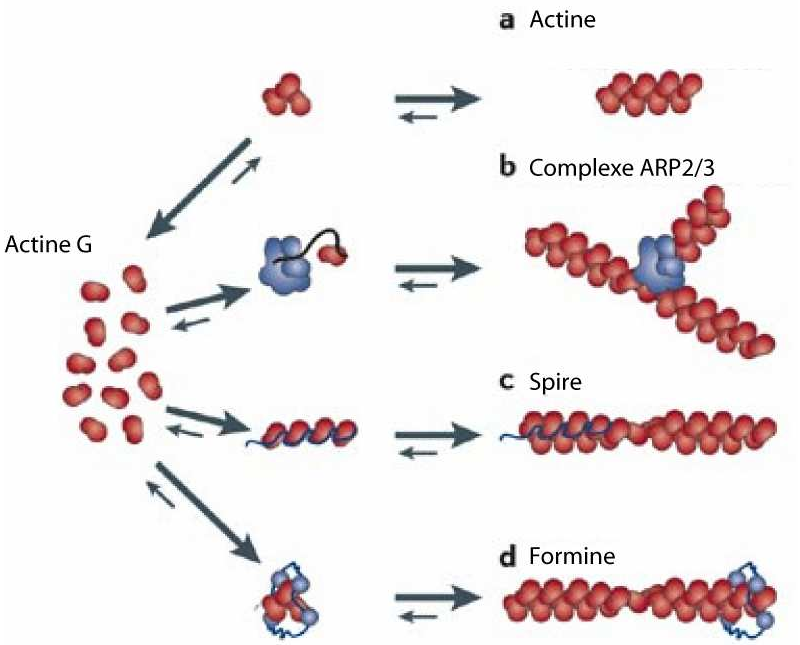


Figure 4 **Les différents nucléateurs des filaments d'actine**
La figure est adaptée de (Goley *et al.* 2006).

1.2.2 Les microtubules

Les microtubules sont constitués de tubuline, un hétérodimère composé de 2 sous-unités, α et β formées chacune de 450 acides aminés. Ces molécules de tubuline s'assemblent pour former des protofilaments en présence de GTP (Figure 5). La sous-unité β d'un dimère de tubuline est liée à la sous-unité α du dimère suivant. Ces protofilaments linéaires disposés côte à côte, constituent un microtubule. Un microtubule est généralement constitué de 13 protofilaments et peut avoir une taille de 5 à 50 μm pour un diamètre de 15 à 25 nm. Parmi les 3 cytosquelettes, les microtubules sont les seuls filaments qui possèdent une lumière centrale d'à peu près 10 nm.

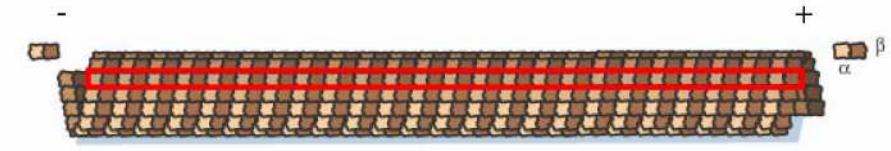


Figure 5 **Formation d'un microtubule à partir de 13 protofilaments.**
Un protofilament, indiqué en rouge, est constitué d'assemblage de doublets de tubuline alpha et béta (adapté de (Howard *et al.* 2003)).

Les microtubules sont des filaments polarisés dont les 2 extrémités, dénommées « + » et « - », sont caractérisées *in vitro* par des vitesses de polymérisation différentes. L'extrémité « + » est

déterminée par la présence de tubuline β et par une polymérisation plus rapide que l'extrémité « - » caractérisée par la présence de tubuline α (Figure 5). Tout comme l'actine, les microtubules sont capables de réaliser le remplacement de leur composants par le phénomène du « *treadmilling* » (Bergen *et al.* 1980).

Dans la cellule, on observe le phénomène appelé « instabilité dynamique des microtubules » (Mitchison *et al.* 1984). L'instabilité dynamique est un comportement particulier qui alterne des phases dites de « catastrophes » (décroissance brusque) et des phases dites de « sauvetages » (croissance).

L'organisation et la croissance des microtubules dans la cellule eucaryote animale est initiée par des MTOC (*MicroTubule Organizing Center*). Le principal MTOC est une structure unique appelée centrosome. Il a été décrit en 1888 par Boveri (décrit dans (Sathananthan *et al.* 2006)) comme un organe important dans la division cellulaire. En effet, le centrosome est à l'origine du fuseau mitotique qui permet l'alignement des chromosomes puis leur ségrégation lors de la mitose (revue dans (Scholey *et al.* 2003)). Néanmoins, le centrosome a également un rôle dans la progression du cycle cellulaire et ceci indépendamment de son rôle d'organisateur des microtubules (Rieder *et al.* 2001). D'un point de vue structural, le centrosome est constitué de 2 centrioles perpendiculaires localisés généralement près du noyau. Un centriole est une structure cylindrique de 0,4 à 0,5 μm comportant 9 triplets de microtubules (voir Figure 10 p 38). La structure est entourée du matériel péri-centriolaire qui contient tous les éléments nécessaires à son fonctionnement (Schnackenberg *et al.* 1999) comme par exemple une forme de tubuline atypique, la tubuline γ (Ou *et al.* 2004), la péricentrine ou encore la ninéine (Delgehyr *et al.* 2005). Dans les plantes, les microtubules jouent également un rôle important dans la forme de la cellule. Cependant, les cellules de plantes ne contiennent pas de centrosomes. La nucléation des microtubules se fait directememnt au niveau du cortex de la cellule (Shaw *et al.* 2003) (revue dans (Ehrhardt *et al.* 2006)).

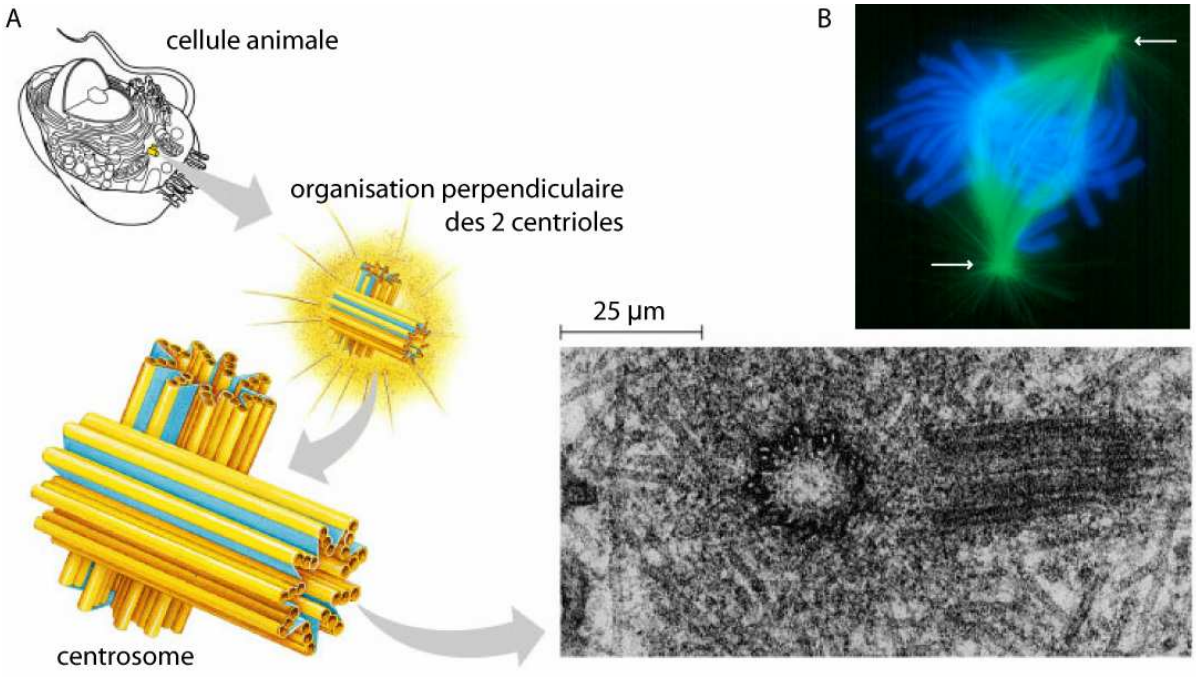


Figure 6 **Le centrosome.**
(A) Vue schématique et en microscopie électronique (à droite) du centrosome dans la cellule animale. Dans la cellule, il existe un seul centrosome composé de 2 centrioles perpendiculaires (adapté de Biology 7ème édition par Neil A. Campbell et Jane B. Reece). (B) Illustration dans une cellule de poumon en mitose par marquage fluorescent de la tubuline (vert) du fuseau mitotique de microtubules générés par les centrosomes (flèches blanches) et de l'alignement des chromosomes (DAPI, bleu) lors de l'étape de métaphase (adapté de http://www.wadsworth.org/bms/SCBlinks/web_mit2/res_mit.htm).

Tout comme le cytosquelette d'actine, les microtubules interagissent avec un certain nombre de protéines et notamment des protéines motrices (revue dans (Schliwa *et al.* 2003)). L'organisation des microtubules peut être comparée à de véritables rails orientés et permettant le transport de vésicules et d'organelles au sein de la cellule (Hirokawa 1998). Il existe ainsi 2 types de moteurs moléculaires en fonction du sens de circulation ; les kinésines permettent de manière générale un déplacement vers l'extrémité « + » des microtubules alors que les dynéines sont orientées vers l'extrémité « - ».

1.2.3 Les filaments intermédiaires

A la différence des microfilaments et des microtubules, les filaments intermédiaires sont composés de protéines différentes en fonction du type cellulaire (Tableau 1). Les protéines sont classées en 6 types et correspondent chez l'homme à près de 67 gènes (Coulombe *et al.* 2001).

Type	Nom	Tissu/Type cellulaire
I	Kératine (acide)	Epithélium
II	Kératine (neutre et basique)	Epithélium
III	Vimentine	Mésenchyme
	Desmine	Muscle
	GFAP*	Astrocyte
	Périférine	Neurone
	Syncoiline	Muscle
IV	Neurofilament (L, M, H)**	Neurone
	α-internexine	Neurone
V	Lamine	Noyau
VI	Nestine	Cellules souches

* GFAP : Glial Fibrillary Acidic Protein

** L : Low, M : Medium, H : High (poids moléculaire)

Tableau 1 Les différents types de protéines formant les filaments intermédiaires et leur distribution cellulaire.

La figure est adapté de (Coulombe *et al.* 2001).

Chaque protéine a une organisation similaire et comprend un domaine central structuré en hélice α d'une taille d'environ 310 acides aminés et 2 extrémités de taille et de structure variables.

La formation des filaments intermédiaires est le résultat d'un assemblage successif de structures de plus en plus complexes. Ainsi, les unités de base dimérisent entre elles (domaine coiled-coil), ces dimères forment à leur tour des tétramères, puis l'assemblage des tétramères créent les protofilaments… pour aboutir à un filament intermédiaire qui correspond à une fibre de 10 nm de diamètre environ (Figure 7).

Ces filaments se distinguent des microtubules et des microfilaments sur plusieurs points : leur assemblage ne requiert ni ATP, ni GTP, ils ne sont pas polarisés et ils ne peuvent effectuer de « *treadmilling* ».

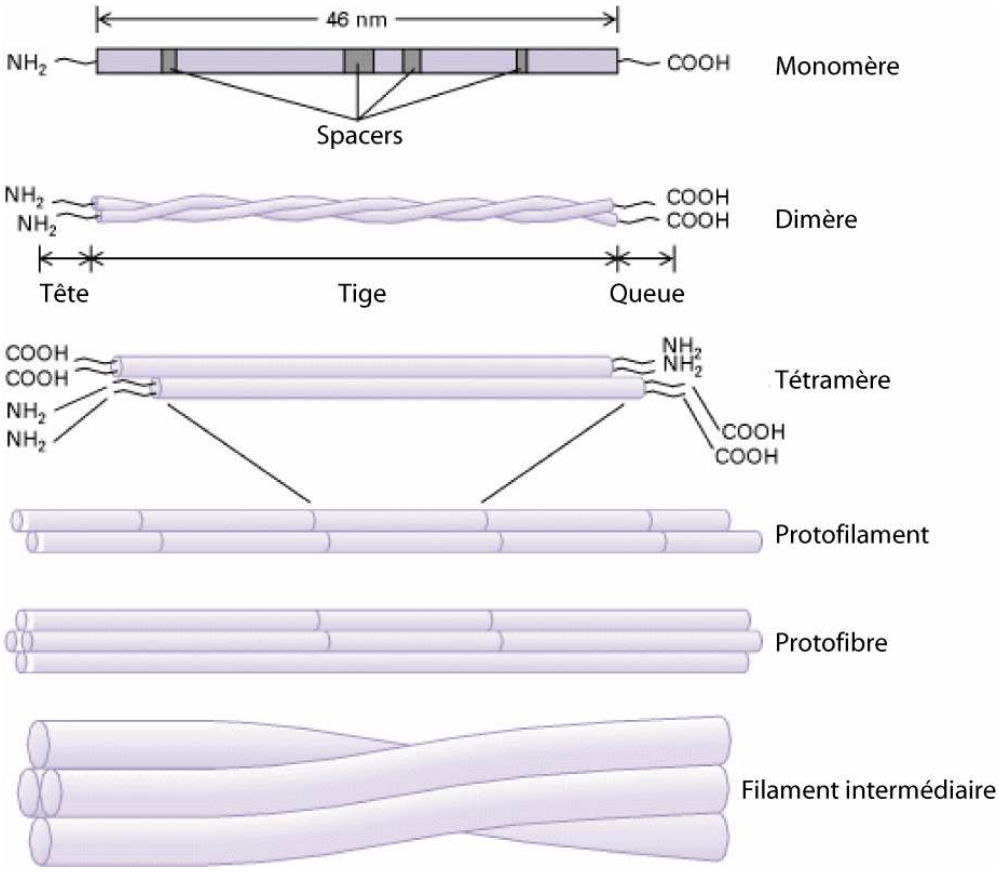


Figure 7 Formation d'un filament intermédiaire.
La formation d'un filament intermédiaire est le résultat de l'association des unités successives (adapté de http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=mcb.figgrp.5550).

Ils ont longtemps été perçus comme un réseau statique de filaments nécessaires à la résistance aux stress mécaniques (Coulombe *et al.* 2000). Bien que ces fonctions soient bien réelles, tout porte à croire que les filaments intermédiaires soient également capables d'une certaine dynamique (commenté dans (Helfand *et al.* 2004)). Par exemple, la vimentine, constituant des filaments intermédiaires dans les fibroblastes, est constamment désassemblée et réassemblée (Goldman *et al.* 1999; Martys *et al.* 1999). De plus, un certain nombre de liens avec les autres cytosquelettes implique les filaments intermédiaires dans des fonctions communes.

Les filaments intermédiaires ne possèdent pas de centres organisateurs connus. Ils sont cependant localisés autour du noyau de la cellule eucaryote et en fonction du type cellulaire en lien étroit avec les jonctions cellulaires comme les desmosomes. Ils ont un rôle majeur dans le maintien de la structure de la cellule et dans la résistance des tissus au stress mécanique (Goldman *et al.* 1996).

Ainsi l'importance des filaments intermédiaires réside également dans leur capacité à interagir avec les 2 autres cytosquelettes (Fuchs *et al.* 1998; Chang *et al.* 2004). Leur composition moléculaire différente en fonction du type cellulaire les place en position

centrale face aux 2 autres cytosquelettes si bien conservés pour engendrer des interactions tissus spécifiques. Cette interaction a des répercutions importantes sur la forme des cellules générant ainsi une force de cohésion importante entre la membrane de la cellule et son cytosquelette. Parmi les protéines de liaison avec l'actine et les microtubules, on notera la desmoplakine, la plectine (Svitkina *et al.* 1996), BPAG1 (Fuchs *et al.* 1998) et la myosine V (Rao *et al.* 2002).

1.3 Structures cellulaires du cytosquelette

En étroit contact avec la membrane plasmique, le cytosquelette sous-jacent forme des structures cellulaires ayant des fonctions diverses. Ces spécialisations de la membrane et du cytosquelette sont fréquemment associées à un type cellulaire ou à un organisme en particulier. Bien qu'ayant une organisation de base du cytosquelette similaire (p. ex. extension membranaire), ces structures peuvent assurer des fonctions très diverses grâce à la présence de protéines dont l'expression est tissu-spécifique. L'assemblage de ces structures spécialisées est contrôlé dans l'espace et dans le temps par les protéines du cytosquelette. Le répertoire des gènes impliqués dans ces structures doit être capable d'interagir avec la membrane plasmique en s'y fixant, d'organiser des faisceaux ou réseaux de filaments, générer la force nécessaire pour repousser la membrane et assurer le renouvèlement et le maintien de la structure.

Nous nous limiterons ici aux structures impliquant l'actine et les microtubules.

1.3.1 Les structures basées sur l'actine

Le cytosquelette d'actine est à l'origine de plusieurs structures membranaires basées sur des organisations différentes de ses filaments. Ainsi, il existe des extensions de la membrane qui s'appuient par exemple, sur des réseaux branchés de filaments d'actine (lamellipodes et filopodes) ou des faisceaux de filaments parallèles (soies, stéréocils et microvillosités) (Figure 8). Les filaments sont associés entre eux au moyen de différentes ABP (pour une revue voir (Revenu *et al.* 2004)).

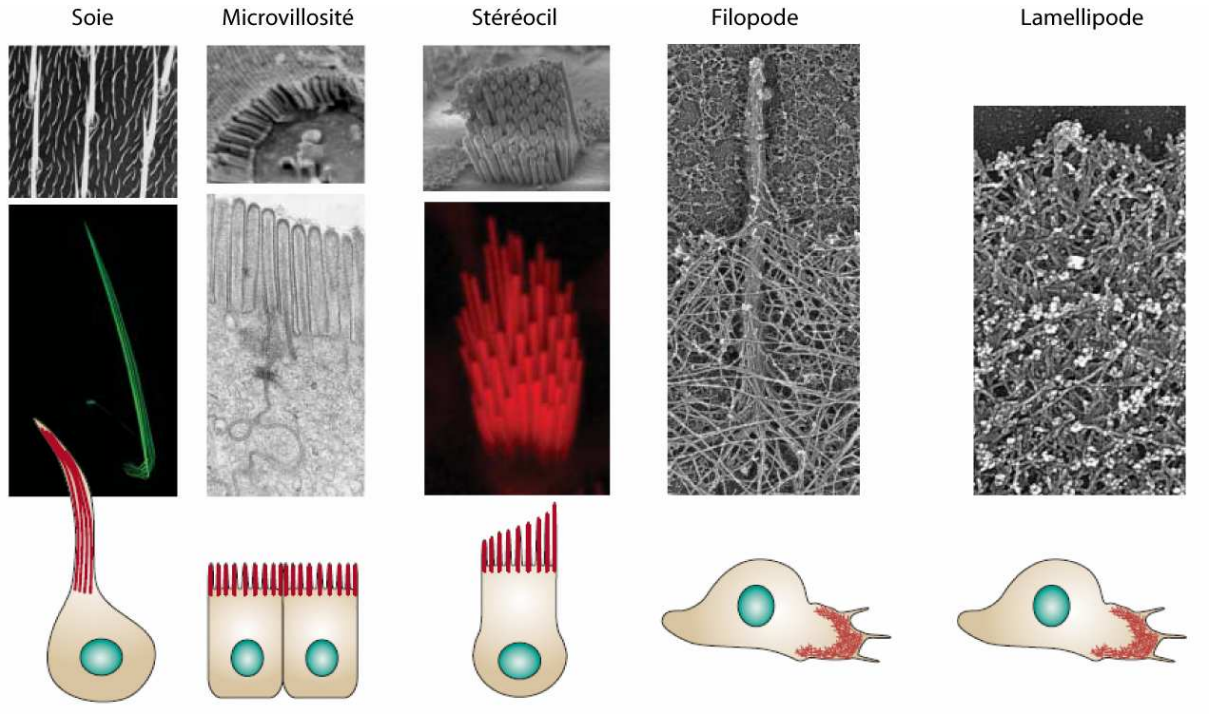


Figure 8 **Les structures membranaires établies à partir du cytosquelette d'actine.**
Pour chaque type de structure, une vue schématique (partie basse) ainsi qu'une photo prise au microscope électronique (partie haute) sont présentées (adapté de (Revenu *et al.* 2004)). Les soies, les microvillosités et les stéréocils sont constitués de faisceaux de filaments d'actine alors que les lamellipodes sont eux constitués de réseaux de filaments branchés. Entre ces 2 structures, les filopodes sont constitués de faisceaux de filaments parallèles émergeant d'un réseau de filaments branchés.

- Les soies sont disposées sur les cellules mécanoréceptrices du thorax de *Drosophila melanogaster.* Elles sont constituées de faisceaux parallèles de filaments d'actine, appelés aussi « bundle », d'environ 400 µm de long. Une soie est le résultat de l'assemblage de 11 faisceaux d'à peu près 500 à 700 filaments par des protéines dite de pontage comme la fascine et forked (Tilney *et al.* 1995).

- Les microvillosités possèdent une taille de 1 à 2 µm pour 0,1 µm de large et sont formées à la surface exposée des cellules. Ce sont des protrusions fines contenant des faisceaux parallèles de filaments d'actine (20 à 30 filaments par faisceau) organisés par exemple par la villine (Bretscher *et al.* 1979) ou la fimbrine (Bretscher *et al.* 1980). Elles sont localisées dans plusieurs types de cellules polarisées comme les cellules épithéliales intestinales, les cellules du rein, les hépatocytes, les lymphocytes et les cellules de Schwann. La fonction des microvillosités des cellules hématopoïétiques diffère de celles des cellules épithéliales. Dans les cellules épithéliales, elles sont utilisées afin de maximiser leurs surfaces d'échange (dans les cellules absorptives intestinales ou de rein, elles forment une « bordure en brosse »).

Dans les leucocytes, les microvillosités participent à l'adhésion à la paroi des vaisseaux.

- Les stéréocils sont d'une taille de 1,5 à 5,5 μm et sont également constitués de faisceaux parallèles de filaments d'actine pouvant contenir jusqu'à 900 filaments d'actines (revue dans (Tilney *et al.* 1992)). Les protéines responsables de leur assemblage sont l'espine et la fimbrine (Tilney *et al.* 1989; Zheng *et al.* 2000). Les stéréocils sont localisés dans les cellules de l'oreille interne de la plupart des vertébrés et sur le thorax de *Drosophila melanogaster.* Ce sont des mécanorécepteurs responsables par exemple, de la détection des variations sonores.
- Les filopodes sont des protrusions fines contenant des faisceaux de filaments d'actine parallèles eux-mêmes basés sur un réseau branché de filaments. Ce sont des extensions de la membrane impliquées dans la motilité et dans les interactions avec les autres cellules.
- Les lamellipodes sont des structures larges et planes enrichies en réseaux branchés de filaments d'actine. Le branchement des filaments est rendu possible par le complexe de nucléation ARP2/3 (Svitkina *et al.* 1999). Ces extensions sont situées dans la partie de la membrane leader du mouvement de la cellule. Elles sont instables et hautement dynamiques.

La formation de ces structures est directement liée à la dynamique des filaments d'actines et à leur polarité. Ainsi l'addition de monomères aux extrémités proches de la membrane permet de générer la force nécessaire pour repousser la membrane.

A l'exception des filopodes, le contrôle de la taille de l'ensemble de ces structures représente un axe de recherche particulièrement intéressant. Par exemple, la taille constante et homogène de toutes les microvillosités intestinales ou la corrélation entre la taille des stéréocils et la position des cellules qui les portent sont des illustrations frappantes. Cette caractéristique est semble-t-il directement liée à la concentration et à la nature des ABPs. En effet, il a été montré par exemple, que la surexpression de la villine, une protéine stabilisatrice des microfilaments, provoque l'élongation de microvillosités de tailles identiques à celles des cellules épithéliales intestinales sur la face dorsale de cellules fibroblastiques (Friederich *et al.* 1989). De la même manière, l'espine semble, par sa capacité à former des faisceaux parallèles d'actine, contrôler la longueur des microvillosités ou des stéréocils (Sekerkova *et al.* 2004; Sekerkova *et al.* 2006).

En dépit de leur apparente stabilité, ces structures sont à l'image du cytosquelette d'actine, très dynamiques. En effet, le *treadmilling* permet un renouvellement constant des monomères

d'actine au sein des filaments (Pollard *et al.* 2003). Ceci est notamment le cas pour les structures (filopodes et lamellipodes) impliquées dans le mouvement de la cellule. Ce renouvellement a également été décrit pour les microvillosités (Stidwill *et al.* 1984) et les stéréocils (Tilney *et al.* 1992) et plus récemment, vérifié par des expériences de FRAP (*Fluorescence Recovery After Photobleaching*) qui ont permis de mesurer le recouvrement de la fluorescence après irradiation de zones précises pour chacune des 2 structures (Schneider *et al.* 2002; Tyska *et al.* 2002).

Un certain nombre de pathologies qui sont associées à ces structures sont directement liées à des gènes du cytosquelette. Par exemple, le syndrome de Wiscott-Aldrich, une maladie génétique rare liée au chromosome X et responsable de graves déficits fonctionnels au niveau des plaquettes et des cellules immunitaires, a pour cause des mutations dans le gène WASp régulateur de la polymérisation de l'actine (Kenney *et al.* 1986). Les mutations dans les protéines de transport comme les myosines non conventionnelles sont également responsables des pertes d'audition (Gibson *et al.* 1995; Weil *et al.* 1995) et de l'acuité visuelle dans le syndrome de Usher (revue dans (Petit 2001)). Suivant le même plan de construction, ces structures de la membrane plasmique et du cytosquelette sont présentes sur de nombreux types cellulaires spécialisés et ont des rôles fondamentaux dans la vie cellulaire. Leur dysfonctionnement aboutit souvent à des répercutions fonctionnelles majeures à l'échelle de l'organisme.

1.3.2 Les structures basées sur les microtubules

Les cils sont des structures riches en microtubules qui ont une morphologie similaire à des cheveux ou à des antennes. Il existe 2 dénominations pour décrire la même structure. Ainsi, le terme générique cil est employé lorsque les cils sont observés en grand nombre à la surface des cellules et le terme flagelle lorsqu'il n'y en qu'un seul ou un petit nombre. Les flagelles correspondent également à une forme de cil dédiée à la motilité des cellules. Les cils existent dans les eucaryotes unicellulaires, dans les spermatozoïdes ou encore dans les cellules épithéliales (Figure 9). Leur fonction est généralement associée à la motilité ou à la perception de stimuli à la fois chimiques ou mécaniques.

Leur apparition est extrêmement précoce dans l'évolution des eucaryotes. Les cils sont présents dans la plupart des organismes eucaryotes (voir http://www.bowserlab.org/primarycilia/cilialist.html). Il existe néanmoins quelques exceptions comme *Cyanidioschyzon merolae, Arabidopsis thaliana, Dictyostelium discoideum* et *Saccharomyces cerevisiae* (Cavalier-Smith 2002) qui ne possèdent ni cil, ni flagelle.

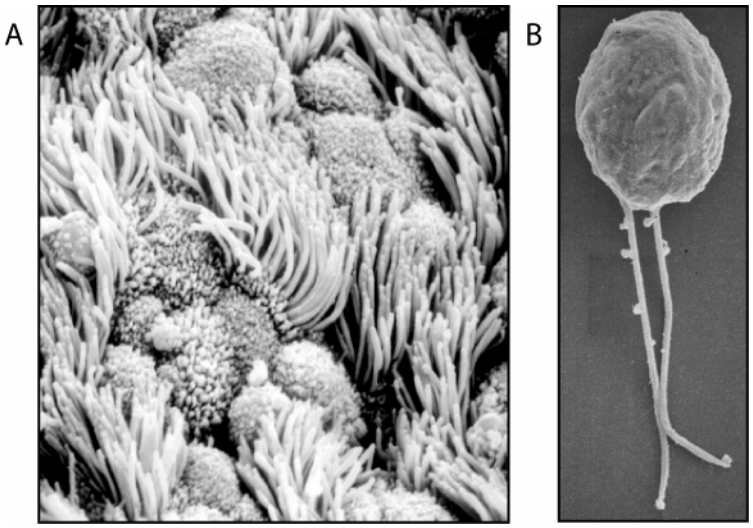


Figure 9 **Exemple de cellules ciliées en microscopie électronique.**
(A) Des cellules ciliées de la trachée d'une souris (Rosenbaum *et al.* 2002). (B) ***Chlamydomonas reinhardtii*** un organisme unicellulaire possédant 2 flagelles (Silflow *et al.* 2001).

La taille des cils est comprise entre 3 et 30 μm et jusqu'à 200 μm pour les flagelles, leur diamètre étant d'environ 0,2 à 0,3 μm. Le cil est composé de l'axonème qui contient les filaments longs de microtubules et d'un centrosome ou corpuscule basal à sa base (Figure 10). Au sein de l'axonème, les filaments sont organisés selon 2 manières ; la première dite « 9+2 » correspond à 9 doublets de microtubules encerclant une paire de microtubules centrale et la seconde « 9+0 » correspond à la même organisation sans la paire centrale (Figure 10). Chaque doublet périphérique est constitué d'un microtubule A formé de 13 protofilaments et d'un microtubule B formé de 10 ou 11 protofilaments selon les espèces. Le doublet central est constitué de 2 microtubules de 13 protofilaments chacun.

A l'image des microtubules cytoplasmiques initiés à partir des centrosomes, la formation du cil est basée sur une structure similaire, le corpuscule basal. Le corpuscule basal est constitué de 9 triplets de microtubules (A, B et C) à 13 protofilaments à partir desquels les filaments de l'axonème vont s'étendre.

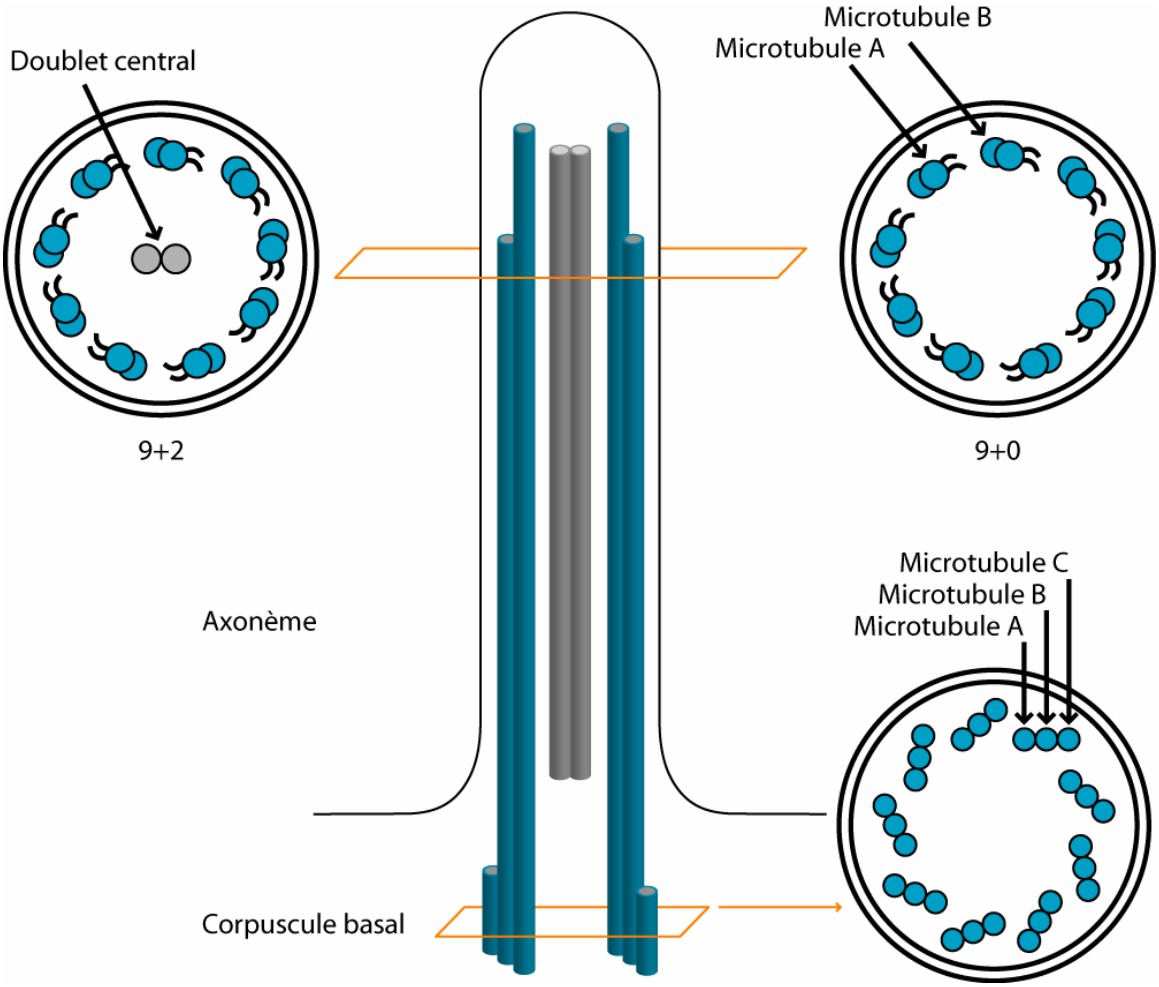


Figure 10 Représentation schématique de la structure d'un cil.
L'axonème et le corpuscule basal sont indiqués. Des coupes transversales des types « 9+2 » et « 9+0 » ainsi que du corpuscule sont représentées. Un cil est constitué de 9 doublets externes de microtubules et selon le cas d'un doublet central. Les bras de dynéine sont des protéines responsables de la force de mouvement générée entre les doublets de microtubules (inspirée de (Inglis *et al.* 2006)).

De manière générale, la forme « 9+2 » est associée à des cils motiles et la forme « 9+0 » à des cils sensibles aux stimuli. Cette séparation a priori claire est cependant sujette à certaines exceptions. Par exemple, il a été observé des cils « 9+0 » dans les gamètes des diatomés (Manton *et al.* 1970) ou encore chez la souris dans l'établissement de la polarité droite-gauche par le déclenchement d'un courant (flux nodal) orienté de la droite vers la gauche ((Nonaka *et al.* 2002) et revue dans (Tabin 2006)).

L'importance physiologique des cils a été longtemps négligée, au point de considérer cette structure comme un vestige de l'évolution (Webber *et al.* 1975). La présence de cellules ciliées dans des types cellulaires impliqués dans des fonctions majeures d'un organisme, comme les cellules rénales, les gonades, les cellules de l'oreille interne et les cellules photoréceptrices de l'œil, laisse cependant suggérer un rôle plus fondamental. (Figure 11).

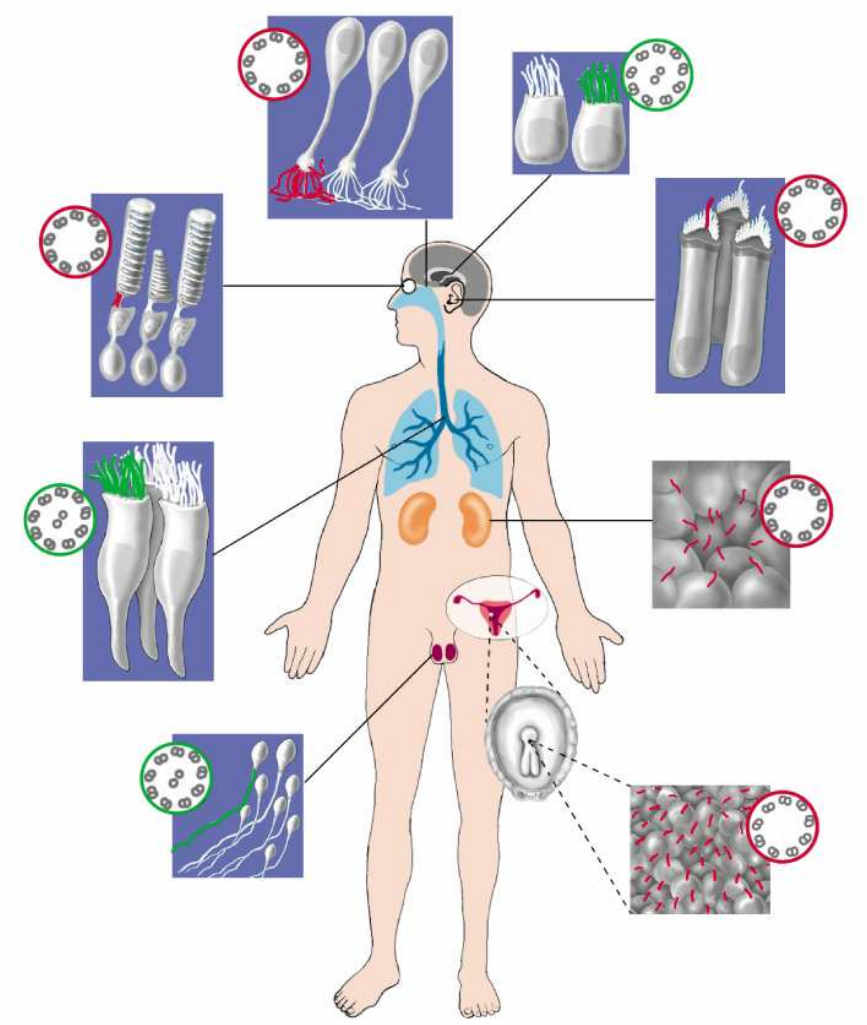


Figure 11 **Répartition des cellules ciliées chez l'homme.**

Aussi, des études récentes ont permis d'une part de montrer son importance fonctionnelle dans certains tissus (Praetorius *et al.* 2001) et d'autre part, de révéler que plusieurs de ses composants sont directement responsables de maladies ayant des conséquences dramatiques chez l'homme (revue dans (Hildebrandt *et al.* 2005; Badano *et al.* 2006)). On citera par exemple, la nephronophthisis ([OMIM 256100], référence dans la banque de maladies génétiques OMIM, *Online Mendelian Inheritance in Man* (Hamosh *et al.* 2005)), le syndrome d'Alstrom [OMIM 203800], le syndrome de Meckel-Gruber [OMIM 249000] et le syndrome de Bardet-Biedl [OMIM 209900].

L'ensemble des points que nous venons d'aborder montre l'importance des cytosquelettes dans la cellule. Ce système, dont les composants majeurs sont identifiés, est régulé de manière fine dans le temps et dans l'espace. Dans ce contexte, l'analyse globale des composants des 3 cytosquelettes par différentes approches, comme les profils phylogénétiques ou les profils d'expression, est essentielle afin de mieux comprendre les relations qui existent entre ses différents éléments.

Chapitre 2 - La bioinformatique

« Je n'ai pas peur des ordinateurs. J'ai peur qu'ils viennent à nous manquer. »

Isaac Asimov

2.1 Définition et historique

La bioinformatique est la discipline scientifique récente qui englobe la collecte, l'organisation et l'analyse de données biologiques au moyen de l'outil informatique. Le terme bioinformatique, aussi appelé en anglais « computational biology » ou « *in silico* biology » regroupe sous le même mot, 2 approches. La première approche correspond à la sensibilité plus informatique du domaine et consiste en l'élaboration d'algorithmes et le développement de programmes pour extraire l'information biologique ainsi que la création de banques de données. La seconde approche correspond à la bio-analyse dont le but est centré sur l'analyse de ces données et leur signification dans un contexte biologique. Ces 2 approches, bio-analytique et informatique, sont complémentaires et sont naturellement liées l'une à l'autre dans leur évolution (Tableau 2). En effet, l'étude d'un système biologique complet n'est rendue possible que par l'existence d'outils adaptés et performants et la bonne compréhension des systèmes biologiques permet de mettre au point des programmes adaptés et efficaces.

La bioinformatique est une discipline jeune qui date de la seconde moitié du vingtième siècle (Tableau 2). Elle doit son essor à l'émergence et à la diffusion de l'informatique qui a permis de disposer d'ordinateurs puissants capables de stocker et de traiter un nombre croissant de données. La bioinformatique doit également son développement à la compréhension des objets biologiques majeurs (ADN, ARN et protéine) et l'avènement d'outils de prédictions (Tableau 2). L'évolution conjointe de ses 2 approches de la bioinformatique en fait actuellement une discipline indispensable à la compréhension du vivant.

Date	Auteur	Evènement
1953	Watson	Modèle en double hélice de l'ADN (Watson *et al.* 1953)
	IBM 650	Premier ordinateur commercial
	Sanger	Détermination de la séquence de la chaîne A et B de l'insuline (Sanger *et al.* 1953; Sanger *et al.* 1953)
1956	Anfinsen	La structure tridimensionnelle d'une protéine est fonction de sa séquence (Anfinsen *et al.* 1956)
	IBM	Commercialisation des premiers disques durs
1958	Crick	Enonciation du dogme central de la biologie moléculaire (Crick 1958)
1961-1965	Nirenberg et Matthaei	Déchiffrage du code génétique (Matthaei *et al.* 1962)
1965	Dayhoff	1er atlas de séquences et structures de protéines (Dayhoff 1965)
	Moore	Loi de Moore sur l'évolution de la puissance de calcul
1967	Fitch	Construction d'arbres phylogénétiques (Fitch *et al.* 1967)
1969	ARPANET	Premières interconnexions universitaires
1970	Needleman et Wunsch	Algorithme d'alignement optimal global entre deux séquences de protéines (Needleman *et al.* 1970)
1974	Cerf et Kahn	Développement du concept d'Internet et du protocole TCP
	Chou	Algorithme de prédiction de structures secondaires de protéines (Chou *et al.* 1974)
1977	Sanger	Méthode de séquençage de séquence nucléique (Sanger *et al.* 1977)
1978	Sanger	Séquençage du bactériophage phiX174 (5386 pb) (Sanger *et al.* 1978)
1980	EMBL	Création de la banque européenne de séquences nucléiques
1981	Smith et Waterman	Algorithme d'alignement optimal local de séquences (Smith *et al.* 1981)
	IBM	Premier ordinateur sous le nom de PC (Personal Computer)
1982	GenBank	Création de la banque américaine de séquences nucléiques
1983	Mullis	Invention de la réaction de polymérisation en chaîne (PCR)
1984	Gouy	ACNUC : logiciel d'interrogation de banques de séquences (Gouy *et al.* 1984)
1985	Lipman et Pearson	FASTA (Lipman *et al.* 1985)
	Sony et Philips	Création d'un nouveau support numérique le CD (Compact Disc)
1986	Swiss-Prot	Création de la banque de séquences protéiques
	DDBJ	Création de la banque japonaise de séquences nucléiques
	T. Roderick	Apparition du terme « genomic »
1987	Applied Biosystems	Commercialisation du premier séquenceur automatisé
	Burke	Création du vecteur de clonage YAC (Burke *et al.* 1987)
	McKusick	1ère carte génétique du génome humain (McKusick *et al.* 1987)
	Puces à ADN	Apparition de la technologie
1988	HUGO	Coordonne le décryptage mondial du génome humain
	Higgins	CLUSTAL : programme d'alignement multiple de séquences (Higgins *et al.* 1988)
	Peterson	Utilisation de la Taq polymérase pour la PCR (Peterson 1988)
1989	Fields	Système double-hybride permettant de détecter les interactions cellulaires

		entre deux protéines
	Internet	Internet succède à ARPANET
1990	Altschul	BLAST (Altschul *et al.* 1990)
	Berners-Lee	Publication du premier document HTML
	Collins	Mise au point du clonage positionnel (Collins 1990)
	HGP	« Human Genome Project » visant à décrypter l'intégralité du génome humain
1991	Adams	1er séquençage à grande échelle d'ADNc (EST) (Adams *et al.* 1991)
	Roberts	GRAIL : programme de localisation de gènes
1992		Séquence du chromosome III de *Saccharomyces cerevisiae*
1993	Cohen	Première carte physique du génome humain (Cohen *et al.* 1993)
	Boguski	dbEST : banque de données internationale d'EST (Boguski *et al.* 1993)
	Etzold	SRS : logiciel d'interrogation de banques (Etzold *et al.* 1993)
1995	Fleischmann	Séquençage du premier organisme vivant, la bactérie *Haemophilus influenzae* (Fleischmann *et al.* 1995)
	DVD Forum	Création d'un nouveau support numérique de stockage haute capacité, le DVD (Digital Video Disc ou Digital Versatile Disc)
1996	Walsh	Séquençage du premier génome eucaryote *Saccharomyces cerevisiae* (Walsh *et al.* 1996)
	Affymetrix	Commercialisation de la première puce à ADN
1997	Blattner	Séquençage de *Escherichia coli* (Blattner *et al.* 1997)
1998		Séquençage du 1er organisme pluricellulaire, *Caenorhabditis elegans* (1998)
	W3C	Création du format XML
1999	Dunham	Séquence du chromosome 22 humain (Dunham *et al.* 1999)
2000	Stover	Séquence du génome de *Pseudomonas aeruginosa* (Stover *et al.* 2000)
	Dennis	Séquence du génome d'*Arabidopsis thaliana* (Dennis *et al.* 2000)
	Adams	Séquence du génome de *Drosophila melanogaster* (Adams *et al.* 2000)
	Ashburner	Création de la banque d'annotation Gene Ontology (Ashburner *et al.* 2000)
2001	Lander	Séquence préliminaire du génome humain par HGP (Venter *et al.* 2001)
	Venter	Séquence préliminaire du génome humain par Celera Genomics (Venter *et al.* 2001)
	AMD et Intel	1ers processeurs cadencés à 1GHz pour PC
	Ensembl	Genome browser Ensembl (www.ensembl.org)
	NCBI	Genome browser du NCBI (www.ncbi.nlm.nih.gov/mapview)
2002	Waterston	Séquence préliminaire du génome de la souris (Waterston *et al.* 2002)
	Kent	BLAT pour la recherche de séquence génomique (Kent 2002)
	UCSC	Genome browser à l'UCSC (Université de Santa Cruz) (Kent *et al.* 2002)
2003	Galagan	Séquence du génome complet *Neurospora crassa* (Galagan *et al.* 2003)
	Stein	Séquence du 2ème nématode *C. briggsae* (Stein *et al.* 2003)
2004	Gibbs	Séquence du génome complet du rat (Gibbs *et al.* 2004)
	Jaillon	Séquence du génome complet de *Tetraodon nigroviridis* (Jaillon *et al.* 2004)
	Dujon	Séquence du génome complet de 4 nouveaux champignons (Dujon *et al.* 2004)
	Hillier	Séquence du génome complet de *Gallus gallus* (Hillier *et al.* 2004)
	Matsuzaki	Séquence du génome complet de *Cyanidioschyzon merolae* une algue rouge (Matsuzaki *et al.* 2004)

2005	Berriman et El-Sayed	Séquence du génome complet de 2 trypanosomes (Berriman *et al.* 2005; El-Sayed *et al.* 2005)
	Eichinger	Séquence du génome complet de *Dictyostelium discoideum* (Eichinger *et al.* 2005)
	Abrahamsen et Xu	Séquence de génomes complets de 2 Cryptosporidies (Abrahamsen *et al.* 2004; Xu *et al.* 2004)
	Ivens	Séquence complète de *Leishmania major* (Ivens *et al.* 2005)
	Loftus	Séquence complète de *Entamoeba histolytica* (Loftus *et al.* 2005)
	Do	ProbCons : programme d'alignement multiple (Do *et al.* 2005)
	AMD et Intel	1ers processeurs double-cœur pour PC.
2006-		Séquençage de plusieurs primates, du cochon, de la vache, du cheval, du kangourou, de l'éléphant, du mouton, du chien, du chat, du lapin, de la grenouille, du poisson zèbre, etc. (soit plus de 600 génomes eucaryotes complets en cours de séquençage)

Tableau 2 Chronologie non exhaustive des évènements marquants en biologie, informatique et bioinformatique.
Les événements de la biologie, de l'informatique et de la bioinformatique sont représentés respectivement en noir, jaune et vert (traduit et adapté de http://bio.cc/Bioinformatics/history_of_bioinformatics.html, http://villemin.gerard.free.fr/Wwwgvmm/Histoire/Informat.htm).

Depuis l'énoncé du dogme central en biologie moléculaire en 1958 par Francis Crick (Crick 1958; Crick 1970) marquant les fondements de la biologie moléculaire et la définition des objets majeurs en biologie jusqu'à la publication des séquences complètes du génome de la levure *Saccharomyces cerevisiae* et de l'homme à la fin du XXème siècle de grands changements ont été opérés en biologie.

Figure 12 Dogme central de la biologie moléculaire énoncé par Francis Crick en 1958.
Les flèches de couleurs représentent les voies « normales » de l'information, les flèches noires représentent d'autres voies possibles.

En à peine 50 ans, la recherche en biologie est passée de l'étude d'un seul gène à l'étude de génomes et de protéomes d'organismes complets. Ce bouleversement a nécessité des moyens de calculs de plus en plus importants, le rassemblement des banques de données existantes comme GenBank/EMBL/DDBJ, le développement de programmes performants (BLAST, BLAT, ProbCons) et le développement de nouvelles techniques (les puces à ADN). Cette nouvelle échelle a bouleversé les méthodes d'analyse et généré de plus en plus de données. L'analyse de cette grande quantité de données n'est rendue possible que par l'utilisation de

la bioinformatique. En effet, la bioinformatique est un élément essentiel pour le passage de la biologie à l'ère post génomique et aux études à haut débit.

2.2 Les banques de données de séquences : la pierre angulaire

Le besoin d'accéder rapidement et facilement à l'information a toujours été une demande de la part des biologistes et un moteur de développement de la bioinformatique. Ce besoin s'est fait encore plus pressant depuis l'utilisation de technologies à haut débit et notamment du séquençage à haut débit de génomes ou de séquences exprimées par les organismes. Cet accès rapide, et donc forcément organisé, se traduit par la création de banque de données. Les premières banques de données de séquences en biologie moléculaire sont apparues d'abord sous la forme de livres. Rapidement, le choix du support informatique s'est imposé de lui même pour pouvoir réaliser des recherches et des analyses plus rapides et particulièrement via des programmes informatiques dédiés. Au début des années 70, il existait uniquement des banques de séquences protéiques et de structures tridimensionnelles. Il faudra attendre la fin des années 70 pour voir apparaître les premières banques nucléiques. Déjà s'esquisse la séparation entre les banques généralistes, dans lesquelles sont stockées les séquences correspondantes aux objets de bases en biologie moléculaire (ADN, ARN et protéine) provenant de tous les organismes, et des banques à valeur ajoutée qui concentrent leurs efforts sur des points particuliers comme un organisme en particulier ou une thématique donnée, en y apportant des informations supplémentaires.

Actuellement, les maîtres mots des banques de données sont « l'unification », la standardisation et l'interconnexion. En effet, de gros efforts sont réalisés pour standardiser puis unifier les banques entre elles. L'exemple le plus récent est l'unification de toutes les banques protéiques majeures sous une seule appellation, UniProt (*United Protein Databases*) que nous présenterons dans le chapitre suivant. Bien entendu, ces unifications impliquent des efforts colossaux au niveau international entre tous les organismes impliqués dans le stockage et la diffusion de ces données. Néanmoins, la fusion des banques n'est pas toujours applicable, ce sont alors les efforts d'interconnexion entre les banques qui sont importants avec des liens dits croisés entre 2 banques qui traitent du même objet biologique, mais contiennent des informations complémentaires.

Dans ce chapitre, nous aborderons, dans un premier temps, les banques de données dites « généralistes » et dans un deuxième temps, les banques de données à valeur ajoutée.

2.2.1 Les banques dites « généralistes »

Cette section présente les principales collections généralistes de séquences nucléotidiques et protéiques ainsi que les centres de saisie qui leur sont associés.

2.2.1.1 Les banques de nucléotides

Il existe actuellement trois banques généralistes de séquences nucléotidiques publiquement accessibles de par le monde : la banque de l'EMBL ou EMBL-Bank (Hamm *et al.* 1986; Baker *et al.* 2000 Kanz, 2005 #282) en Europe, GenBank (Benson *et al.* 2000; Benson *et al.* 2006) aux Etats-Unis et la DDBJ (Tateno *et al.* 2000; Okubo *et al.* 2006) au Japon. EMBL-Bank a été créée en 1980 à Heidelberg (Allemagne) et est maintenue depuis 1994 par l'EBI (*European Bioinformatic Institute*), à Cambridge (Angleterre). GenBank a été mise en place en 1979 au LANL (*Los Alamos National Laboratory*), à Los Alamos (Etats-Unis). Depuis 1992, elle est maintenue au NCBI (*National Center for Biotechnology Information*), à Bethesda (Etats-Unis). Enfin, la DDBJ a débuté son activité en 1984. Cette banque a été créée, et est toujours maintenue, au NIG (*National Institute of Genetics*) à Mishima (Japon). Ces trois centres ont également en charge la saisie et la distribution des données.

Un effort collaboratif important a toujours été de mise entre ces 3 banques qui constituent les plus grands centres de dépôt et de consultation de séquences nucléotidiques du monde. En effet, depuis la formalisation de leur entente en février 1987 sous la forme de l'*International Nucleotide Sequence Database* (http://www.insdc.org), un nombre impressionnant de séquences a été soumis (Figure 13). En août 2005, un nouveau cap symbolique fut franchit puisque l'information stockée au sein des 3 centres atteignait 100 Gigabases (100 000 000 000 bases) (http://www.nlm.nih.gov/news/press_releases/dna_rna_100_gig.html).

Les données proviennent en quasi-totalité de soumissions directes effectuées par les auteurs, ceci par l'intermédiaire du réseau Internet. En effet, la plupart des revues de biologie moléculaire n'acceptent de publier des articles se référant à des séquences que si celles-ci sont dotées d'un numéro d'accession fourni par les banques. Il convient donc, dès l'obtention d'une nouvelle séquence, de soumettre celle-ci à l'un des trois centres de saisie. Les données restantes sont extraites de la littérature scientifique (essentiellement à partir de documents tels que des livres ou des thèses). Il existe également des procédures de soumission automatique pour des séquences provenant de brevets.

La collaboration accrue entre les 3 banques se traduit par des échanges de séquences quotidiens assurant une synchronisation quasi parfaite entre les 3 centres de dépôts. La conséquence de ces échanges est qu'en pratique, ces trois banques n'en font qu'une. Ceci

amène régulièrement les responsables des centres de saisie à se poser la question de l'utilité de la maintenance de trois banques différentes. Il existe ainsi depuis longtemps un projet de fusion d'EMBL, de GenBank et de la DDBJ en un seul système.

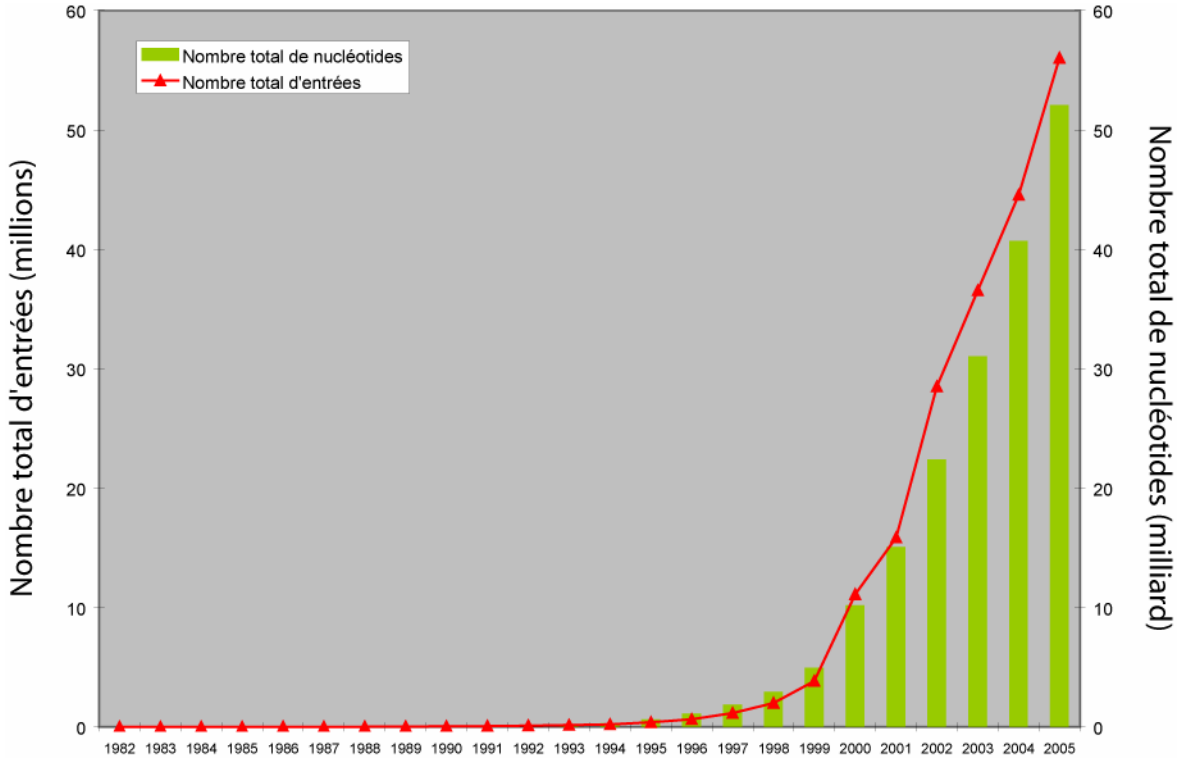


Figure 13 Evolution du nombre d'entrées et du nombre de nucléotides stockés par la banque GenBank depuis 1982.
Les données sont extraites du NCBI (http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html).

EMBL, GenBank et la DDBJ sont distribuées par les centres sous la forme d'un ensemble de fichiers plats regroupant les séquences en fonction de critères taxonomiques (procaryotes, Virus, Primates, …) ou de leur origine (EST, STS et brevets). A l'intérieur de ces fichiers, chaque séquence est contenue dans une structure appelée « entrée », une entrée comprenant une quantité variable d'informations liée à la séquence considérée. Les informations en question sont introduites au niveau de « champs » bien définis.

Le format de stockage utilisé peut être différent d'une banque à l'autre. Citons l'exemple des banques EMBL (Annexe 1) différent de celui de GenBank et de la DDBJ (Annexe 2). Cependant, cette différence ne porte que sur la façon de représenter les données. Le format EMBL facilite, par la simplicité de sa structure, le développement de programmes accédant de façon automatique aux données des banques (voir 2.3.1 La recherche textuelle). En effet, dans ce format, les champs sont identifiés à l'aide d'un code à deux lettres localisé dans les deux premières colonnes du fichier (par exemple « ID » correspond au nom de l'entrée).

De nouvelles versions de ces banques sont proposées avec une périodicité de deux mois pour GenBank et de trois mois pour EMBL-Bank. Par ailleurs, les trois centres de saisie procèdent à des mises à jour quotidiennes de leurs banques respectives. Il est important de noter que chaque séquence soumise est de fait archivée par ces banques de séquences et ne pourra être modifiée que par les auteurs de la dite séquence.

2.2.1.2 Les banques de protéines

La première banque informatique de séquence de protéines date de la fin des années 70 avec la création de la NBRF/PIR (Protein Identification Resource) grâce au *Georgetown University Medical Center* et de la *National Biomedical Research Foundation,* résultant de la version imprimée de l'Atlas of Protein Sequence and Structure de Margaret Dayhoff (USA). En 1986, une version reformatée et corrigée de la NBRF/PIR nommée Swiss-Prot voit le jour à l'EMBL (Amos Bairoch). Depuis plus de 30 ans, l'importance de ces banques de séquences n'a fait que croître et elles constituent désormais la pierre angulaire de la compréhension des fonctions des protéines dans la vie cellulaire.

A l'image de l'association des 3 grandes banques de séquences nucléiques, le *National Institute of Health* (NIH) annonçait, en octobre 2002, le financement d'une base de connaissance unique et universelle sur les protéines. Totalisant 15 millions de dollars pour un programme d'une durée de trois ans, qui a d'ors et déjà été renouvelée, cette subvention devait permettre la création d'une base de données unique sur les protéines en associant les deux groupes européens le *Swiss Institute of Bioinformatic* (SIB) et l'*European Bioinformatics Institute* (EBI) et un groupe américain, la *National Biomedical Research Foundation*. Cette association récente comparée à celle effectuée par les banques GenBank, EMBL et DDBJ va cependant plus loin car une véritable banque unique existe et est distribuée sous le nom de UniProt. UniProt (*United Protein Databases*) est la réunification des trois banques de séquences protéiques majeures (Swiss-Prot, TrEMBL et PIR) en une seule et unique ressource (Wu *et al.* 2006). Dans les points qui suivent je décrirai les différents composants d'UniProt.

2.2.1.3 Swiss-Prot

La banque Swiss-Prot, créée en juillet 1986 par Amos Bairoch à Genève, est maintenue et distribuée conjointement par le SIB et l'EBI. Cette banque de données de séquences protéiques se veut un atlas des protéines et constitue la référence des banques de protéines annotées. Le format de données adopté par Swiss-Prot suit de très près celui en vigueur à l'EMBL (Annexe 3).

Chaque entrée de Swiss-Prot n'est pas annotée par les auteurs de la séquence, mais par les annotateurs de Swiss-Prot eux-mêmes, ceci permettant de donner une cohérence beaucoup plus grande à son contenu. Les séquences présentes dans cette banque proviennent de quatre origines : la traduction des gènes annotés dans EMBL, certaines protéines issues d'autres banques protéiques, la consultation de publications scientifiques et quelques soumissions directes par les auteurs.

Une des caractéristiques principales de Swiss-Prot est la qualité extrêmement élevée de ses annotations. Les différentes catégories d'informations figurant dans la banque comprennent notamment la ou les fonctions des protéines, les modifications post-traductionnelles connues, les mutations, les sites et domaines structuraux ou fonctionnels identifiés, les références bibliographiques, les structures secondaires et quaternaires, les similarités avec d'autres protéines, les positions conflictuelles pour chaque entrée, etc. Toutes ces annotations proviennent à la fois d'une consultation régulière de la bibliographie et de l'apport d'informations par des « experts » sur certaines familles de protéines.

Une autre qualité de cette banque est sa redondance minimale, en effet les différentes versions d'une même entrée sont fusionnées et ne portent qu'un seul identifiant. Un autre atout de cette banque est également l'introduction d'un nombre très important de références croisées avec d'autres banques de données (>60) comme par exemple avec les banques nucléiques (GenBank/EMBL/DDBJ) ou vers la banque de structure 3D (PDB). Cette spécificité se révèle particulièrement utile avec le développement des systèmes d'interrogation de banques (voir 2.3.1 La recherche textuelle).

La banque Swiss-Prot constitue ainsi une véritable exception dans le monde des banques de séquences généralistes dans le sens où elle privilégie la qualité et la richesse des annotations à l'exhaustivité de sa collection de séquences. Ainsi, dans l'état actuel des connaissances, elle ne comprend pas à elle seule toutes les protéines d'un organisme. Elle ne contient, par exemple, que 14573 séquences de protéines humaines pour environ 24000 à 30000 gènes potentiels.

Cependant, l'arrivage continuel de nouvelles séquences issues de projets de séquençage de génomes pénalise la banque Swiss-Prot dont la richesse des annotations ne peut suivre une telle explosion de données. Ces éléments se reflètent parfaitement dans la croissance quasi linéaire du nombre d'entrée dans cette banque (Figure 14).

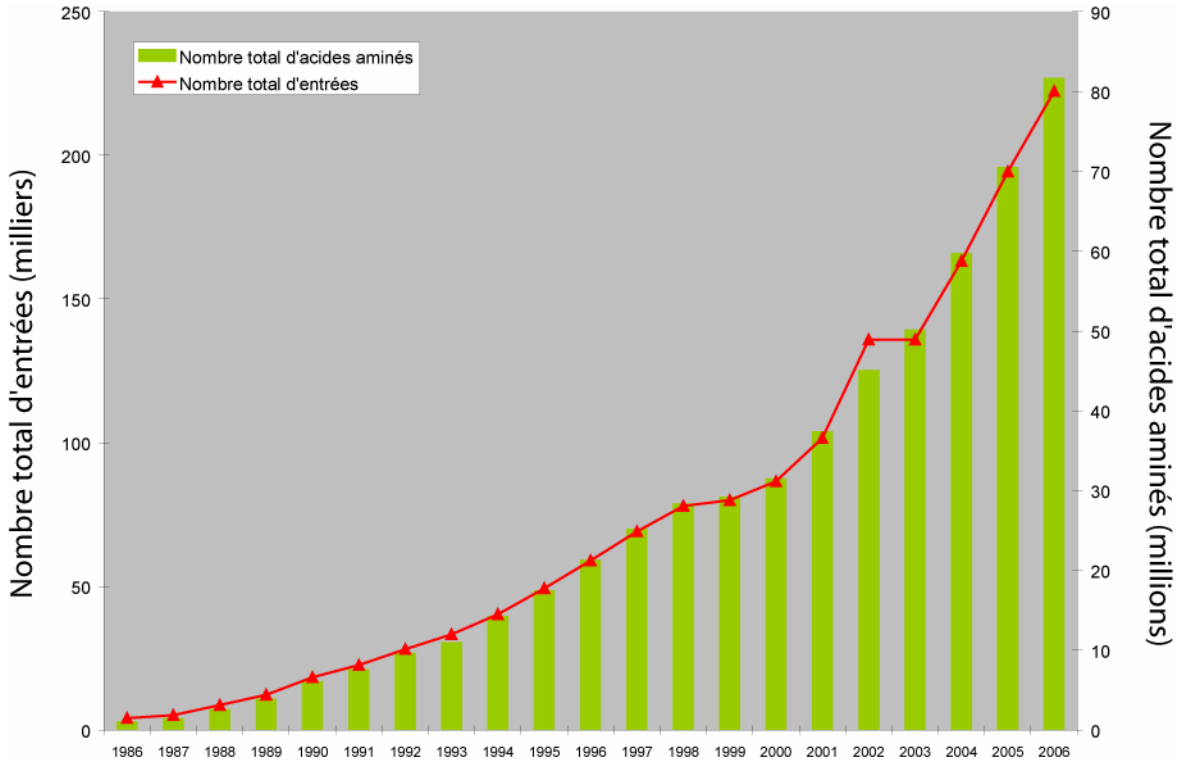


Figure 14 Evolution du nombre d'entrée de la banque Swiss-Prot depuis sa création en 1986.
Les données sont extraites du site Expasy (http://www.expasy.org/sprot/relnotes/#SPstat).

2.2.1.4 TrEMBL

Introduite en 1996, la banque TrEMBL pour *Translation from EMBL* est décrite comme un supplément de Swiss-Prot. En effet, TrEMBL, distribuée par l'EBI, contient la traduction de toutes les parties codantes annotées figurant dans la banque EMBL, à l'exception des protéines figurant dans Swiss-Prot. TrEMBL constitue ainsi un dépôt de séquences non validées et faiblement annotées, qui vont par la suite être examinées par les annotateurs de Swiss-Prot pour y être intégrées. Sa croissance est fulgurante et correspond au flux important de données générés par les projets de séquençage des génomes complets (Figure 15).

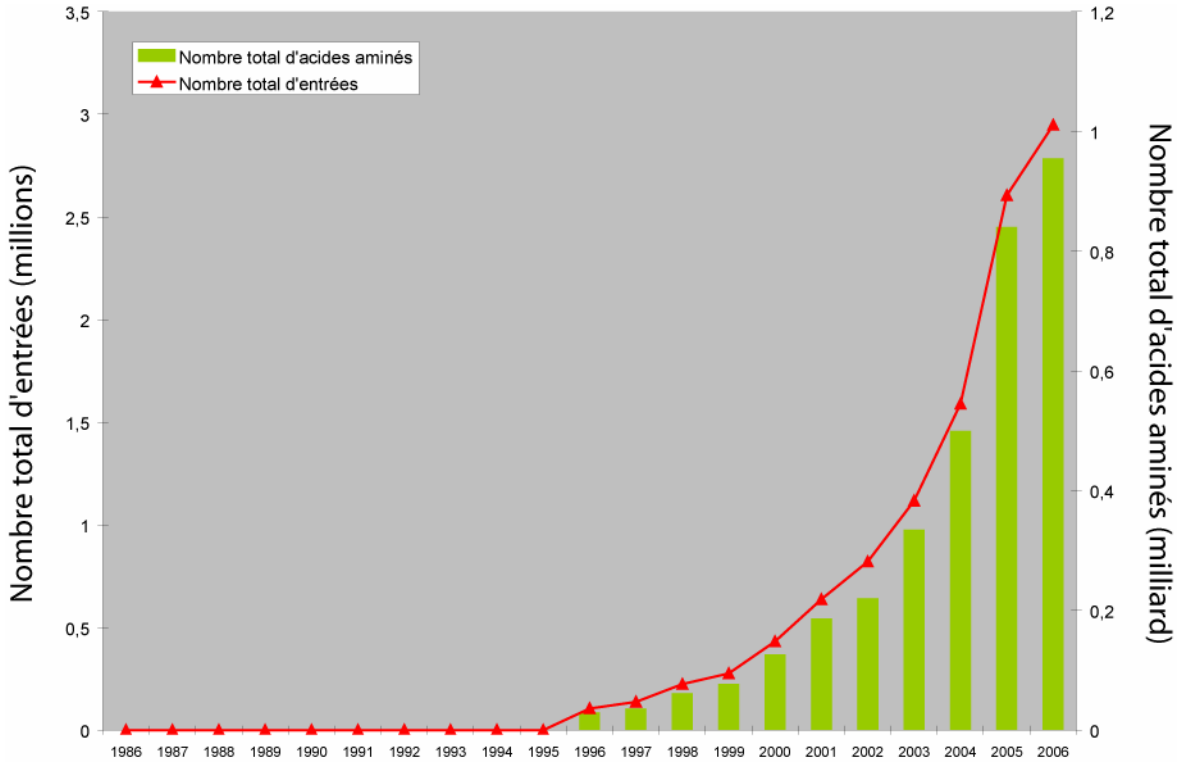


Figure 15 **Evolution du nombre d'entrée dans la banque TrEMBL depuis sa création en 1996.**
Les données sont extraites à partir de la documentation disponible sur le site de l'EBI (http://www.ebi.ac.uk/trembl/Documents/old_trembl_rel_notes.html).

TrEMBL se caractérise ainsi par une redondance importante, des séquences annotées automatiquement et des séquences souvent incomplètes et se veut désormais un complément parfait de Swiss-Prot.

2.2.1.5 PIR

Comme nous l'avons indiqué plus haut, les origines de la banque PIR (*Protein Information Resource*) sont anciennes puisque la toute première version remonte au milieu des années 60. Depuis 1988, cette banque de données a pris une dimension internationale puisqu'elle est maintenue, publiée et distribuée conjointement par la NBRF (*National Biomedical Research Foundation*) aux Etats-Unis, le MIPS (*Munich Information Center for Protein Sequences*) en Allemagne et la JIPID (*Japon International Protein Information Database*) au Japon.

Le but de cette collection est de fournir des informations exhaustives et non redondantes organisées selon des critères taxonomiques et de similarité (Wu *et al.* 2003). Si l'exhaustivité semble effectivement atteinte, il reste encore un taux de redondance non négligeable.

Les données proviennent de trois sources : les publications scientifiques, les soumissions des auteurs et la traduction des parties codantes annotées présentes dans les banques nucléotidiques. Quelques références croisées ont été mises en place avec les banques de

séquences nucléotidiques et quelques banques à valeur ajoutée, mais en nombre nettement moins important que dans Swiss-Prot.

Alors que dans Swiss-Prot, la classification des protéines en familles est réalisée essentiellement en utilisant les motifs PROSITE, PIR utilise une approche bien différente pour construire ses superfamilles. Cette approche est basée sur des similarités de séquence mais aussi de fonction (Barker *et al.* 1996).

2.2.2 Les banques dites à valeur ajoutée

2.2.2.1 PDB

La PDB (*Protein Data Bank*) est la principale banque internationale de structures tridimensionnelles (Kouranov *et al.* 2006). Cette banque a été fondée en 1971 par le BNL (*Brookhaven National Laboratory)* et contenait 7 structures. Depuis 1998, elle est sous la tutelle du RSCB (*Research Collaboratory for Structural Bioinformatics*).

Les structures de protéines constituent l'essentiel des entrées de la PDB avec 38096 entrées en août 2006, mais on y trouve également des structures de molécules d'ARN (511) et d'ADN (1068), de complexes protéines-acides nucléiques (1519). Ces structures sont déterminées expérimentalement par cristallographie aux rayons X, par RMN ou encore microscopie électronique. On peut noter que plus de 90% des structures déposées dans la PDB sont résolues par la technique de cristallographie aux rayons X.

Les entrées de la banque comprennent des informations sur les structures primaires et secondaires des molécules considérées, les coordonnées atomiques, souvent des détails des expériences (conditions de cristallisation, empilement cristallin, statistiques d'affinement, etc.) ainsi que des références bibliographiques.

Bien que le nombre de structures de macromolécules biologiques connues à l'heure actuelle soit très inférieur à celui des séquences (38096 structures dans la banque PDB contre 222289 protéines dans la seule banque Swiss-Prot en août 2006), celui-ci croît actuellement à une vitesse comparable à celle observée pour les séquences protéiques il y a quelques années (Figure 16). Cependant, il faut noter une redondance importante dans la PDB car plusieurs structures 3D peuvent correspondre à la même séquence selon les conditions d'obtention de la structure ou la finesse de sa résolution.

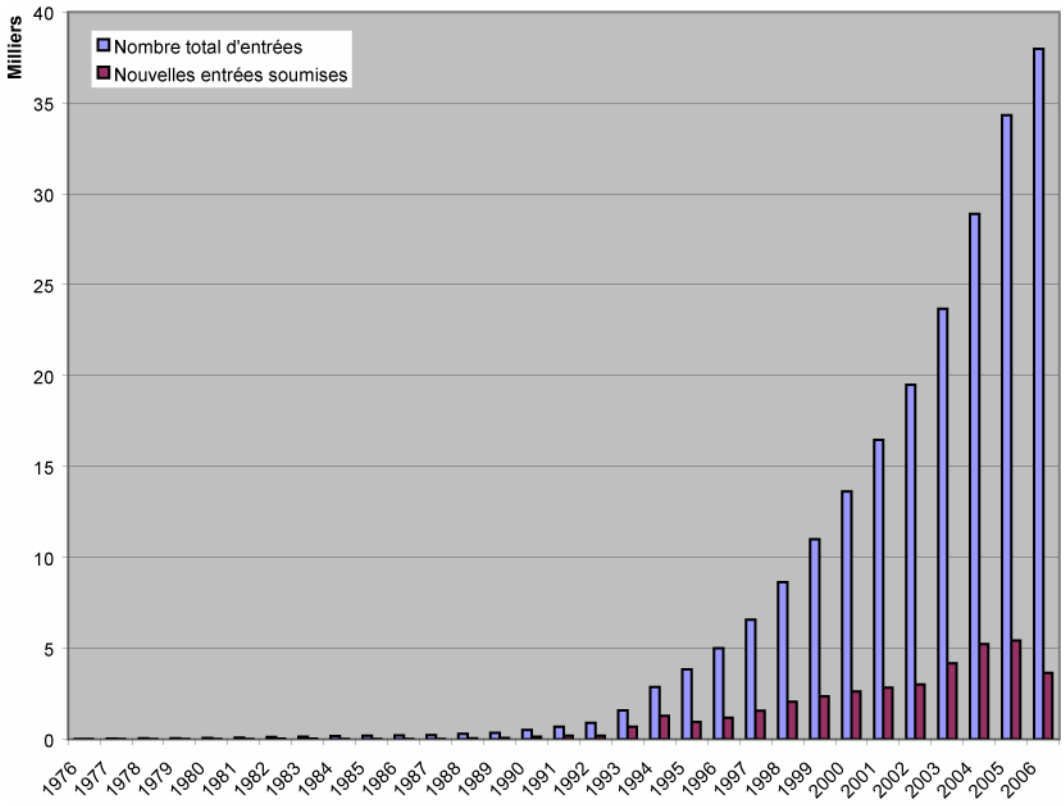


Figure 16 **Evolution du nombre d'entrée dans la banque de structures 3D, la PDB.**
Les données sont extraites à partir du site web de la Protein Data Bank pour le mois de juin 2006 (http://www.rcsb.org/pdb/contentGrowthChart.do?content=total&seqid=100).

2.2.2.2 RefSeq

RefSeq pour *Reference Sequence database* est une collection de séquences distribuée et générée par le NCBI depuis 2003, regroupant à la fois des séquences protéiques et des séquences nucléiques (Pruitt *et al.* 2005). L'idée est de distribuer des séquences de référence stables et utilisables pour différentes études fonctionnelles et médicales. Le point fondamental de cette banque est la validation constante appliquée aux séquences. La banque est disponible par organisme ou par taxon et permet donc de disposer d'une collection plus ou moins complète des séquences exprimées par un organisme donné (l'homme par exemple). Au niveau nucléique, RefSeq dérive de la banque GenBank, mais se distingue de cette dernière par plusieurs points. En effet, à la différence de GenBank qui se veut être un dépôt de toutes les séquences nucléiques existantes et contient de ce fait une redondance importante et nécessaire pour garder le caractère original des séquences soumises, RefSeq contient des séquences validées, annotées, non redondantes et reflète la synthèse des connaissances actuelles pour chacune de ses entrées. Au niveau protéique, la même politique de validation est appliquée.

2.2.2.3 UniGene

UniGene est une collection de séquences d'ESTs (*Expressed Sequence Tags*) et d'ARNm complets organisés en clusters et représentant chacun un gène unique connu ou inconnu pour un organisme donné. UniGene est distribuée par le NCBI depuis 1995 et représente une vue unifiée du transcriptome d'un organisme. Une entrée UniGene contient des données liées à son expression, une liste de protéines similaires, une localisation génomique et la liste des séquences qui constitue le cluster. Elle est constituée de 2 banques ; UniSeq qui contient les séquences consensus identifiées par cluster et UniGene a proprement dit qui contient les clusters calculés.

2.2.2.4 Gene Ontology (GO)

L'ontologie est une description formalisée des concepts d'un domaine de la connaissance. Elle contient des concepts (ou classes), des propriétés (ou attributs) de chaque concept décrivant les caractéristiques et les attributs d'une classe et des restrictions sur les propriétés (ou restrictions de rôles). La Gene Ontology (GO) (Ashburner *et al.* 2000) (http://www.geneontology.org) a vu le jour en 1998. Son objectif principal, comme son nom l'indique, est la gestion des informations liées aux gènes. GO est un système suffisamment large pour décrire l'ensemble des fonctions biologiques de toutes les espèces et suffisamment profond pour distinguer les spécificités d'une protéine particulière des autres membres de la famille.

GO est subdivisée en trois parties principales :

- Molecular Function : la (ou les) fonction d'un produit de gène (exemple : electron transporter activity). Elle décrit la tâche accomplie par les protéines individuelles.
- Biological Process : le (ou les) rôle biologique général de complexes et de fonctions moléculaires (exemple : electron transport). Il décrit le processus biologique dans lequel intervient le produit du gène considéré.
- Cellular Component : la (ou les) structure subcellulaire, localisation ou complexe macromoléculaire (exemple : extracellular space) dans lequel intervient le produit du gène considéré.

GO est devenue le standard dans l'annotation des génomes et est intégré comme référence croisée dans bon nombre d'autres banques de données, notamment Swiss-Prot.

Le vocabulaire contrôlé de GO permet des requêtes à différents niveaux. Il est ainsi aisé de rechercher tous les produits de gènes de la souris impliqués dans la transduction du signal.

2.2.2.5 Interpro

Interpro est une banque intégrée constituée d'informations sur les familles de protéines, les domaines et les sites fonctionnels identifiés (Apweiler *et al.* 2001). Elle est distribuée par l'EBI depuis 1999. La 13ème release permet de couvrir près de 75% des protéines disponibles dans la banque de protéines UniProt. Les éléments identifiés dans Interpro pour des protéines connues peuvent ensuite être utilisée pour décrire des protéines inconnues. Interpro est connectée avec GO et UniProt.

2.2.3 La qualité des données disponibles

Après des années de mise en place des moyens de collecte et d'interrogations efficaces et d'unification des données de séquences, se posent de plus en plus le problème de la mise en place de procédures pour assurer la qualité des données. Ceci se heurte à un certain nombre de problèmes

Il est important de souligner les efforts colossaux entrepris pour construire et maintenir à jour l'ensemble de ces banques de données. Il est également important de rappeler que, dans un grand nombre de cas, les séquences protéiques sont uniquement des prédictions à partir des séquences nucléiques déposées dans les banques et des efforts sont nécessaires pour valider ces prédictions. En bioinformatique, nous avons tendance à utiliser le maximum de ressources différentes pour valider les prédictions effectuées. Ainsi la diversité des banques peut être un atout mais également une contrainte.

La question du format des banques qui sont presque spécifiques à chaque banque est un frein évident si on ne dispose pas des outils adéquats. La constance de chaque format est également un problème. Un des exemples probant est la décision dans UniProt de fusionner en une seule entrée l'ensemble des séquences identiques correspondant au même gène chez plusieurs organismes (par exemple les actines). Ce changement majeur n'aura pas duré tant les conséquences étaient désastreuses sur les programmes utilisant les banques et sur l'analyse de leurs données. D'autres conséquences sont apparues plusieurs mois après la décision de re-séparer ces entrées puisqu'un certain nombre de données ont été perdues, notamment pour les actines.

Du point de vue des efforts consentis pour améliorer la qualité des données disponibles, on citera, par exemple, SwissProt (incluse dans UniProt) et RefSeq. En créant RefSeq, le NCBI a entrepris récemment des efforts importants pour assurer la haute qualité de ses séquences en les validant manuellement. Cependant, cette banque bien que très utilisée n'en reste pas moins isolée par le manque de liens vers les autres banques déjà existantes. Ceci pose un

autre problème dans le monde des banques de données, l'utilisation d'identifiants différents entres les différentes entrées pour un même gène ou une même protéine.

Enfin, un dernier point est la constance des identifiants des banques. La banque UniGene est un exemple de la versatilité des banques de données. UniGene est constituée de 2 banques ayant chacun leur identifiants ; les clusters d'un côté et les séquences de l'autre. Ainsi, il est fréquent de perdre, d'une version à une autre de la banque, le lien entre les identifiant de séquence et des clusters.

Les banques de données en biologie sont des outils fondamentaux et leur utilisation en bioinformatique est incontournable. Néanmoins, une connaissance approfondie de ces objets est essentielle afin les exploiter au maximum et éviter toute surprise. Ce problème est particulièrement pernicieux dans les approches à haut débit, où l'introduction d'erreurs ou d'approximations non-détectées à une étape quelconque du processus peut avoir des conséquences dramatiques pour l'interprétation finale.

2.3 La fouille de données

Depuis 1994, l'extraordinaire développement d'Internet a entraîné des modifications fondamentales dans les possibilités offertes aux utilisateurs et permet en particulier d'accéder aux banques de données et aux outils mis à disposition par les grands centres de bioinformatique. Cependant, bien que ces outils restent utilisables pour l'étude d'une petite quantité de séquence, des analyses à haut débit nécessitent souvent des développements supplémentaires. Ainsi des versions locales des banques de données sont indispensables à la réalisation de procédures d'analyses automatiques et systématiques.

Ainsi, afin de rendre les bases de données plus facilement exploitables et permettre aux utilisateurs d'en extraire de l'information, des logiciels spécifiques ont été développés.

Une entrée dans une banque de données biologique est définie par 2 entités ; les informations liées à la description de l'entrée, par exemple sa définition (« Actin béta »), et sa séquence. L'exploitation de ses données se fait classiquement de 2 manières ; la première appelée recherche textuelle permet de rechercher par mots clés directement dans les champs des entrées stockées dans les banques, la deuxième, la recherche par similarité, utilise uniquement les séquences des entrées et permet de déterminer la ou les séquences les plus proches d'une séquence dite « appât ».

2.3.1 La recherche textuelle

L'ancêtre de ces systèmes d'interrogation est le système ACNUC développé au Laboratoire de Biométrie et Biologie Evolutive il y a plus de vingt ans (Gouy *et al.* 1984). Ce système permet de structurer les informations de toute banque de séquences utilisant les formats GenBank, EMBL, Swiss-Prot et PIR et de les interroger sous la forme de requêtes multicritères.

Le système SRS (*Sequence Retrieval System*) développé au début des années 1990 par Etzold *et al* à l'EMBL, a ensuite été racheté par LION bioscience et plus récemment par biowisdom. Il permet d'interroger à l'aide d'une même interface pratiquement n'importe quelle collection de séquences disponible sous la forme de fichiers texte (Etzold *et al.* 1993). A l'heure actuelle, plus de 1300 banques de données différentes dispersées sur près de 40 sites Web publics sont interrogeables sous SRS. La liste des serveurs SRS dans le monde est disponible (http://downloads.lionbio.co.uk/publicsrs.html) et on peut citer en particulier ceux de l'EBI, de Pasteur, du Sanger et bien entendu de l'IGBMC et très prochainement du Centre de Recherche Public-Santé à Luxembourg.

Les concepteurs de SRS ont développé leur propre langage de programmation, ICARUS (*Interpreter of Commands and RecUrsive Syntax*) qui permet d'indexer toute collection structurée. Les fichiers à plat des banques de données sont parcourus par des interpréteurs syntaxiques et les champs contenant les données sont alors indexés. Ces index sont ensuite directement utilisés par « getz », le programme d'interrogation associé à SRS. SRS permet également s'il existe des champs en communs de réaliser des liens croisés entre les banques. Ces liens peuvent être directs comme par exemple entre GenBank et Swiss-Prot ou indirects comme par exemple entre GenBank et la PDB qui nécessitera de « passer » par une banque intermédiaire (Swiss-Prot) (Figure 17).

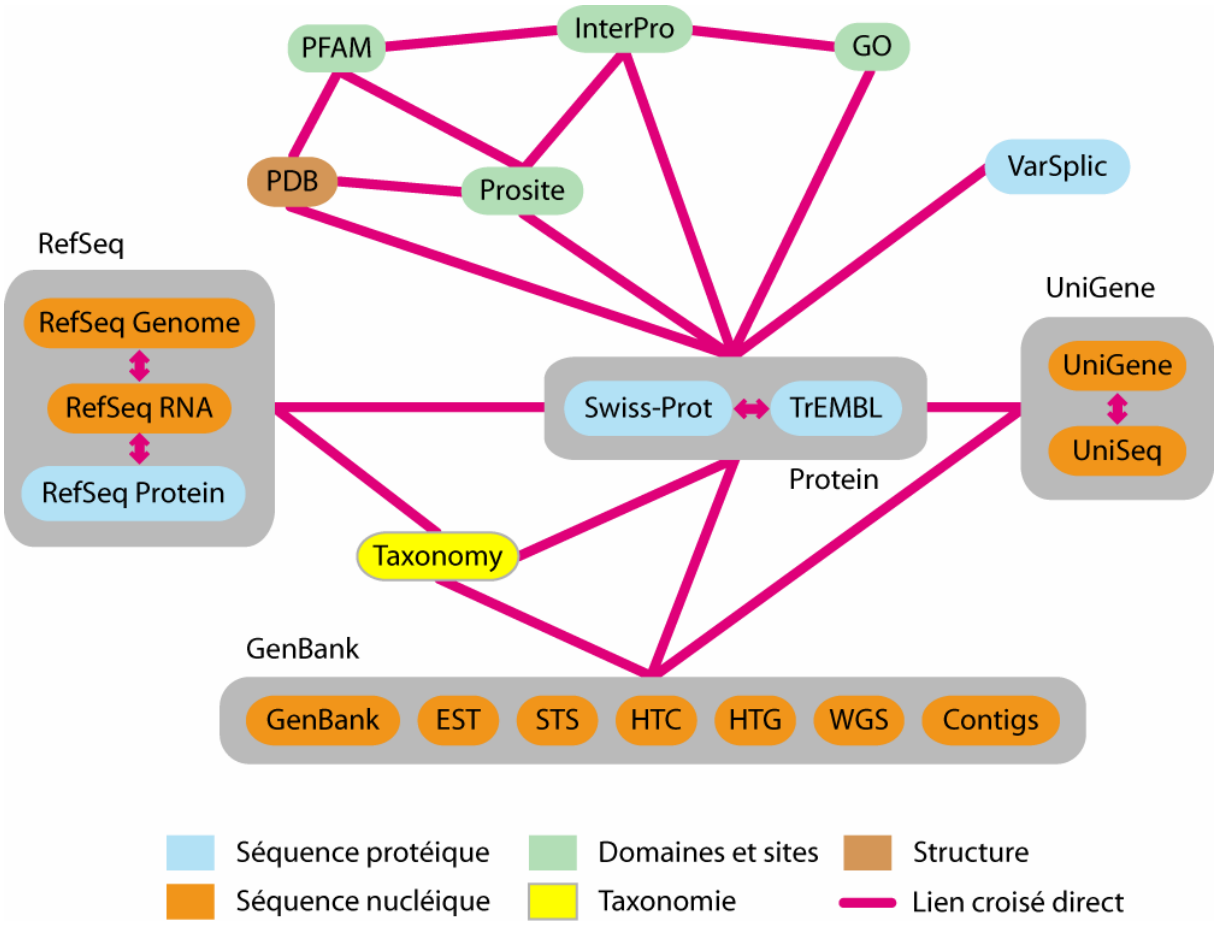


Figure 17 Relations entres les banques de données disponibles sur le serveur SRS à l'IGBMC.
Certaines banques présentes dans ce schéma ne sont pas décrites dans cette thèse. PFAM et PROSITE sont des banques de motifs et de sites fonctionnels. VarSplic est une sous-banque de la banque UniProt (SwissProt et TrEMBL) qui contient les séquences protéiques des variants d'épissage connus.

SRS peut être utilisé en mode ligne de commande au moyen de getz (utile pour l'incorporer dans des scripts automatiques) ou au travers de son interface web, SRSWWW et le programme « wgetz ». C'est cette dernière interface qui est la plus utilisée par les biologistes pour accéder aux informations d'une entrée.

Les requêtes sous SRS peuvent être extrêmement complexes car utilisant une syntaxe qui permet la combinaison de critères et de banques, des associations multiples et des liens croisés. En outre, SRS permet de réutiliser les résultats d'une requête pour en effectuer une autre. Il est à noter que la suite de programme d'analyse de séquence EMBOSS (Rice *et al.* 2000) est désormais accessible et intégrée à SRS.

2.3.2 La recherche de similarité

La recherche de similarité ou la comparaison entre une séquence inconnue et toutes les séquences d'une banque de données est une des premières étapes à laquelle un biologiste est confronté s'il veut identifier ou mieux caractériser sa séquence d'intérêt. La recherche de similarité fait appel à l'alignement d'une séquence contre une autre, processus par lequel les séquences sont comparées afin d'obtenir le plus de correspondances (identités ou similarités) possibles entre les lettres qui les composent. La recherche de similarité fait appel aux alignements 2 à 2 ou « pairwise » entre une séquence appât et une séquence d'une banque. Il s'oppose à l'alignement multiple qui contient plusieurs séquences (voir 2.4 L'alignement multiple).

Les alignements de séquences visent à établir des liens entre les séquences (voir 2.5 Un exemple d'application : l'annotation des protéines pour plus de détails).

Il existe plusieurs programmes basés sur des méthodes optimales, qui cherchent à trouver l'alignement du meilleur score. Ces algorithmes de Needleman-Wunsch (Needleman *et al.* 1970) ou Smith-Waterman (Smith *et al.* 1981) sont extrêmement sensibles, mais bien trop lents et couteux en mémoire. Des programmes basés sur des méthodes dites heuristiques permettent, à partir d'*a priori*, d'obtenir un résultat satisfaisant et rapidement. Ils privilégient la recherche de régions de forte identité et accélèrent ainsi la recherche. A la fin des années 1990, 2 programmes majeurs, FASTA et BLAST, ont été décrits et constituent les outils de base de la recherche de similarité dans les banques de séquences. Ils combinent à la fois efficacité et rapidité de leurs requêtes.

2.3.2.1 FASTA

Le principe de l'algorithme FASTA proposé par Pearson et Lipman (Lipman *et al.* 1985) est de ne considérer que les séquences présentant une région de forte similitude avec la séquence recherchée. Il applique ensuite localement à chacune de ces meilleures zones de ressemblance un algorithme d'alignement optimal. Cet algorithme se décompose et se déroule en quatre étapes (Figure 18) :

- (A) Les régions les plus denses en similarités entre les deux séquences sont recherchées. Ces régions sont appelés points chauds ou « *hot spots* ». C'est le paramètre « *ktup* » qui détermine le nombre minimum de résidus consécutifs identiques. Généralement : *ktup* = 2 pour les protéines - *ktup* = 6 pour l'ADN. Recherche des meilleures diagonales : plusieurs "*hot spots*" dans une même région génère des diagonales de similarité sans insertion ni délétions. Ces diagonales sont les régions ayant le plus de similarité. Elles sont représentées par un graphique de points ou « *dotplot* ».
- (B) Les dix meilleures diagonales sont réévaluées à l'aide d'une matrice de substitution et les extrémités de ces diagonales sont coupées afin de conserver les régions ayant les plus hauts scores seulement. Cette recherche de similitude est faite sans insertion ni délétion. Le score le plus élevé obtenu est appelé le score « *init1* ». Il est attribué à la région ayant le plus fort score parmi les 10 analysées.
- (C) Les diagonales trouvées à l'étape 1 dont le score dépasse un certain seuil (« cutoff »), sont reliées entre elles pour étendre la meilleure similarité. Ces nouvelles régions contiennent des insertions et/ou des délétions. Le score des nouvelles régions est calculé en combinant le score des diagonales réunies diminué d'un score de pénalité de jonction des diagonales. Le score le plus élevé obtenu à cette étape s'appelle le score « *initn* ». Cette étape permet d'éliminer les segments peu probables parmi ceux définis à l'étape précédente.
- (D) La région initiale qui a généré le score « init1 » est de nouveau évaluée avec un algorithme de programmation dynamique sur une fenêtre de résidus dont la largeur est déterminée par le paramètre « *ktup* ». Le nouveau score est « *opt* ». Les séquences de la base de données sont classées selon leurs scores « *initn* » ou « *opt* ». Les séquences sont alignées avec la séquence cible à l'aide de l'algorithme de Smith & Waterman : le score final est le score Smith & Waterman.

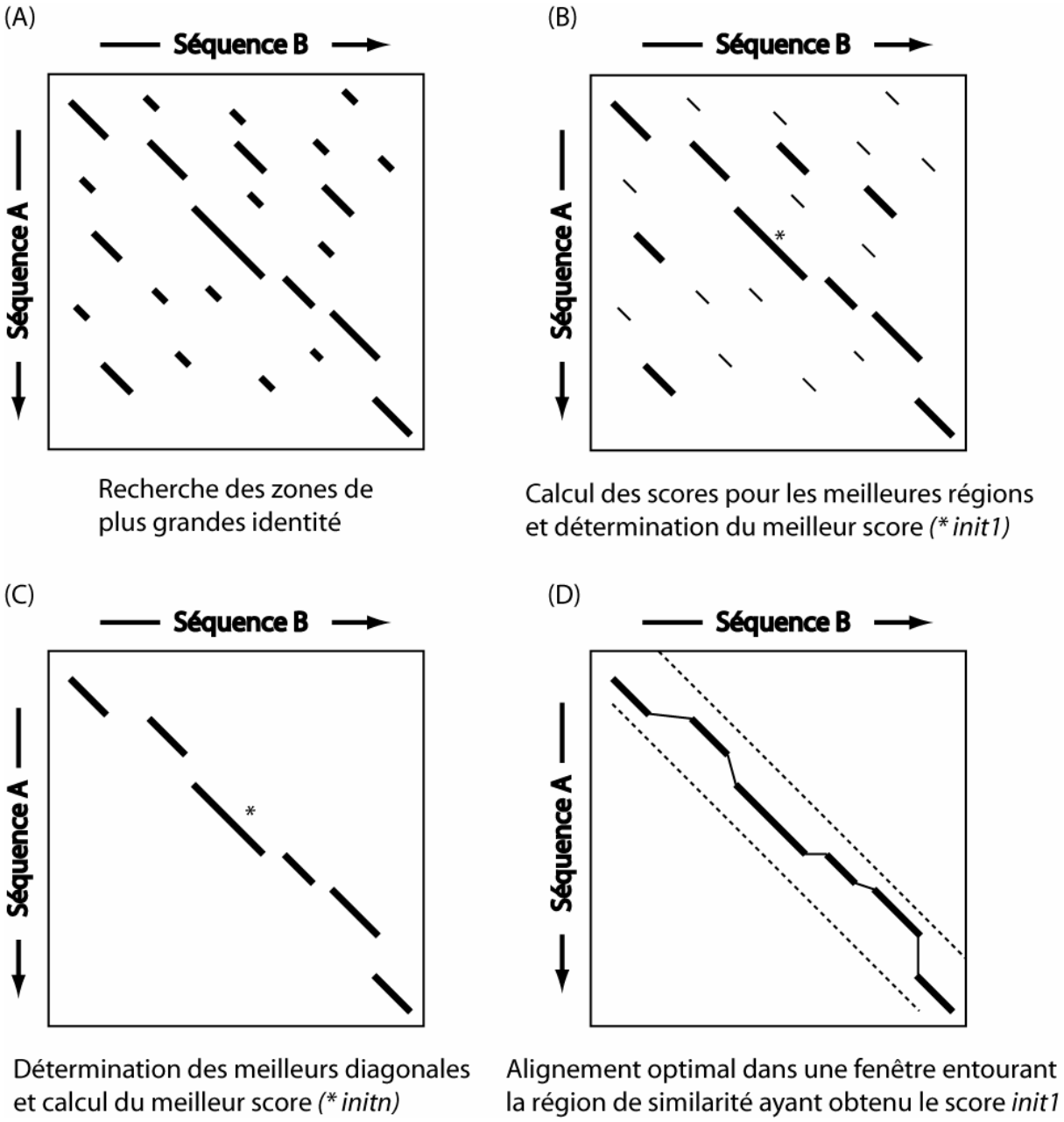


Figure 18 **Algorithme du programme FASTA.**

Cet algorithme a toutefois un inconvénient, bien que plus rapide que les algorithmes optimaux il n'en reste pas moins plus lent que BLAST. Cependant, sa sensibilité accrue le rend plus apte à la comparaison de séquences nucléotidiques et l'identification d'une séquence nucléotidique « appât » sur un génome complet, par exemple.

2.3.2.2 BLAST

BLAST (Basic Local Alignement Search Tool) (Altschul *et al.* 1990) est le programme de recherche de similarité le plus utilisé pour les comparaisons protéine-protéine. Il fait la même hypothèse que pour FASTA, c'est-à-dire que les meilleurs alignements doivent

contenir quelque part des petits segments strictement identiques ou de score très élevé. Ces éléments constituent les points d'ancrage à partir desquels l'alignement est étendu. L'algorithme initial de BLAST ne permet ni les insertions ni les délétions, mais il est très rapide et attribue une valeur statistique au score obtenu. L'algorithme initial a été modifié plusieurs fois pour répondre à différents besoins. Ainsi, BLAST2 autorise les insertions et les délétions (mais la statistique n'est plus exacte) alors que PSI-BLAST est une version qui construit des motifs consensus stockés sous forme de matrices qui seront utilisés lors de recherches et d'alignements itératifs (Altschul *et al.* 1997).

De plus, des filtres comme les programmes SEG (Wootton *et al.* 1996) et DUST (Tatusov et Lipman données non publiées) ont été conçus pour éliminer les régions répétitives et segments de « faible complexité » qui bruitent les résultats. Pour cela, la séquence « appât » est tout d'abord comparée à une banque de données contenant des séquences représentatives des familles surreprésentées dans les banques. Les sous-fragments de la séquence « appât » appartenant à ces familles ainsi que les régions poly X sont alors masqués (et ne seront pas utilisés) avant d'effectuer la recherche de similitude sur la banque complète.

Du point de vu algorithmique BLAST se décompose également en quatre étapes (Figure 19) :

- (A) La séquence à aligner est décomposée en « mots » d'une longueur W fixe (3 par défaut pour les protéines, 11 pour les nucléotides) et une banque de mots proches ou similaires aux mots de la séquence initiale est créée.
- (B) Chacun de ces mots est alors recherché à l'identique dans toutes les séquences de la base.
- (C) Les alignements autour de ces mots trouvés sont alors étendus tant que le score de cet alignement ne chute pas d'une valeur X ou qu'une extrémité ne soit atteinte. Les alignements sont alors appelés HSPs (*High-scoring Segment Pair*). Le HSP de plus haut score est appelé MSP (*Maximal Segment Pair).*
- (D) Enfin, les HSPs qui ont un score supérieur à un seuil S sont sélectionnés et affichés. Si il existe plusieurs HSPs par séquence détectée leur score sont combinés.

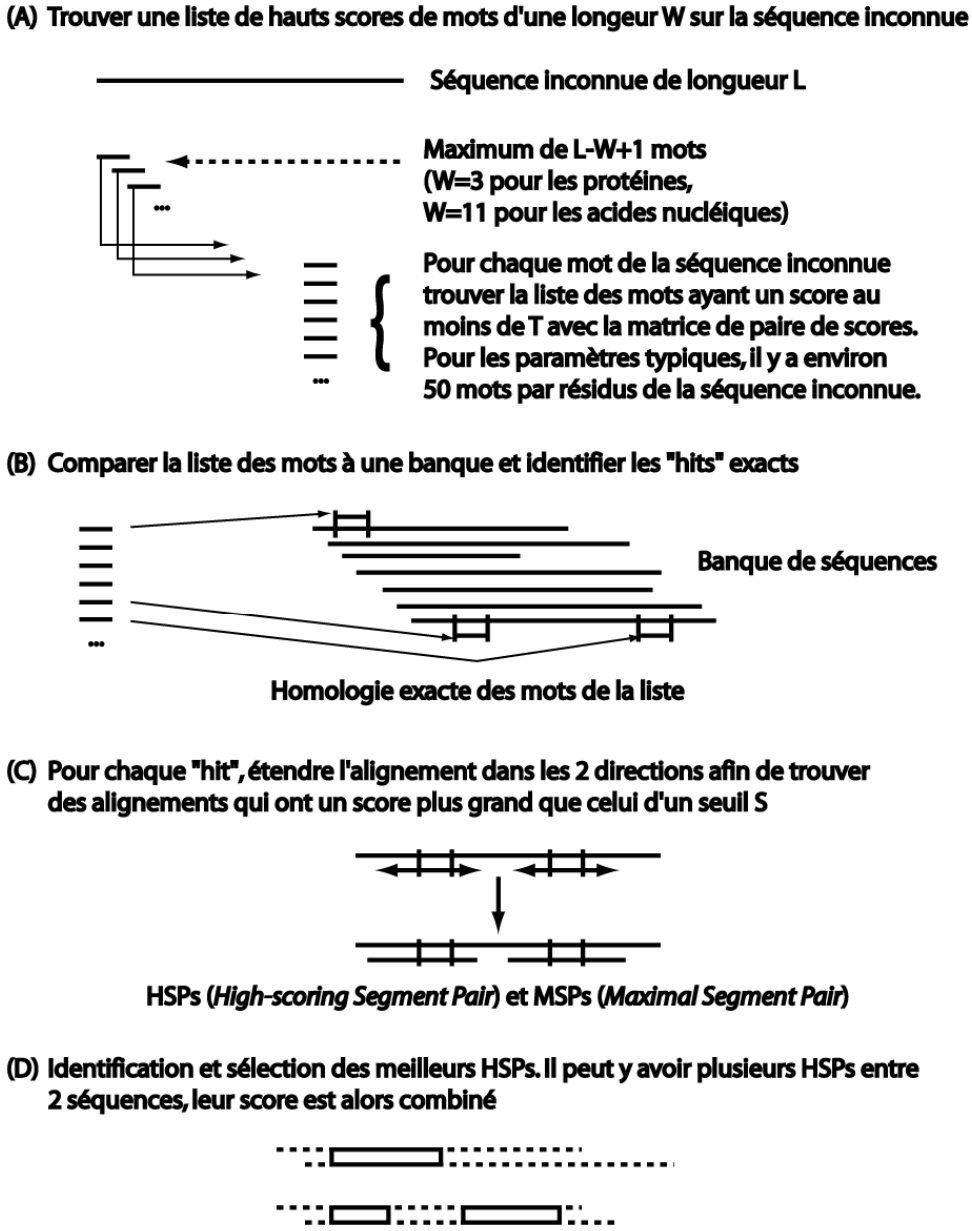


Figure 19 **Algorithme de BLAST.**

Enfin, l'interprétation des résultats ou l'analyse de la signification des alignements est un point capital. Cette signification est évaluée statistiquement en fonction de la longueur et de la composition de la séquence, de la taille de la banque et de la matrice de scores utilisée. BLAST calcule ainsi un score, l'« E-value » qui correspond à la probabilité d'observer au hasard cet alignement dans les banques de séquences considérées. Ainsi plus la « E-value » est faible, plus l'alignement est significatif (Tableau 3). Cependant, la décision finale repose aussi et (peut-être surtout ?) sur une inspection visuelle du résultat par l'expérimentateur.

E-value	Signification
$E<1e^{-100}$	appariement exact, même séquence, même origine
$1e^{-100} < E < 1e^{-50}$	séquences quasiment identiques (allèles, mutations, espèces voisines)
$1e^{-50} < E < 0,1$	un éventuel lien entre la séquence requête et celles qui ont été trouvées
$E > 0,1$	séquences de l'alignement à rejeter, sans lien avec la séquence requête

Tableau 3 **Interprétation des valeurs d'expect calculés par BLAST.**

On peut noter que pour des séquences requêtes très courtes, la "E-Value" est très faible, on ne peut donc pas se baser sur ces valeurs pour estimer la signification des alignements (par exemple pour des alignements d'oligonucléotides). Les séquences alignées et détectées par BLAST sont appelés également « Subject » ou « hit ».

2.3.2.3 Les différentes possibilités

Les programmes FASTA et BLAST permettent de comparer l'ensemble des types de séquence à l'ensemble des types de banques (Tableau 4). Les programmes utilisés dépendent de la séquence appât, de la banque à testée et surtout de l'utilisation souhaitée.

Programme	Séquence appât	Banque	Exemple d'utilisation
BLASTP/FASTA	Protéine	Protéine	Recherche de protéines homologues
BLASTN/FASTA	ADN	ADN	Recherche d'ARNm proches
TBLASTN/TFASTA	Protéine	ADN*	Localiser une protéine sur un génome
BLASTX/FASTX	ADN*	Protéine	Définir les cadres de lecture
TBLASTX/TFASTX	ADN*	ADN*	idem que TBLASTN et BLASTX

*traduit en séquence protéique dans les 6 cadres de lecture

Tableau 4 **Les différents programmes BLAST et FASTA ainsi que leurs utilisations.**

2.4 L'alignement multiple

L'étude d'une séquence protéique dans le contexte de sa famille constitue une étape cruciale dans les possibilités offertes au bioanalyste. Cet outil est malheureusement souvent négligé. Dans ce cadre, l'alignement multiple de séquences complètes ou MACS (Lecompte *et al.* 2001), est donc un outil central dans la biologie moderne, mais à quoi correspond ce type d'alignement ? Un alignement multiple est l'alignement de plusieurs séquences en même temps. Il tient en général compte de la totalité de la longueur des séquences. Il s'oppose aux alignements 2 à 2 (voir 2.3.2 La recherche de similarité). Dans un alignement multiple, si les séquences ne sont pas suffisamment similaires sur certaines parties ou si leurs longueurs sont différentes, il faut introduire des insertions créant ainsi des lacunes ou « gap » (Figure

20). Généralement, il contient les séquences d'une famille de protéines en y intégrant, s'ils existent, les membres issus du même organisme que la séquence initiale et ceux issus de différents organismes.

Un MACS permet en outre d'intégrer pour chaque séquence à la fois une analyse horizontale en tenant compte de sa longueur totale et une analyse verticale en l'intégrant dans son contexte évolutif.

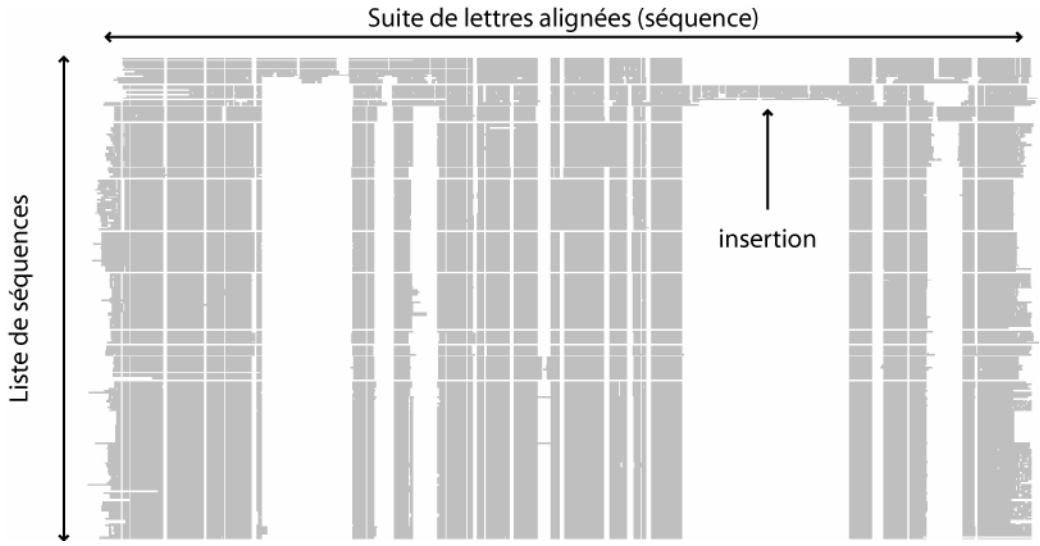


Figure 20 **Vue schématique d'un alignement multiple de séquences.**
L'alignement multiple contient des séquences alignées sur toute leur longueur. Les séquences d'une famille de protéines peuvent contenir des protéines issues de plusieurs organismes. L'alignement des protéines sur toute leur longueur, peut selon les cas, nécessiter la création d'insertion dans un groupe de séquence correspondant à des lacunes ou « gaps » dans d'autres.

L'alignement multiple permet d'adresser bon nombre des questions de la biologie moderne comme :

- La visualisation de l'organisation en domaines de la protéine et la présence d'insertions ou de délétions.
- La validation de la séquence protéique en permettant de détecter les erreurs de séquençage, les décalages du cadre de lecture, les erreurs de prédiction de la structure du gène (codon initiateur, site d'épissage exon/intron).
- De replacer la protéine au sein de sa famille et donc dans son contexte évolutif (Phillips *et al.* 2000).
- De distinguer les paralogues des orthologues.
- De prédire la fonction par la propagation de ce type d'information à des séquences non annotées.

- De préciser les prédictions de structures secondaires aussi bien au niveau 2D (Lee *et al.* 2006) qu'au niveau 3D (Yang *et al.* 2000; Al-Lazikani *et al.* 2001).
- D'analyser les conservations au sein d'une famille de protéines, notamment les résidus strictement conservés dans toutes les séquences (résidus ou motifs contenant une importance structurale ou fonctionnelle) ou encore les résidus conservés spécifiquement dans un sous-groupe de séquences (résidus discriminants de ce sous-groupe) (del Sol Mesa *et al.* 2003).

2.5 Un exemple d'application : l'annotation des protéines

Une des applications majeures de l'ensemble des banques et des outils décrits précédemment réside dans l'annotation des séquences. L'annotation d'une séquence nucléique ou protéique consiste à lui attribuer des éléments qui vont permettre de la définir, de la classer, de mieux comprendre son rôle dans la cellule, son fonctionnement moléculaire et ses interactions avec d'autres protéines, par exemple.

On distingue plusieurs stratégies d'annotation, la première dite « classique » s'appuie sur la recherche de similarité, la deuxième consiste à intégrer la séquence dans le contexte de sa famille de gènes ou de protéines, enfin une troisième consiste à utiliser les propriétés communes à d'autres séquences.

2.5.1 Stratégie classique

De manière classique, la première étape de l'annotation consiste à comparer les séquences protéiques inconnues, ou prédites à partir de nouvelles séquences nucléotidiques, à une banque généraliste de séquences protéiques bien caractérisée et à transférer la fonction des protéines similaires aux séquences inconnues. Pour cela, on utilise l'ensemble des programmes définis précédemment (voir 2.3 La fouille de données p55) (BLASTP, FASTA, ou encore PSI-BLAST, ce dernier permettant de détecter des similarités entre des protéines très éloignées). Seuls les alignements présentant une valeur d'expect inférieure à une valeur seuil sont pris en compte. Cette valeur seuil, fixée arbitrairement, est extrêmement variable d'un protocole à l'autre. De la même manière, le transfert de la fonction peut obéir à différentes règles :

- transfert pur et simple de la fonction du meilleur « hit » à la séquence à annotée.
- comparaison des meilleurs hits pour en vérifier la cohérence et choisir la plus informative.

- attribution d'une fonction plus ou moins précise suivant le degré de similarité, allant de la description précise de la fonction moléculaire (par exemple, le numéro complet d'une enzyme suivant la classification officielle) à une description très générale (« actin » ou « actin-like »).

Le report automatique de la fonction d'une protéine à une autre sur la base d'un alignement local peut générer de nombreuses erreurs (Galperin *et al.* 1998; Devos *et al.* 2001). L'un des problèmes majeurs est l'absence de prise en compte de l'organisation en domaines de la protéine. L'intégration des informations fournies par l'alignement multiple ainsi que les données complémentaires (domaines fonctionnels, mutations, structures secondaires…) sont des outils totalement complémentaires et indispensables à cette approche.

2.5.2 Une famille de protéine

La prédiction de fonction par similarité repose sur l'hypothèse que des gènes « suffisamment proches » sont susceptibles d'être homologues, c'est-à-dire dériver d'un même gène ancestral, et d'avoir conservé la même fonction ou, à tout le moins, le même repliement tridimensionnel. Pratiquement, ceci veut dire que la proximité phylogénétique entre deux séquences est proportionnelle à leur similarité. Les séquences possèdent ainsi une séquence ancêtre commun à partir de laquelle des mutations, insertions et délétions se sont accumulées au cours de l'évolution pour aboutir à ces 2 protéines. L'homologie entre deux séquences peut laisser supposer, mais sans le prouver, que les protéines codées assurent une fonction similaire ou proche. Or, l'évolution d'une famille de gènes peut comporter de nombreux événements. Dès 1970, Fitch (Fitch 1970) a introduit différents termes pour classer les différents types d'homologie :

- gènes orthologues : gènes homologues ayant évolué à partir d'un gène ancestral (gène X_0) par spéciation (gènes X_1 et X_2 de la Figure 21).
- gènes paralogues : gènes homologues ayant évolué à partir d'un gène ancestral par duplication (gènes Y_0 et Z_0 de la Figure 21) au sein de la même espèce.
- gènes xénologues : gènes homologues présents chez un même organisme, l'un ayant été hérité verticalement, l'autre par transfert horizontal, c'est-à-dire acquisition du matériel génétique d'une autre espèce (gènes X et Y de la Figure 21).

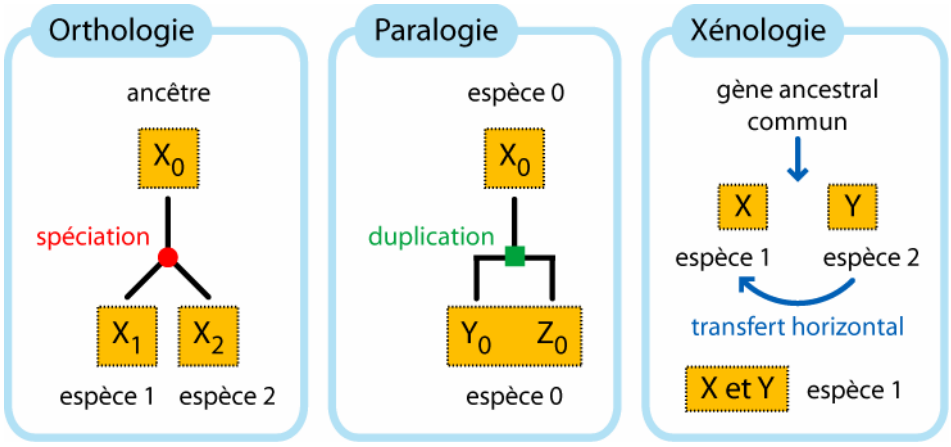


Figure 21 **Les différentes relations d'homologie.**

2.5.3 D'autres méthodes

D'autres approches permettent de générer des informations précieuses sur la protéine. Le « diagnostique de séquence », selon l'expression de Tsoka et Ouzounis (Tsoka *et al.* 2000), consiste à prédire certaines caractéristiques de la protéine en se basant sur sa seule composition (présence d'hélices transmembranaires…). Enfin, une nouvelle catégorie de méthodes rassemblées sous l'expression générale « prédiction par non-homologie » ont vu le jour (revues dans (Eisenberg *et al.* 2000) et (Marcotte 2000)). Ces méthodes reposent principalement sur l'utilisation du contexte génomique ou de la distribution phylogénétique des gènes pour inférer des fonctions ou, tout au moins, des liens fonctionnels entre protéines (voir 3.2 Les outils de la génomique comparative).

Chapitre 3 - La génomique comparative

« Les espèces qui survivent ne sont pas les espèces les plus fortes,
ni les plus intelligentes, mais celles qui s'adaptent le mieux aux changements. »
Charles Darwin

La génomique comparative est l'analyse et la comparaison des génomes de différentes espèces. Son but est une meilleure compréhension de l'évolution des espèces, la détermination de la fonction des gènes, l'organisation des éléments d'un génome (taille des gènes, structure des gènes…) et la mise en évidence de gènes d'intérêts comme de nouveaux gènes impliqués dans des maladies infectieuses ou héréditaires (Figure 22).

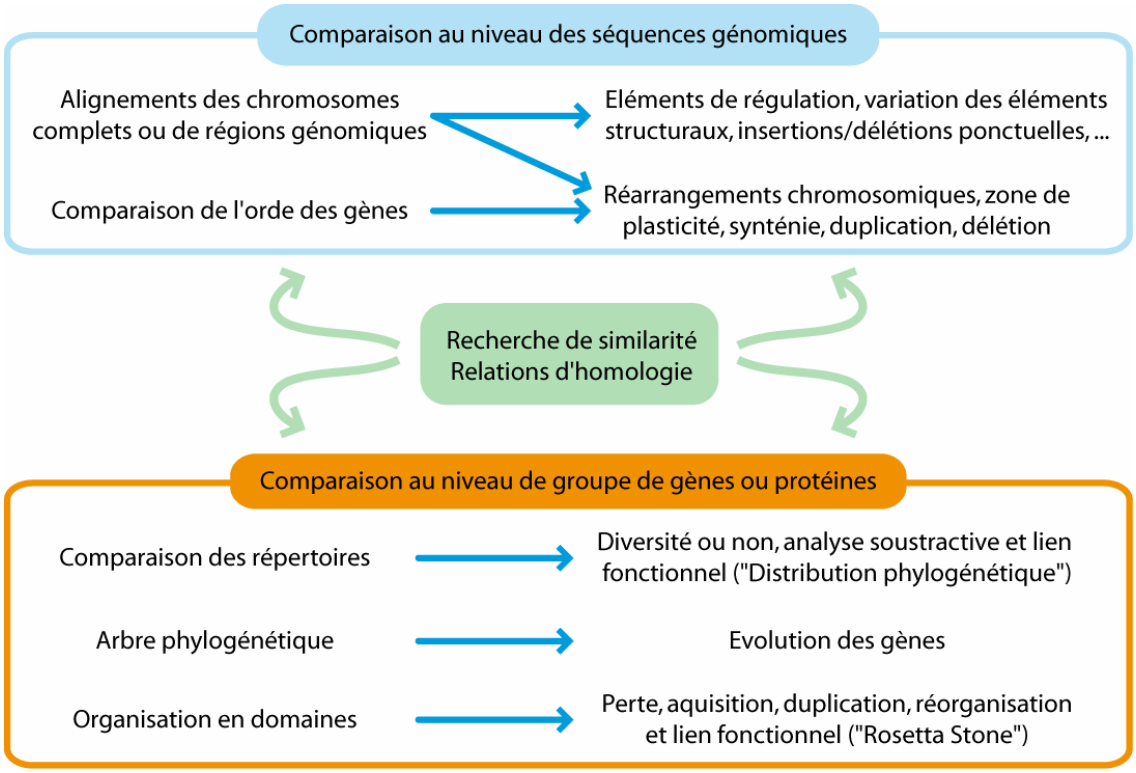


Figure 22 Schéma représentant certains niveaux de comparaison en génomique comparative.
La recherche de similarité et les relations d'homologie apparaissent centrales pour l'ensemble des applications de la génomique comparative.

Les éléments cruciaux au développement de la génomique comparative sont la disponibilité des séquences complètes de génomes et de protéomes d'organismes couvrant un maximum

de branches de l'arbre de la vie, la possibilité de comparer de façon rapide et fiable des séquences similaires et d'établir les relations d'homologies entre les séquences. Son importance dans la biologie post-génomique est indéniable tant le nombre d'applications auxquelles elle contribue sont essentielles.

Nous verrons donc dans ce chapitre tout d'abord, la matière première indispensable à la génomique comparative c'est-à-dire les génomes et les enseignements que nous pouvons en dégager, puis nous nous attarderons sur certaines des techniques de génomique comparative et leurs applications.

3.1 Les génomes complets

Disposer de la séquence complète du génome d'un organisme est un évènement majeur puisqu'il permet a priori de disposer de l'ensemble des informations nécessaire à sa compréhension. Cette information est non seulement utile pour replacer l'organisme en question dans l'arbre du vivant et comprendre son évolution mais elle permet également de contribuer à une meilleure compréhension des autres organismes complets en les comparants entre eux. En 1990, Woese *et al.* ont proposé (Woese *et al.* 1990) une séparation du domaine du vivant en 3 grands domaines différenciant ainsi les archées (Archaea), les bactéries (Bacteria) et les eucaryotes (Eukaryota) (Figure 23).

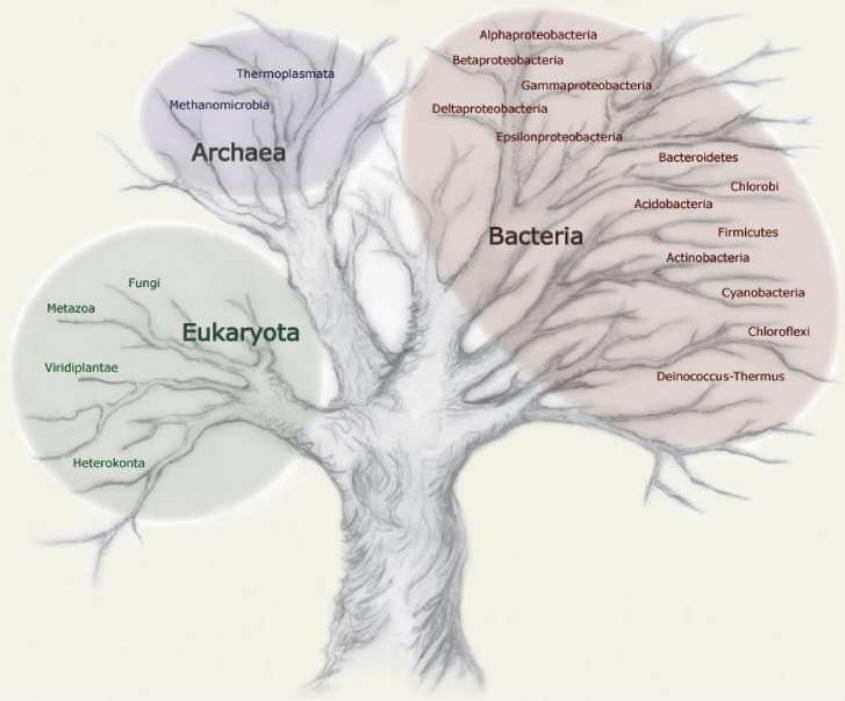


Figure 23 **Arbre de la vie**
L'arbre comprend les 3 grands domaines de la vie ; les bactéries, les archées et les eucaryotes (source http://genome.jgi-psf.org/tre_home.html)

Les points qui suivent nous permettront d'aborder le nombre de génomes complets disponibles, la taille et la structure des génomes, le nombre et la structure des gènes. Au cours de ces 3 parties, nous couvrirons l'ensemble des règnes du vivant tout en favorisant l'étude des organismes eucaryotes.

3.1.1 Les génomes eucaryotes

Chez les eucaryotes supérieurs, le caractère « complet » d'une séquence génomique est difficile à apprécier en raison de l'importance de régions centromériques et télomériques essentiellement constituées de séquences d'ADN fortement biaisées en composition et hautement répétées. Fonctionnellement, ces séquences se trouvent souvant dans l'hétérochromatine qui se définit comme l'ensemble des régions très condensées de la chromatine, inactives pour la transcription et répliquées tardivement. Ainsi, la taille estimée du génome nucléaire de *Drosophila melanogaster* est d'environ 180 Mb dont un tiers est représenté par de l'hétérochromatine localisée dans les régions centromériques (Adams *et al.* 2000).

En raison de leur composition, les régions hétérochromatiques peuvent être difficiles à cloner et à séquencer, la plupart des programmes de séquençage de génomes d'eucaryotes supérieurs ne s'intéressent par conséquent qu'à la portion euchromatique. L'euchromatine se définit comme l'ensemble des régions non condensées dans la chromatine. Il est bon de souligner que les génomes complets publiés n'ont pas tous le même stade d'avancement de séquençage et d'assemblage, leur statut allant du « draft » comme dans le cas par exemples, de *Gallus gallus*, de *Plasmodium yoelii* ou d'*Orysa sativa* à la séquence complète comme dans le cas de *Saccharomyces cerevisiae*. Par ailleurs, de nombreux génomes de mitochondries et de chloroplastes ont également été totalement séquencés.

Le génome de la levure *S. cerevisiae* a été le premier génome eucaryote séquencé entièrement, au terme d'une collaboration internationale entre plus de 600 chercheurs (Goffeau *et al.* 1996). Depuis, un nombre important d'autres génomes nucléaires d'eucaryotes ont été séquencés plus ou moins complètement et rendus publiques (Tableau 5).

Tableau 5 **Liste des génomes eucaryotes complets.**

	Organisme	Taille du génome (Mb)	Nombre de gènes estimés	Référence
Chordé	*Homo sapiens*	2900	23299	(Lander *et al.* 2001) (Venter *et al.* 2001)
	Mus musculus	2600	24948	(Waterston *et al.* 2002)
	Rattus norvegicus	2750	21022	(Gibbs *et al.* 2004)
	Gallus gallus	1000	23000	(Hillier *et al.* 2004)
	Xenopus laevis	3100		
	Xenopus tropicalis	1700	28000	
	Tetraodon nigroviridis	342	27918	(Jaillon *et al.* 2004)
	Danio rerio	1626,1	21503	
	Takifugu rubripes	320	22208	(Aparicio *et al.* 2002)
	Ciona intestinalis	162	14182	(Dehal *et al.* 2002)
Echinoderme	*Strongylocentrotus purpuratus*	800	27350	
Arthropode	*Drosophila melanogaster*	122	13676	(Adams *et al.* 2000)
	Anopheles gambiae	278,2	13683	(Holt *et al.* 2002)
Nématode	*Caenorhabditis elegans*	103	19893	(Consortium 1998)
	Caenorhabditis briggsae	102	19507	(Stein *et al.* 2003)
Ascomycète	*Schizosaccharomyces pombe*	12,5	4929	(Wood *et al.* 2002)
	Saccharomyces cerevisiae	12,2	5807	(Goffeau *et al.* 1996)
	Candida glabrata	12,3	5283	(Dujon *et al.* 2004)
	Kluyveromyces lactis	10,6	5329	(Dujon *et al.* 2004)
	Ashbya gossypii	9,2	4718	(Dietrich *et al.* 2004)
	Debaryomyces hansenii	12,2	6906	(Dujon *et al.* 2004)
	Yarrowia lipolytica	20,5	6703	(Dujon *et al.* 2004)
	Aspergillus fumigatus	29,4	9926	(Nierman *et al.* 2005)
	Neurospora crassa	38,6	10082	(Borkovich *et al.* 2004)
Basidiomycète	*Cryptococcus neoformans*	19,2	7302	(Loftus *et al.* 2005)
Microsporidie	*Encephalitozoon cuniculi*	2,5	1997	(Katinka *et al.* 2001)
Mycetozoa	*Dictyostelium discoideum*	34,0	12500	(Eichinger *et al.* 2005)
Amibe	*Entamoeba histolytica*	23,8	9938	(Loftus *et al.* 2005)
Straménopiles	*Thalassiosira pseudonana*	34	11242	(Armbrust *et al.* 2004)
Euglène	*Trypanosoma brucei*	26,1	9068	(Berriman *et al.* 2005)
	Trypanosoma cruzi	60,4	23216	(El-Sayed *et al.* 2005)
	Leishmania major	34	8272	(Ivens *et al.* 2005)
Alvéolates	*Plasmodium falciparum*	22,9	5268	(Gardner *et al.* 2002)
	Plasmodium yoelii	23,1	5878	(Carlton *et al.* 2002)
	Cryptosporidium parvum	9,1	3952	(Abrahamsen *et al.* 2004)
	Cryptosporidium hominis	9,2	3994	(Xu *et al.* 2004)
	Theileria parva	8,3	4035	(Gardner *et al.* 2005)
	Tetrahymena thermophila	106	27400	
Diplomonodie	*Giardia lamblia*	11,7	9649	
Dicotylédons	*Arabidopsis thaliana*	125	25498	(Arabidospis 2000)
Monocotylédons	*Oryza sativa*	430	42653	(Goff *et al.* 2002; Yu *et al.* 2002)
Rhodophyte	*Cyanidioschyzon merolae*	16,5	5331	(Matsuzaki *et al.* 2004)

3.1.2 Nombre de génomes

En juin 2006, la communauté scientifique dispose de 405 génomes dont la séquence est dite complète (consultation du site GOLD en juin 2006) qui se répartissent en 27 archées, 337 bactéries et 41 eucaryotes (Liolios *et al.* 2006). Ces informations ainsi que la liste des projets en cours sont répertoriées sur un certain nombre de sites web de grands centres de séquençage ou de sites dédiés aux génomes (Tableau 6).

Nom du site	Adresse du site
GOLD	http://www.genomesonline.org
NCBI	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genomeprj
EBI	http://www.ebi.ac.uk/genomes
TIGR	http://www.tigr.org/tdb
SANGER	http://www.sanger.ac.uk/Projects
JGI	http://www.jgi.doe.gov

Tableau 6 **Liste non exhaustive des sites internet majeurs regroupant l'information sur les génomes séquencés ou en cours de séquençage.**

Depuis la révolution introduite par le séquençage de génomes comme celui de l'Homme ou de la drosophile, le séquençage de génomes complets est devenu un outil banalisé. Cependant, la publication ou la disponibilité de la séquence complète d'un génome constitue encore un événement majeur car il est considéré comme un élément indispensable à la compréhension des grands processus biologiques de l'organisme étudié, voire des maladies qui l'affectent. Ainsi, en juin 2006, d'après le site GOLD pas moins de 1665 projets de séquençage sont en cours de réalisation. Parmi ces projets, on distingue 57 projets d'archées, 979 projets de bactéries et 629 projets eucaryotes. On peut ainsi apprécier la croissance quasi exponentielle que prend la courbe du nombre de projet en cours de séquençage (Figure 24).

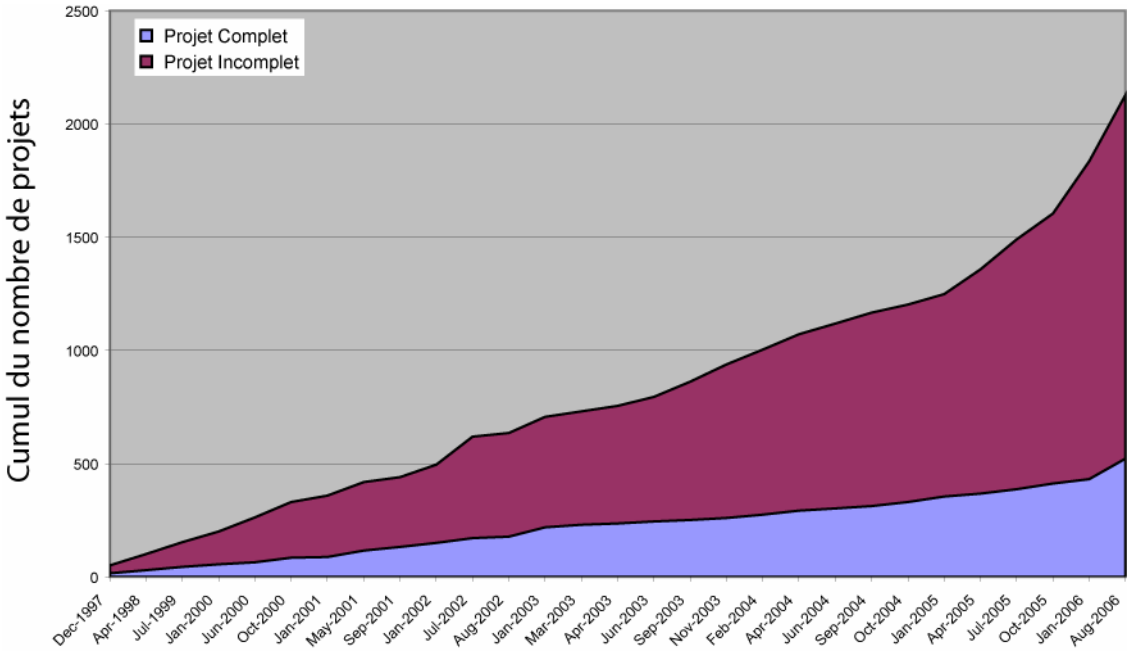


Figure 24 **Evolution du nombre de projets total en fonction du statut final (juin 2006).**

L'analyse un peu plus détaillée de la répartition du nombre de projets référencés sur le site GOLD montre que cette croissance est majoritairement due aux projets bactériens et eucaryotes (Figure 25).

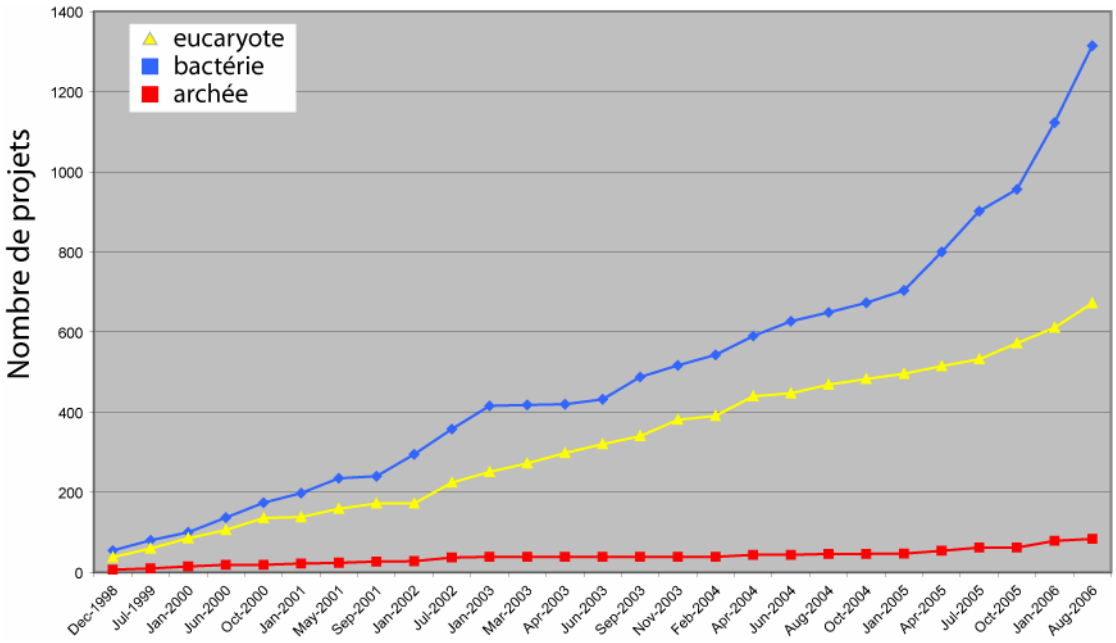


Figure 25 **Evolution du nombre de projets disponibles sur le site GOLD en fonction du groupe phylogénétique (juin 2006).**

3.1.3 Structure et taille des génomes

L'organisation générale ainsi que la taille des génomes est classiquement liée aux types d'organismes. En effet, les procaryotes possèdent des génomes de petites tailles souvent organisés en un unique chromosome circulaire, complémenté éventuellement par un ou plusieurs plasmides. Toutefois, il est possible de trouver plusieurs contre-exemples. Ainsi, les génomes bactériens *Vibrio cholerae* (Heidelberg *et al.* 2000) et *Agrobacterium tumefaciens* (Goodner *et al.* 2001) possèdent deux chromosomes dont un, dans le cas d'*Agrobacterium* est linéaire. Chez les eucaryotes, le génome est généralement d'une taille beaucoup plus conséquente et compacté sous la forme de plusieurs chromosomes.

Cependant, on observe une extraordinaire variabilité de la taille des génomes (Figure 26) couplée à une forte hétérogénéité du nombre et de la taille des chromosomes au sein des êtres vivants connus.

En général, les plus petits génomes sont observés chez les organismes intracellulaires ou symbiotiques telles que, *Buchnera aphidicola* (Tamas *et al.* 2002) (641 Kpb), *Wigglesworthioa glossinidia* (Akman *et al.* 2002) (697 Kpb), *Encephalitozoon cuniculi* (Katinka *et al.* 2001), *Guillardia theta* (551 Kpb) (Douglas *et al.* 2001) ou *Nanoarchaeum equitans* (Waters *et al.* 2003) (500 Kpb). Leur petite taille est généralement attribuée à leur mode de vie intracellulaire.

La séquence contenue dans les plasmides ou « mégaplasmides » constitue également pour certaines espèces un matériel génétique non négligeable tant au niveau de la taille que de la fonction (Mikesell *et al.* 1983). Nous pourrions ainsi citer le cas de la bactérie *Xylella fastidiosa* (Van Sluys *et al.* 2003) dont un des deux plasmides de 51 Kpb possède 64 ORFs. De même, *Borrelia burgdorferi* (Fraser *et al.* 1997) possède, outre un chromosome linéaire (910 Kpb), sept plasmides circulaires (de 9 à 32 Kpb) et dix linéaires (de 17 à 56 Kpb) représentant au total plus de 530 Kpb d'ADN non chromosomique, soit plus d'un tiers de la quantité totale d'information génomique.

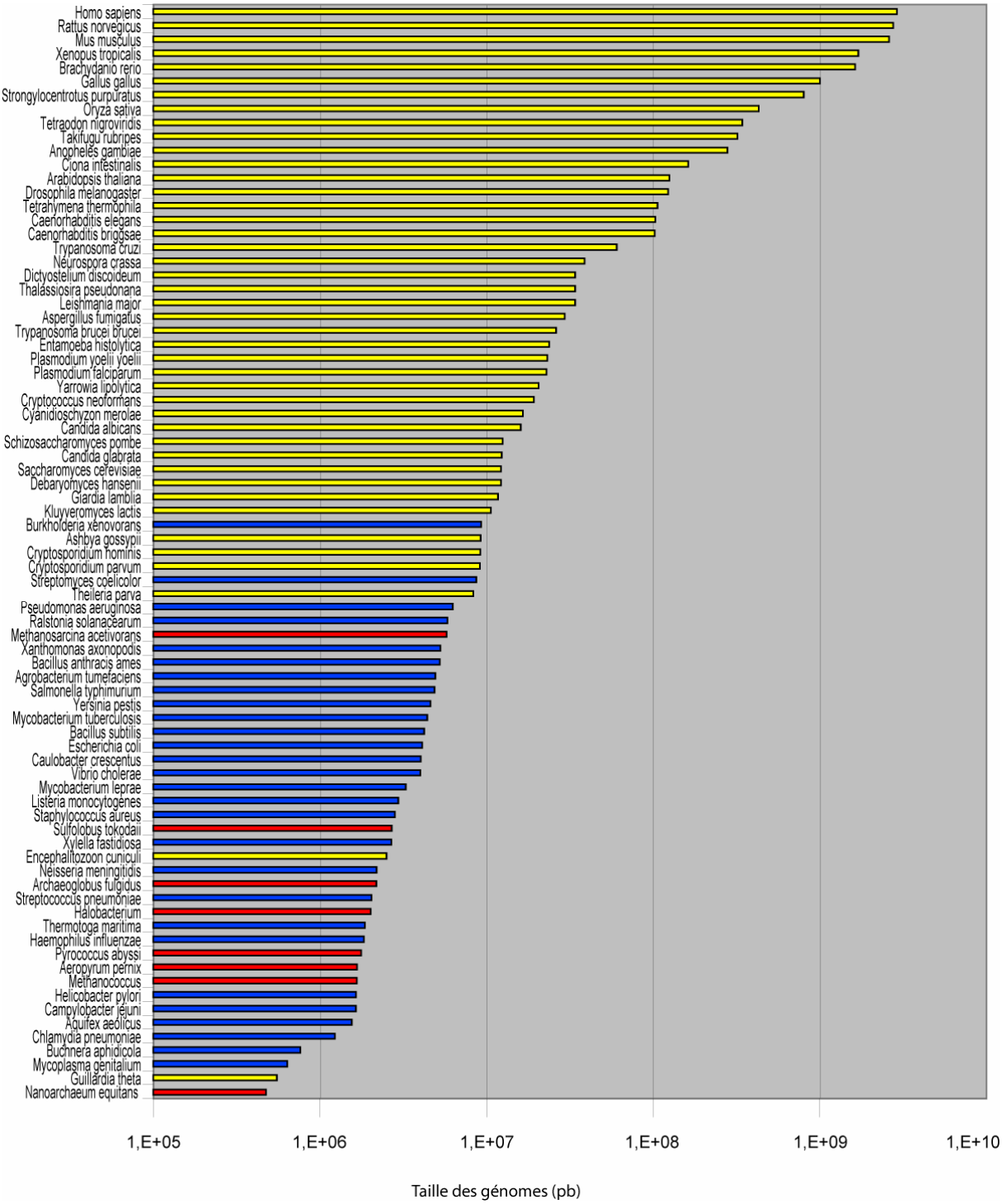


Figure 26 **Taille des génomes complets séquencés.**
L'échelle est logarithmique et regroupe les 3 domaines du vivant ; bactéries (bleu) archées (rouge) et eucaryotes (jaune).

Quelques exemples de génomes eucaryotes, *Ashbya gossypii*, *Cryptosporidium hominis*, *Cryptosporidium parvum*, *Encephalitozoon cuniculi* et *Guillardia theta* ont une taille de génome nucléaire inférieure à certain procaryotes. Néanmoins, la Figure 26 permet d'observer que de manière générale la taille des génomes eucaryotes métazoaires est considérablement plus

grande que celle des procaryotes et semble globalement corréler avec le degré de complexité des organismes.

3.1.4 Structure et nombre de gènes

Une des particularités des procaryotes est l'organisation de leurs gènes en opérons. Un opéron (Jacob *et al.* 1961) est un groupe de gènes positionnés consécutivement sur le génome, comprenant un opérateur (accepteur d'un facteur de transcription), un promoteur commun, et un ou plusieurs gènes contrôlés pour produire un seul transcrit ou ARNm polycistronique. Fréquemment, les gènes transcrits sont impliqués dans un même processus biologique et leur localisation dans un même opéron permet de coordonner simplement leur expression.

Le gène eucaryote présente une structure beaucoup plus complexe que son équivalent procaryote et est qualifié de discontinu ou mosaïque (Figure 27). A l'exception du nématode *C. elegans* (Blumenthal *et al.* 2002) ou de certains gènes impliqués dans des processus bien définis (développement, système immunitaire…), il n'existe pas ou peu de gènes organisés en opéron chez les eucaryotes. Les gènes impliqués dans les mêmes processus biologiques se trouvent en des points disparates du génome, souvent dans des chromosomes différents, ce qui implique que des mécanismes très élaborés seront mis en jeu pour obtenir l'expression coordonnée de ces gènes.

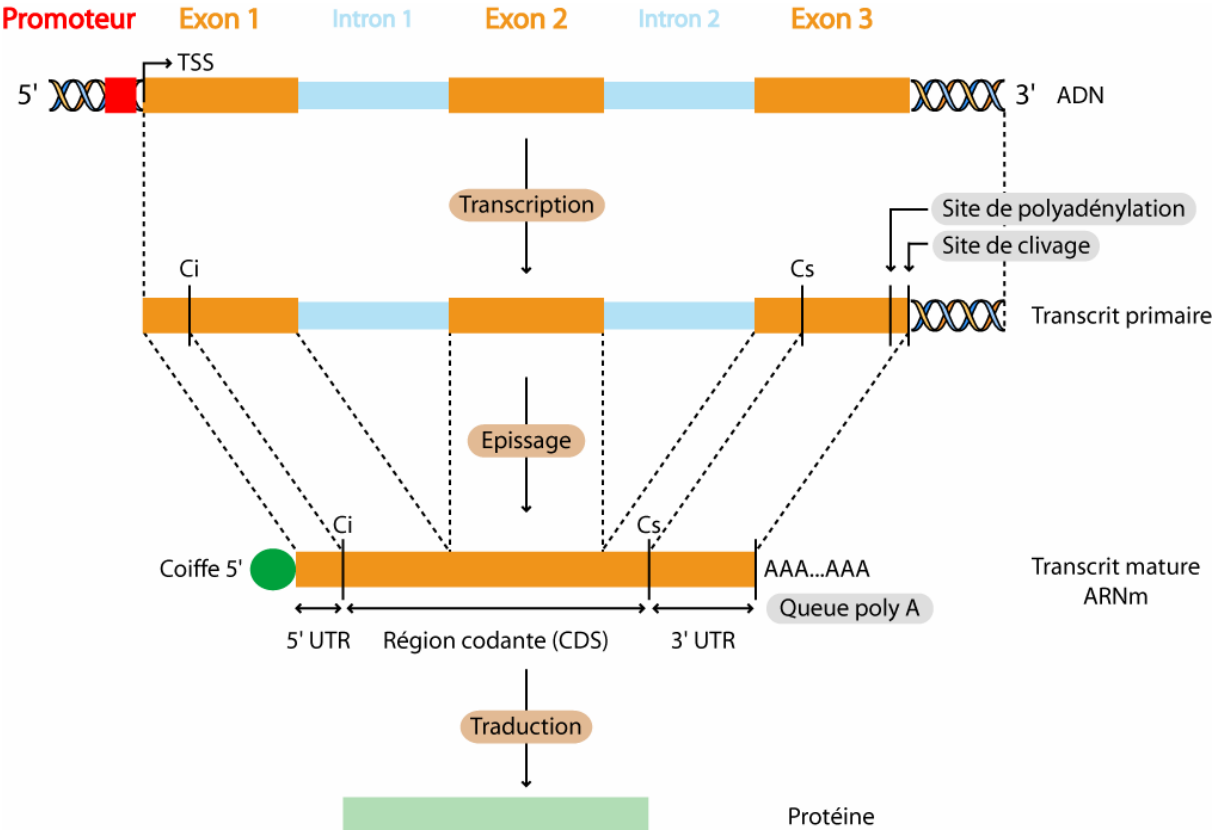


Figure 27 **Structure du gène eucaryote de l'ADN génomique à la protéine.**
Le gène est transcrit de sa forme ADN à sa forme transcrit primaire. L'épissage permet ensuite de supprimer les introns (bleu) pour obtenir le transcrit mature ou ARNm. L'ARNm, composé uniquement des exons (orange), contient la région codante et 2 zones non codantes (5' et 3' UTR). Une coiffe est ajoutée en 5' du transcrit. Le clivage (site de clivage) et l'ajout de queue poly A (site de polyadénylation) sont également réalisés en 3' du transcrit. Enfin, l'ARNm est traduit en protéine. Le codon initiateur de la traduction (Ci) et le codon Stop ou d'arrêt de la traduction (Cs) sont représentés.

La structure d'un gène eucaryote est ainsi qualifiée de discontinue car la séquence transcrite est constituée d'exons et d'introns (Figure 27). L'étape d'épissage permet de retirer les introns et de ne conserver que les exons qui contiennent la partie codante ou CDS (CDS pour *CoDing Sequence*) de la protéine. On peut noter que les exons contiennent également des régions non codantes aux extrémités 5' et 3' de l'ARNm qui sont nommées respectivement « 5' UTR » et « 3' UTR » (*UnTranslated Region*).

L'épissage alternatif consiste à manipuler, pour un gène donné, son répertoire d'exons afin d'obtenir des ARNm différents. Ce processus permet d'augmenter le nombre de protéines codées par un même gène et donc, potentiellement, de fonctions d'un gène (Figure 28). Il peut avoir lieu dans le CDS et entraîner la production de protéines différentes, mais également dans les régions 5' et 3'UTR avec comme conséquence possible, une instabilité du transcrit mature et une baisse du niveau de production de la protéine. Ce phénomène d'épissage alternatif est loin d'être une exception mais plutôt une règle générale puisque,

selon diverses estimations de 50% à 80% des gènes humains possèdent un ou plusieurs variant d'épissage (Modrek *et al.* 2002; Johnson *et al.* 2003; Kampa *et al.* 2004). Cet ensemble contribue ainsi à l'augmentation du répertoire des possibilités d'un organisme, sans avoir besoin de multiplier le nombre de gènes.

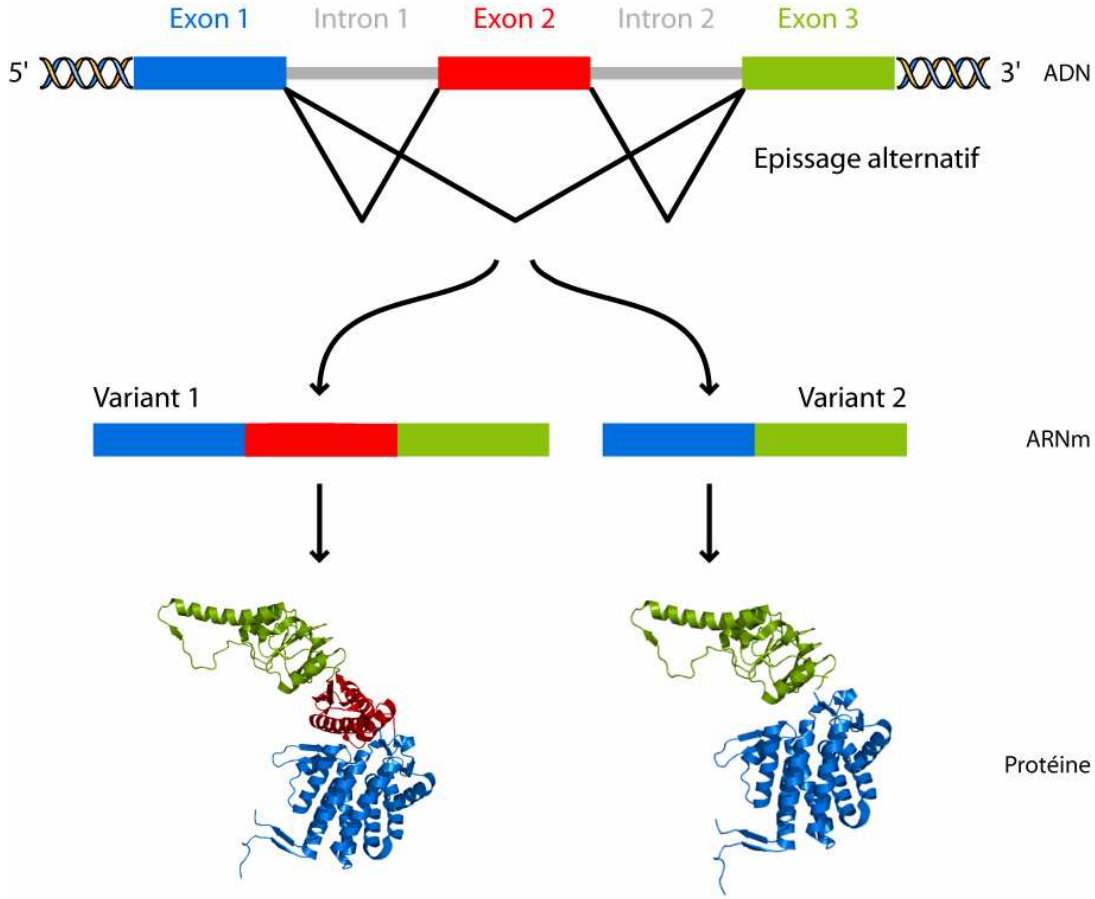


Figure 28 **Exemple de l'épissage alternatif.**
Ce mécanisme permet à partir d'un gène, d'obtenir plusieurs transcrits matures de tailles et de composition différentes. Ces ARNm seront ensuite traduits en protéines qui auront potentiellement des activités et/ou des interactions différentes. L'obtention des variants d'ARNm est réalisée par la composition alternative en exons. Dans l'exemple, le « Variant 1 » est composé des exons 1, 2 et 3 alors que le « Variant 2 » est composé uniquement des exons 1 et 3.

Tout en demeurant dans le cadre du dogme central énoncé par Crick (Crick 1958; Crick 1970), la découverte de cette structure particulière de gène ainsi que celle de l'épissage alternatif (Berget *et al.* 1977; Breathnach *et al.* 1977; Chow *et al.* 1977) ont révolutionné la compréhension de la biologie moléculaire en conférant une place plus importante à l'ARN.

L'analyse globale du nombre de gènes chez les organismes séquencés montre effectivement que les organismes eucaryotes ont en général plus de gènes que les procaryotes. Ainsi, chez les procaryotes, le nombre de gènes varie de 468 pour *Mycoplasma genitalium* à 8702 pour *Burkholderia xenovorans* alors que chez les eucaryotes, on observe de 464 pour *Guillardia theta* à 42653 gènes pour *Oryza sativa*. Il existe cependant certains paradoxes qui ne permettent pas

d'expliquer les différences d'évolution entre le nématode et l'homme ou entre l'homme et le riz sur la simple base du nombre de gènes.

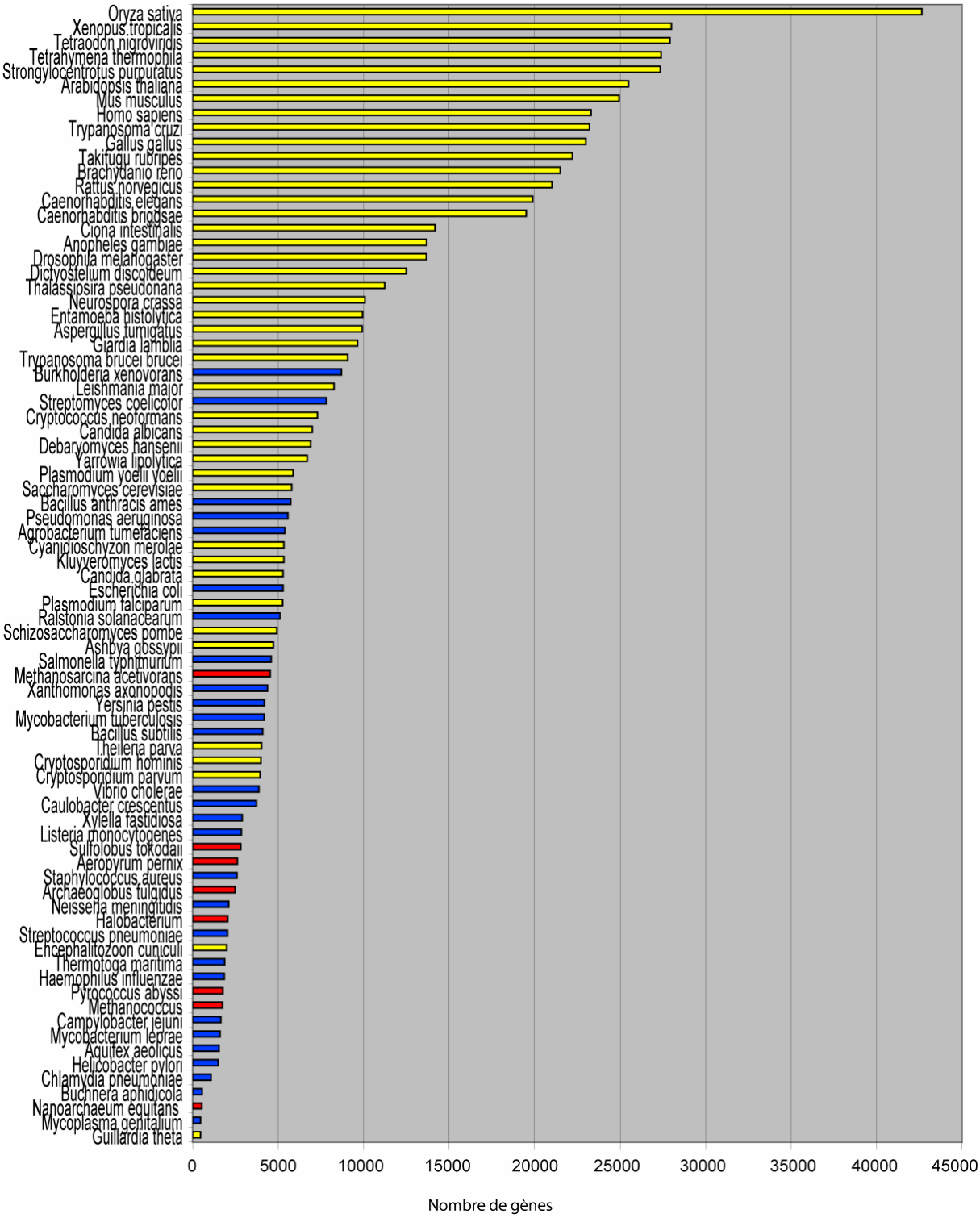


Figure 29 **Nombre de gènes estimés dans les génomes complets séquencés**
Les 3 domaines du vivant sont représentés ; bactéries (bleu) archées (rouge) et eucaryotes (jaune).

Les pseudogènes sont des copies endommagées et souvent, mais pas toujours, inactives de gènes existants. Ils peuvent s'assimiler aux fossiles de notre génome (Jacq *et al.* 1977; Vanin 1985; Mighell *et al.* 2000). Il existe deux types de pseudogènes formés par des mécanismes différents (Figure 30) :

- Les pseudogènes dit *non-processed* correspondent à des duplications partielles ou complètes de gènes. Seule une petite fraction de ces pseudogènes garde sa fonctionnalité génique (Prince *et al.* 2002). Le découpage intron/exon est ainsi conservé.
- Les pseudogènes *processed* qui sont formés par la rétrotransposition d'ARNm matures aléatoirement intégrés dans le génome. A priori, ils ne possèdent donc pas de régions promotrices ni la structure intron/exon. Cependant, quelques cas de gènes « actifs » ne possédant aucun intron et ayant une origine rétrotranspositionnelle ont été décrits dans de nombreux organismes (Brosius 1999). Ces pseudogènes sont alors appelés rétrogènes (Torrents *et al.* 2003).

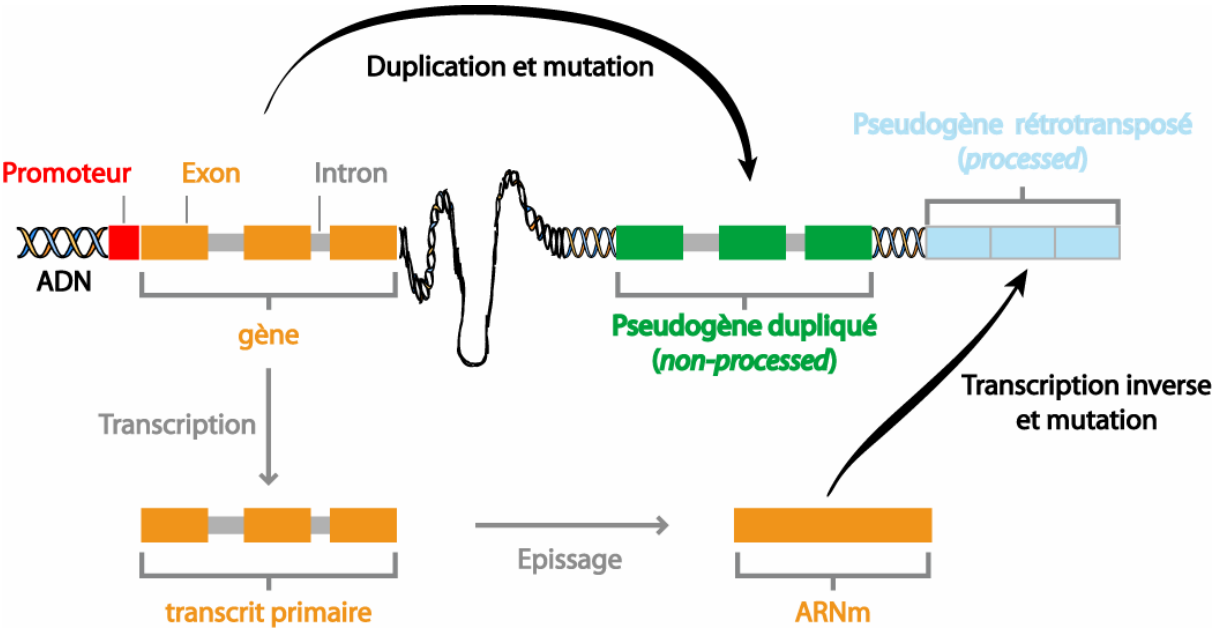


Figure 30 **Processus de fabrication d'un pseudogène.**

La distribution des pseudogènes sur le génome est pour le moment considérée comme aléatoire. L'homme dont le nombre de pseudogènes est estimé entre 14000 et 19000 avec environ 60% de pseudogènes *processed* (pour une revue (Zhang *et al.* 2004)) compte parmi les organismes qui possèdent le plus de pseudogènes. Il en existe également chez certains procaryotes mais dans une moindre mesure. Au niveau de leur distribution, il semblerait que les familles de gènes les plus exprimées dans la cellule aboutissent à plus de pseudogènes. De façon intéressante, parmi les familles de gènes qui dénombrent le plus de

pseudogènes, on observe les gènes codants pour des protéines ribosomales ou des facteurs de transcription mais également pour des familles de protéines comme les chaperonines et celles impliquées dans le cytosquelette (Zhang *et al.* 2003). En particulier, on notera l'actine béta et l'actine gamma, les kératines 18 de type I et 8 de type II qui comptent respectivement 15, 17, 61 et 31 pseudogènes (Zhang *et al.* 2003).

3.1.5 Les enseignements : ce que les génomes ont à nous dire

Le séquençage des génomes d'organismes a pour cible des organismes aussi divers qu'une algue rouge, une bactérie vivant au fin fond des océans, un vers rond ou encore la vache. Cette diversité de projets de séquençage est nécessaire pour disposer d'un panel représentatif de la diversité biologique.

La première chose qui nous frappe lors de l'étude des génomes est à la fois cette grande diversité des phénotypes des organismes décrits, des éléments qui les composent et la grande proximité des éléments de base.

L'analyse de la taille des génomes et du nombre de gènes montre une hiérarchie entre les eucaryotes et les procaryotes (Figure 26 et Figure 29). Ainsi, les génomes eucaryotes sont plus complexes et plus grands que les génomes procaryotes. Néanmoins, il existe, à la frontière entre les deux, une zone floue où certains organismes eucaryotes possèdent moins de gènes ou ont des génomes moins grands que certains procaryotes.

La différence entre ces génomes est donc en partie révélée par la taille des génomes. Cependant, si on compare le rapport entre le nombre de gènes et la taille d'un génome pour ces deux groupes, il est intéressant de noter que s'il existe une relation linéaire entre la taille d'un génome et son contenu en gènes chez les bactéries et les archées (Figure 31), cette relation n'est plus vraie chez les eucaryotes (Figure 32). Pour les eucaryotes, on distingue en particulier deux grands groupes ; à gauche les organismes unicellulaires et plus simples (les levures, les algues…) qui ont une relation quasi linéaire entre le nombre de gènes et la taille de leurs génomes, et à droite les organismes pluricellulaires et plus complexes (l'homme, les poissons, les insectes, les plantes…) qui ont, de façon évidente, perdu toute relation linéaire entre le nombre de gènes et la taille de leurs génomes.

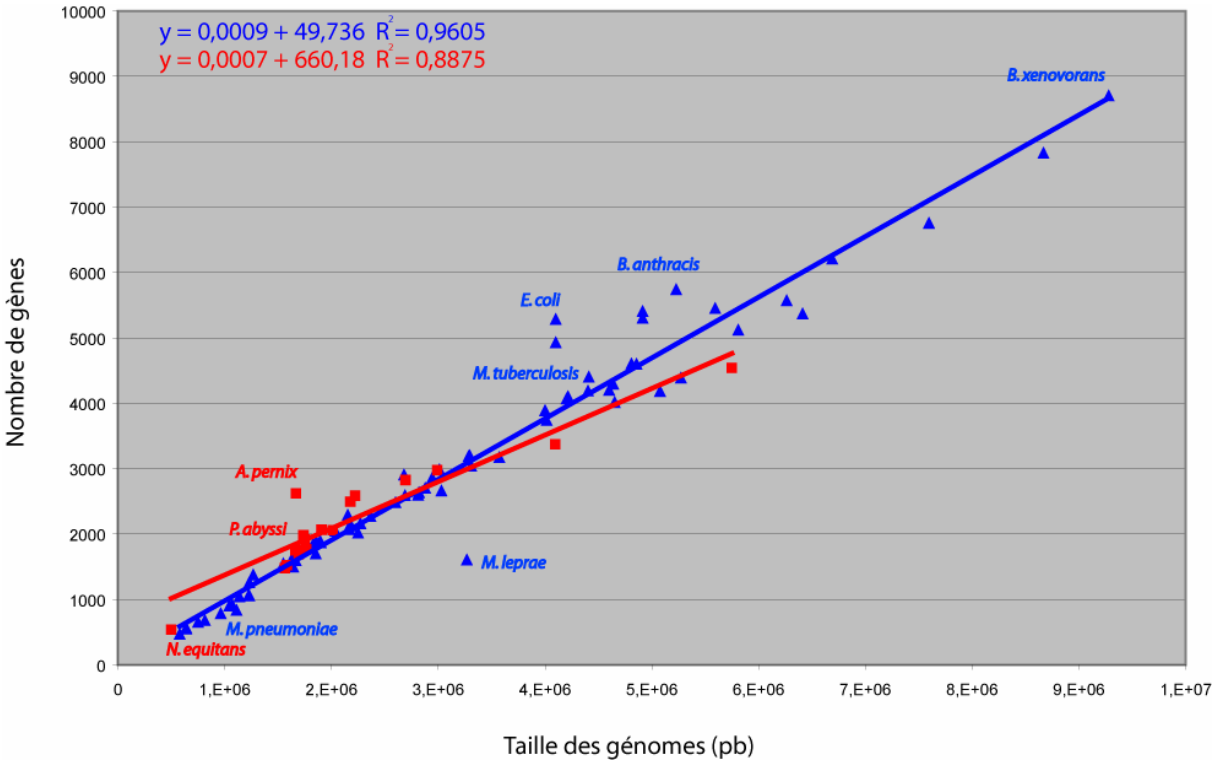


Figure 31 Nombre de gènes prédits en fonction de la taille du génome de bactéries et d'archées disponibles.

Les données de bactéries sont illustrées en bleu (triangle et traits bleus) alors que les archées sont montrées en rouge (carrés et traits rouges). Les deux droites de régression montrent la relation linéaire existante entre le nombre de gènes et la taille des génomes procaryotes.

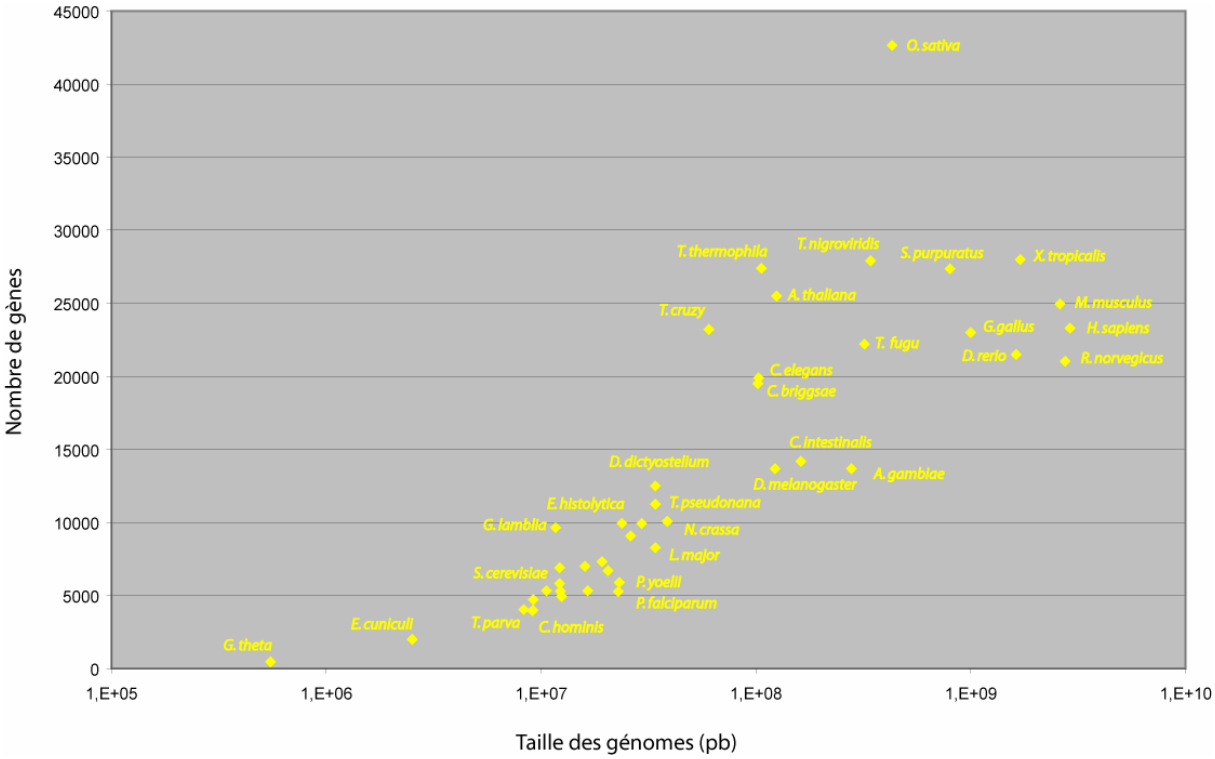


Figure 32 Nombre de gènes prédits en fonction du logarithme de la taille des génomes eucaryotes disponible en juin 2006.

En effet, le rapport entre la fraction codante et non codante des génomes eucaryotes varie dans des proportions considérables, sans commune mesure avec ce qui est observé chez les procaryotes. Les génomes procaryotes sont extrêmement compacts et possèdent peu de régions intergéniques. Dans les grands génomes, tels que le génome humain (plus de 3 Gb), la fraction codante ne représente qu'une part très réduite du génome tandis qu'elle constitue l'essentiel d'un petit génome (2,9 Mb) comme celui d'*E. cuniculi*.

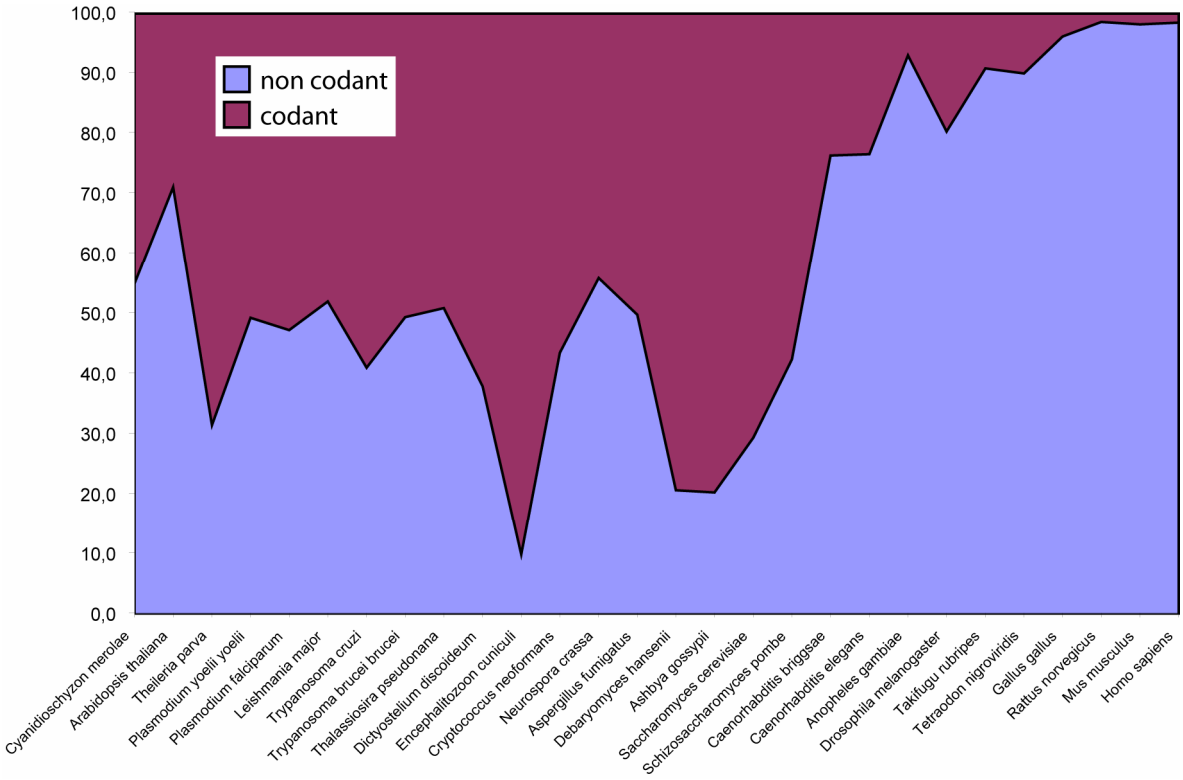


Figure 33 Relation entre l'estimation des fractions codantes et non codantes de différents génomes eucaryotes.

Cette différence entre les génomes procaryotes et eucaryotes s'explique par un certain nombre de mécanismes distincts qui ont amené à une complexification de la structure et des éléments présents dans ces génomes.

Comme décrit précédemment, la structure du gène eucaryote et l'épissage alternatif permettent certainement d'expliquer en partie ces différences, ces deux mécanismes contribuant par ailleurs à augmenter le répertoire des possibilités à partir d'un seul gène.

D'autres éléments peuvent expliquer les différences entre procaryote et eucaryotes, la présence de pseudogènes de façon plus systématique chez les eucaryotes en est notamment une illustration. L'augmentation des régions intergéniques et de la taille des introns (Berget *et al.* 1977; Chow *et al.* 1977), la présence de zones ou d'éléments répétés sont également des éléments distinctifs des organismes eucaryotes.

3.2 Les outils de la génomique comparative

La séquence complète d'un génome est une formidable source de données pour comprendre l'organisme étudié. La comparaison des éléments essentiels des génomes d'organismes des 3 domaines du vivant nous a déjà permis de faire certaines hypothèses sur leur évolution et leur complexification (voir 3.1.5 Les enseignements : ce que les génomes ont à nous dire). Cependant, si la séquence, la carte et la définition des éléments génétiques peuvent être potentiellement parfaitement décrites, leurs implications dans la génèse d'un phénotype ou dans un grand processus biologique sont loin d'être comprises. L'analyse comparative des génomes et protéomes d'organismes différents permet d'aborder différents types de questions (Hardison 2003) au travers d'un choix pertinent d'organismes (Figure 34).

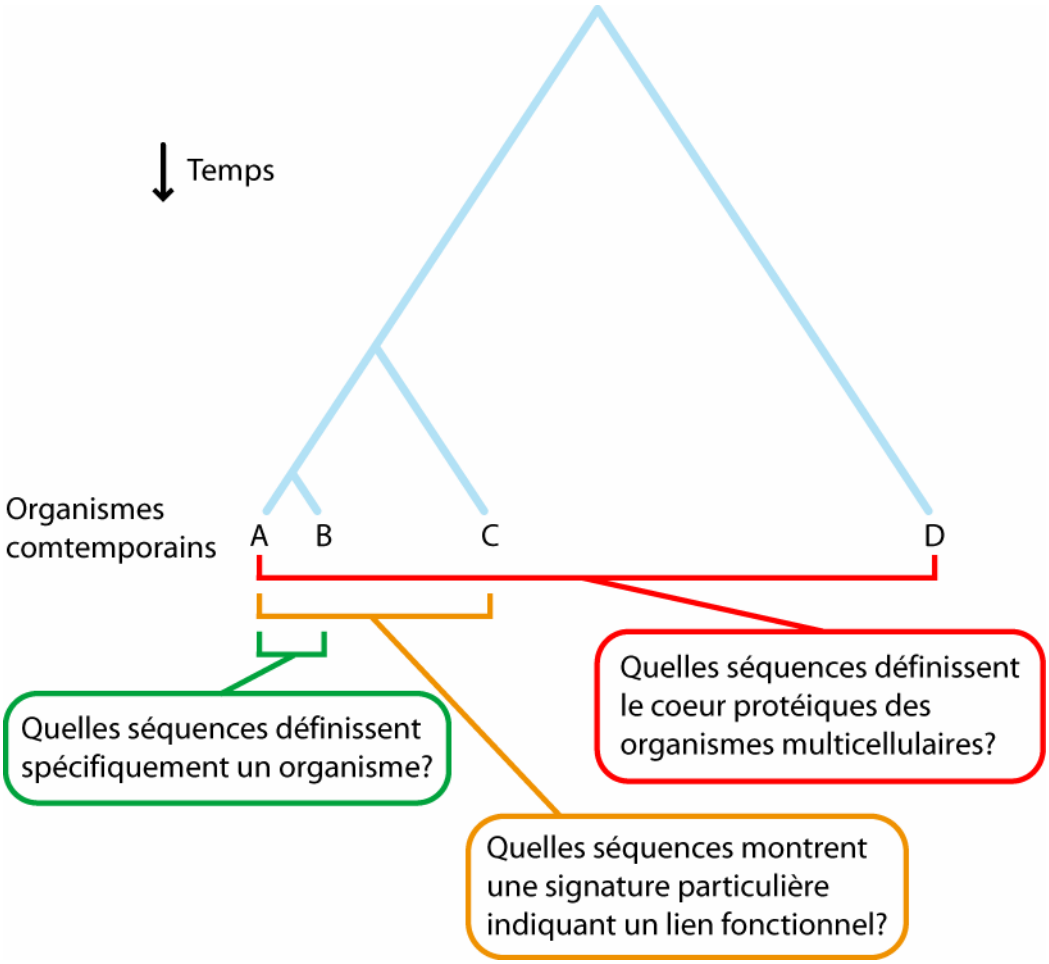


Figure 34 **La comparaison de génomes d'organismes possédant des distances phylogénétiques différentes peuvent répondre à des problèmes différents (adapté de (Hardison 2003)).**

Les outils ou méthodes développés pour répondre à ses questions sont dites de « non-homologie » car elles ne se basent pas directement sur les relations établies entre les séquences mais sur d'autres principes que nous verrons dans les chapitres suivants.

Dans ces chapitres, nous décrirons brièvement 2 méthodes essentielles dans le cadre des travaux présentés : la « soustraction de génomes » et la « distribution phylogénétique », enfin certaines méthodes de « non-homologie » seront rappelées.

3.2.1 Soustraction de génomes

La soustraction de génome est basée sur l'idée que les gènes responsables d'un phénotype particulier sont présents au sein des organismes partageant ce phénotype (qu'ils soient proches ou non !) et absents dans les organismes qui ne le présente pas. Ainsi, la différence entre l'ensemble des gènes communs aux organismes ayant le phénotype et l'ensemble des gènes communs aux organismes qui ne l'ayant pas, doit révéler un groupe de gènes enrichis en fonctions impliquées dans ce phénotype (Huynen *et al.* 1997).

Plusieurs applications ont démontrées l'intérêt de cette méthode, par exemple la comparaison des séquences codantes de *Helicobacter pylori* à celles de *Haemophilus influenzae* et *Escherichia coli* a permis de mettre en évidence une liste de 594 gènes dont 398 de fonctions inconnues et 123 impliquées dans les interactions avec l'hôte (Huynen *et al.* 1998). Considérant la capacité de survie dans un environnement gastrique de *Helicobacter pylori* par rapport aux 2 autres organismes, cette liste de gènes est enrichie en facteurs potentiellement responsables de cette adaptation. Une autre application majeure de cette méthode est l'identification de gènes responsables d'une structure cellulaire : le cil. En effet, à partir d'une question simple : « Quels sont les gènes impliqués dans le cil ? » et la disponibilité de génomes d'organismes pouvant y répondre, le choix judicieux de la comparaison d'organismes ciliés à des organismes non ciliées (Chiang *et al.* 2004; Li *et al.* 2004) a permis de déterminer non seulement les gènes potentiellement importants pour le cil, mais surtout de mettre en évidence un lien fonctionnel inédit avec une maladie génétique, le syndrome de Bardet-Biedl ou BBS. Ainsi, Li *et al.* ont identifiés 688 gènes codant souvent pour des protéines liées à la fonction ciliaire, au sein desquels ils ont pu identifier le gène BBS5. Ce résultat a été obtenu en « soustrayant » des gènes de 2 génomes ciliés, l'homme et chlamydomonas, à ceux présents dans un organisme non cilié, d'Arabidopsis thaliana (Figure 35).

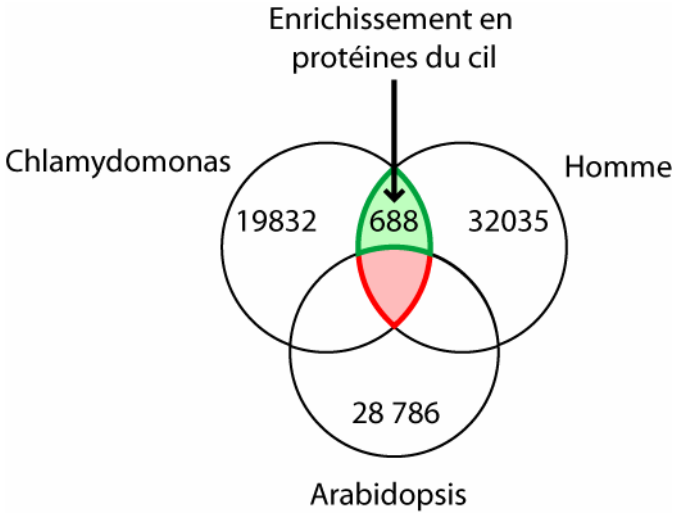


Figure 35 **Exemple de génomique soustractive.**
Diagramme de Venn montrant l'approche comparative pour la création d'un groupe de gènes enrichi en fonctions liées au cil. Les protéomes de l'Homme et de Chlamydomonas ont été comparés, donnant lieu à l'identification de 4348 protéines communes. Ces protéines ont ensuite été filtrées par le protéome de la plante Arabidopsis pour aboutir à 688 protéines spécifiques à l'Homme et à Chlamydomonas, enrichies en fonction liées au cil (adapté de Li, 2004 #244}).

Cette approche a cependant certaines limites, ainsi si l'on reprend l'exemple ayant trait aux cils, certains gènes ciliaires exclus, telles que, la tubuline, impliquée clairement dans la fonction du cil, mais conservée chez tous les organismes, ou encore, les gènes apparus au cours de l'évolution après Chlamydomonas et impliqués dans le cil y sont exclus.

3.2.2 Distribution phylogénétique

La distribution phylogénétique ou « Phylogenetic profiling » a été introduite par Pellegrini *et al.* en 1999 (Pellegrini *et al.* 1999) et correspond à un bilan de Présence/Absence des gènes dans un ensemble d'organismes (Figure 36). Cette méthode est fondée sur l'hypothèse que les protéines qui font partie de complexes structuraux et/ou qui sont impliquées dans une même voie métabolique ont une évolution semblable. Ainsi, au cours de l'évolution les partenaires d'un complexe sont co-présents dans les mêmes orgnismes. Inversement, si le complexe n'est plus utile pour un organisme alors les protéines de ce complexe doivent être co-absentes de cet organisme.

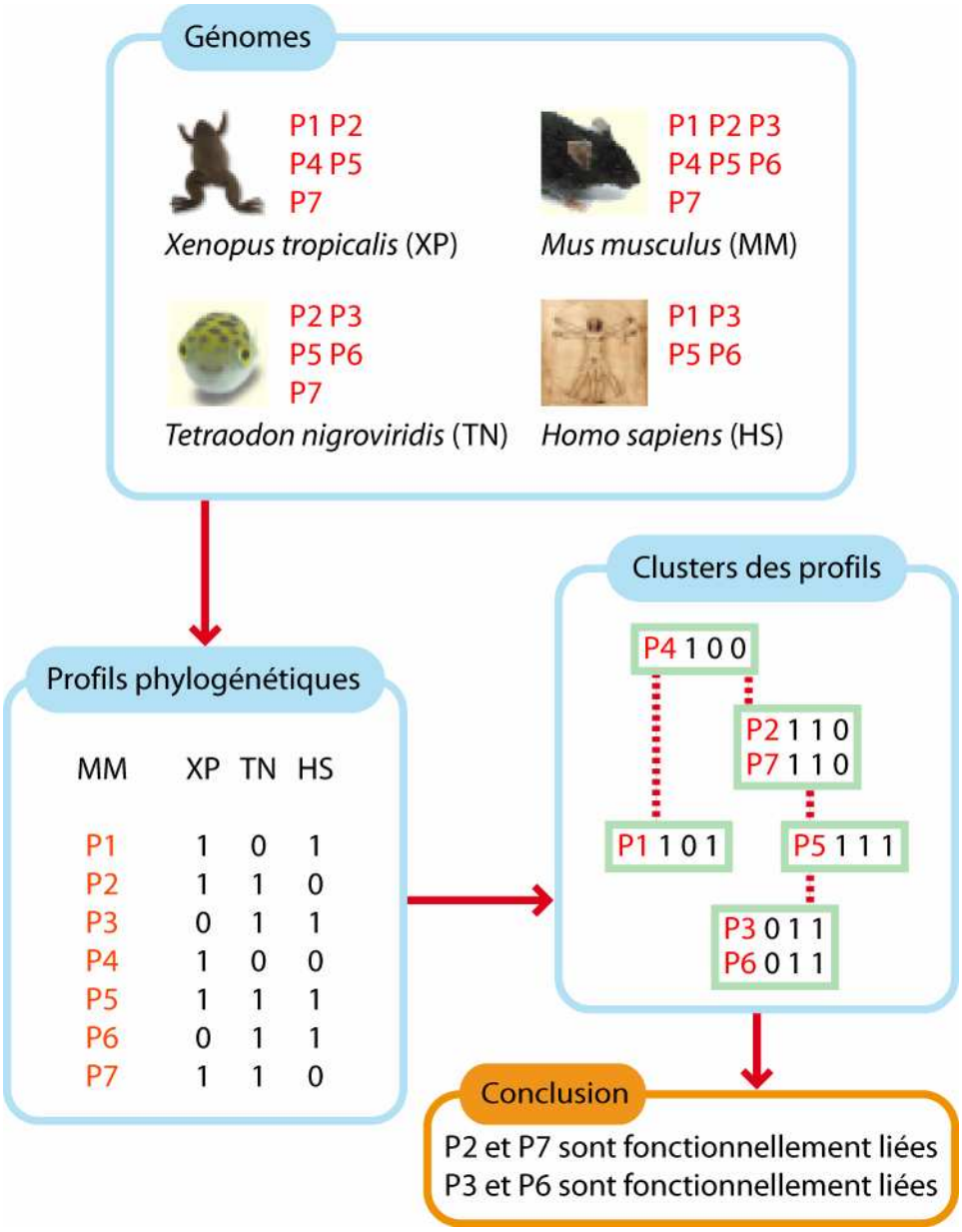


Figure 36 **Schéma représentant la méthode du profil phylogénétique.**
Cas hypothétique, à partir de 4 génomes (***X. tropicalis, M. musculus, T. nigroviridis*** et ***H. sapiens***) contenant un échantillonnage des protéines P1 à P7, on construit leur profils phylogénétiques. Le profil est caractérisé par un « 0 » pour une absence dans un génome donné et par un « 1 » pour la présence. Une étape de « clustering » permet de regrouper les profils proches. Dans ce cas, bien que ni P2 et P7, ni P3 et P6 ne partagent de similarité de séquences, elles partagent le même profil phylogénétique et sont donc fonctionnellement liées (adapté à partir de (Pellegrini *et al.* 1999)).

Une des applications intéressantes de la « Distribution Phylogénétique » est l'annotation de séquences inconnues. Ainsi, en dépit de l'absence de similarité de séquences entre des protéines participant à une même voie métabolique ou à un même complexe, le fait que ces protéines présentent le même profil de présence/absence dans un ensemble d'organismes, permet de mettre en évidence leur lien fonctionnel. Cette technique est ainsi qualifiée de méthode d'annotation par non-homologie.

La distribution phylogénétique peut également être utilisée pour mieux comprendre un groupe de gènes liés par un caractère particulier (p. ex. les gènes du cytosquelette). En effet, les profils pourront nous aider à filtrer les gènes et définir ainsi des sous-groupes de gènes présents/asbents dans certains phyla. On peut, par exemple, définir parmi ces gènes ceux qui sont communs à tous les organismes ou qui « apparaissent » au cours de l'évolution dans certains groupes d'organismes (p. ex. les levures, les vertébrés…).

3.2.3 Brièvement : autres méthodes

Un certain nombre d'autres méthodes basées indirectement sur les notions d'homologie et de recherche de similarité permettent de déterminer s'il existe un lien fonctionnel entre des protéines et d'inférer des fonctions par corrélation :

- la méthode de la « fusion de gènes » ou « pierre de Rosette » (Enright *et al.* 1999; Marcotte *et al.* 1999) est fondée sur le fait que des protéines A et B distinctes chez un organisme peuvent potentiellement interagir ensemble et être fonctionnellement reliées si elles sont retrouvées sous la forme d'un seul polypeptide AB chez un autre organisme.
- la méthode des gènes voisins (Dandekar *et al.* 1998; Overbeek *et al.* 1999) s'inspire de l'organisation des gènes procaryotes en opéron qui est souvent le reflet de l'implication de ces gènes dans un même processus, voire un même complexe. Cette méthode permet d'inférer une relation fonctionnelle entre protéines dont les gènes sont retrouvés en opéron ou au voisinage l'un de l'autre dans les génomes de plusieurs espèces. La conservation de l'ordre des éléments sur le génome est également appelée « synténie ».

Chapitre 4 - La transcriptomique

« Former un couple c'est n'être qu'un ; mais lequel ? »

Oscar Wilde

Le transcriptome se définit comme l'ensemble des gènes transcrits en ARNm dans un organisme. Un certain nombre de techniques permettent son étude tel que i) le séquençage des ESTs pour « Expressed Sequence Tags » qui correspond au séquençage massif de petits fragments d'ADNc (~500 bases), ii) le « *Differential Display* » qui est une des premières méthodes décrites d'analyse de l'expression des gènes basée sur le protocole d'amplification PCR (*Polymerase Chain Reaction*) d'ADNc, iii) la technique du SAGE (*Serial Analysis of Gene Expression*) basée sur le séquençage de étiquette ou *tags* (environ 10 bases) issus de la dégradation d'ADNc par une enzyme et enfin, iv) les puces à ADN. Ce sont les puces à ADN, qui permettent de quantifier le niveau d'expression relatif d'un grand nombre de gènes exprimés simultanément dans une cellule, auxquelles nous nous intéresserons plus particulièrement.

4.1 Historique

Les premières puces à ADN sont apparues au début des années 1990, mais leur concept date de 1987 (cité dans (Bellis *et al.* 1997)). Dérivée de méthodes classiques de la biologie moléculaire comme le *northern blot* ou le *southern blot*, la technique des puces à ADN repose sur le principe développé par Southern (Southern 1974; Southern 1975) de détection par hybridation d'acides nucléiques d'intérêt (typiquement des génomes ou des chromosomes, des ARNm ou des ADN complémentaires des transcrits d'une cellule). Ce principe stipule que deux fragments d'acides nucléiques peuvent s'associer et se dissocier de façon réversible en fonction de la température et de la concentration en sel. Les puces à ADN ont ainsi vu le jour par les efforts combinés de plusieurs technologies comme l'électronique, la robotique et la chimie (technique de dépôt, préparation des lames et greffes de sondes oligonucléotides ou synthèse *in situ*), l'analyse d'image (acquisition des données), l'informatique (stockage des données) et la bioinformatique (interprétation des données). Depuis leur apparition, les puces à ADN suscitent un intérêt considérable avec pour preuve, l'explosion du nombre de publications qui leur sont dédiées depuis 2001 (Figure 37).

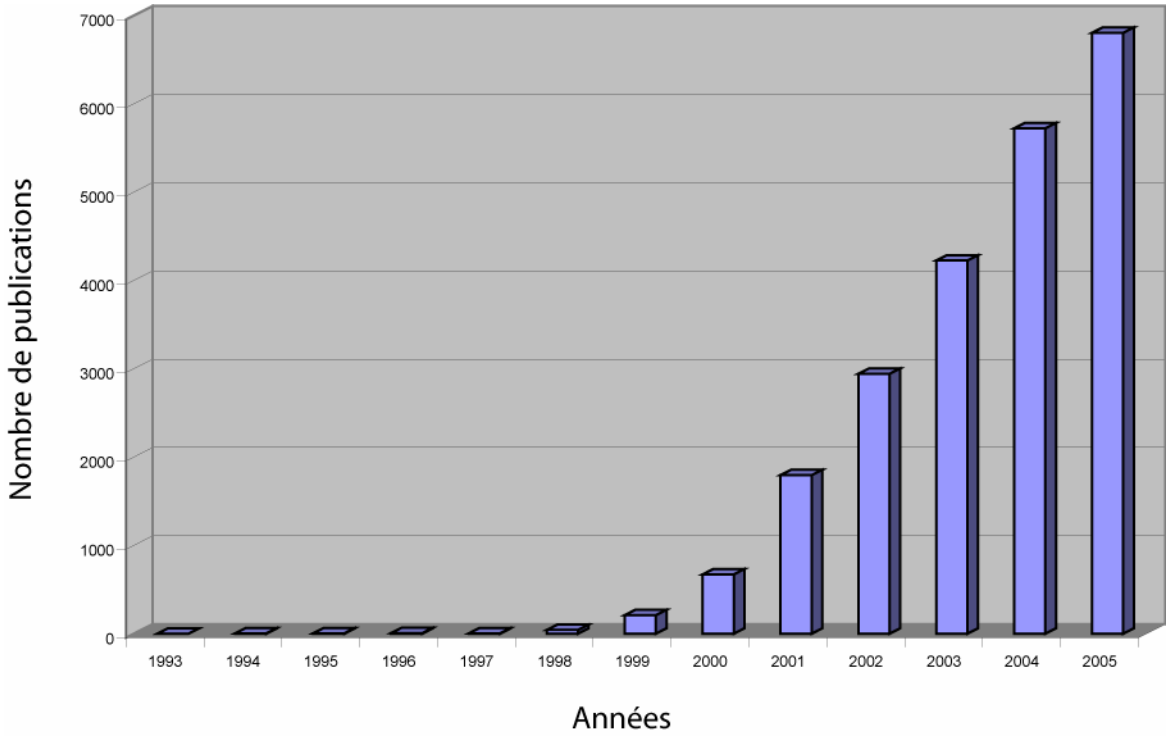


Figure 37 **Nombre de publications concernant les puces à ADN de 1993 à 2005.**
Données recueillies à partir d'une recherche PubMed utilisant les mots clés : « microarray » ou « DNA chip ».

Fondamentalement, on distingue 3 méthodes d'analyse, les *macroarrays*, les *microarrays* et les « véritables » puces à ADN. Les *macroarrays* utilisent des clones d'ADN complémentaire (ADNc) disposés sur des membranes de nylon et les cibles (ADNc à étudier) sont marquées radioactivement. Cette technique possède une densité moyenne. Les *microarrays*, appelées ainsi par opposition aux *macroarrays* car elles proposaient une densité plus importante, comportaient plusieurs milliers de gènes issus de produits de PCR, déposés et fixés sur une lame de verre et des cibles marquées par des molécules fluorescentes. Enfin, les « véritables » puces à ADN associent à un gène, un ensemble d'oligonucléotides synthétisés *in situ* et possédaient une densité élevée. Ces 3 méthodes différentes se distinguent par des supports, des sondes, des techniques de dépôt et des marquages différents. Aujourd'hui, même s'il existe toujours des clivages entre les techniques, le terme général des puces à ADN ou *microarray* est communément utilisé pour décrire n'importe qu'elle type de puce. Les différences qui les caractérisaient sont maintenant combinées en fonction des besoins et des applications. On distinguera désormais les types de puces en fonction du type de sonde, de la méthode de dépôts, des applications etc. On notera que l'on trouve dans la littérature, différents termes synonymes plus ou moins utilisés comme « *DNA chip* » ou « *DNA microarray* », et en français « biopuce » ou « micro-réseau ».

4.2 Principe des puces à ADN et applications

Le principe des puces à ADN est très simple. Il s'agit de disposer, sur un support de quelques centimètres carrés, des fragments d'ADN (les sondes) représentatifs des gènes étudiés. Ce dispositif est ensuite mis en contact avec des ARNm ou ADNc cibles, couplés à un marqueur radioactif ou fluorescent. Les conditions expérimentales (concentration en sel, température, temps d'incubation, etc) sont telles que la formation d'hybrides (phénomène d'hybridation) est possible et peut d'être quantifié grâce aux marqueurs qui leurs sont associées (voir Figure 38 et Figure 39).

On notera que bien que la méthode des puces à ADN constitue un *northern blot* inversé, il existe une confusion de termes. En effet, dans le cas du *northern blot* on appelle « sonde » l'acide nucléique libre et marqué et « cible » l'acide nucléique fixé alors que dans le cas des puces à ADN, c'est l'inverse. Ainsi, la nomenclature depuis 1999 (éditorial de janvier dans Nature Genetics par Phimister) est la suivante ; la sonde ou « probe » correspond aux acides nucléiques fixés sur la puce complémentaires aux cibles potentielles ou « targets » qui représentent l'ensemble des acides nucléiques libres étudiés.

La mesure du transcriptome, c'est-à-dire des niveaux d'expression des gènes se fait de 2 façons selon le type de puces utilisées. La mesure absolue du signal est rendue réalisable sur les puces fabriquées par *Affymetrix* qui mesure un seul échantillon par puce et qui combine une analyse statistique approfondie pour permettre la comparaison des résultats fournis par différentes puces correspondant à différents tissus, stades de développement, états physiologiques... Le protocole général de plateforme *Affymetrix* est décrit dans la Figure 38. Les puces comprennent de 10 à 20 sondes de petites tailles (25 nucléotides) par gène. De plus, à chaque sonde est associée des sondes PM (*Perfect match*) et des sondes MM (*MisMatch*) dont la séquence est identique à la sonde PM à l'exception de la base centrale qui est mutée. Cette stratégie du nombre de sondes par gènes et de la combinaison PM/MM permet de quantifier d'une part, le signal aspécifique (sonde MM) et d'autre part, d'estimer et de valider le niveau d'expression d'un gène.

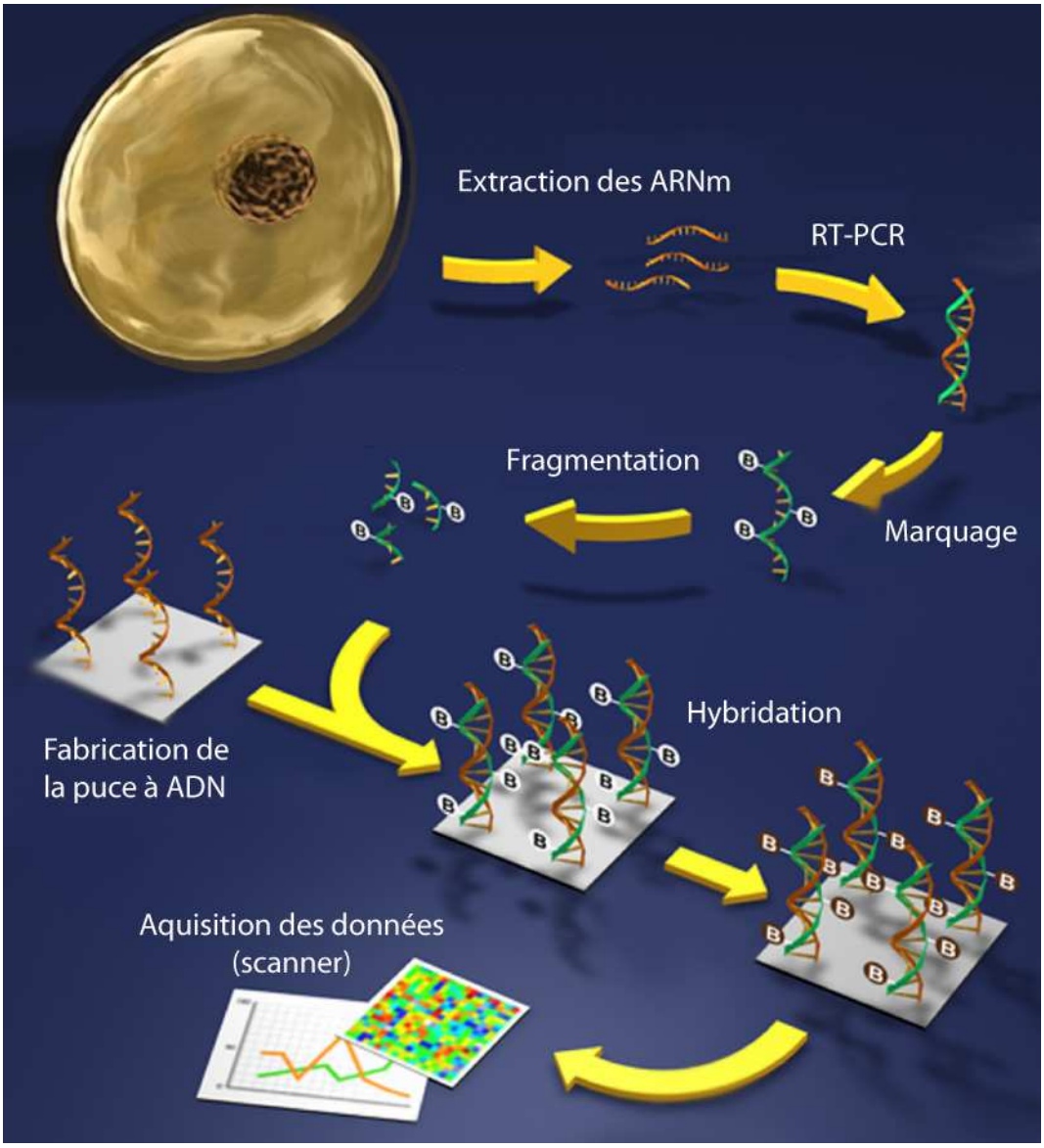


Figure 38 **Principe de la technologie des puces à ADN appliquée par *Affymetrix*.**
Au préalable, la fabrication de la puce à ADN nécessite la synthèse directe des sondes sur la puce. Les ARNm sont extraits des échantillons biologiques à comparer, les ADNc sont obtenus par transcription inverse et marqués par la biotine (B sur fond blanc), fragmentés puis mélangés sur la puce à la streptavidine couplée à fluorescéine pour l'étape d'hybridation (B sur fond rouge). La lecture des lames est réalisée par un scanner qui permet de quantifier le signal d'un seul échantillon. Les données sont ensuite analysées et interprétées et le signal est quantifié (adapté de http://www.scq.ubc.ca/?page_id=247 par Jiane Long).

Dans les cas des autres puces à ADN, la mesure du signal est relative. Ceci permet de mesurer les variations de niveau d'expression d'un organisme dans 2 conditions différentes. En effet, dans ce cas, il est possible d'hybrider 2 échantillons sur la même puce et ceci est rendu possible par l'utilisation de 2 marqueurs distincts (2 fluorochromes par exemple) (Figure 39).

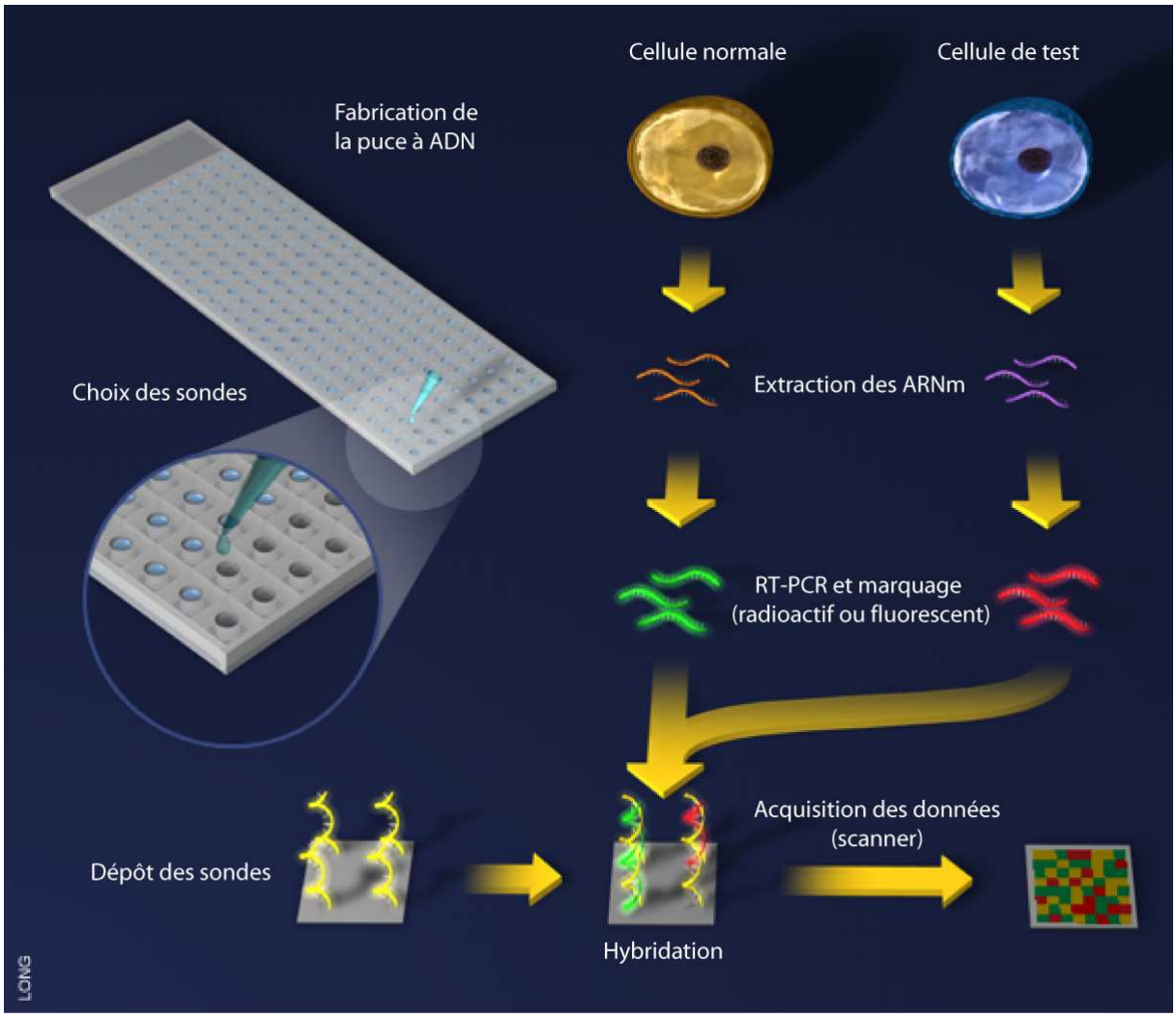


Figure 39 **Principe de la technologie des puces à ADN pour les puces à 2 canaux.**
La fabriquation de la puce à ADN nécessite le dépôt des sondes sur le support. Les ARNm sont extraits des échantillons biologiques à comparer, les ADNc sont obtenus par transcription inverse et marqués (fluorescence ou radioactivité) puis mélangés à la puce pour l'étape de d'hybridation. La lecture des lames est réalisée par un scanner qui permet de quantifier le signal des 2 marqueurs simultanément. Les données sont ensuite analysées et interprétées (adapté de http://www.scq.ubc.ca/?page_id=247 par Jiane Long).

Basés sur ces principes de nombreuses applications sont possibles et sont résumées dans le Tableau 7 (Gerhold *et al.* 1999; Walter *et al.* 2000). Ces puces sont dédiées à des organismes en particulier et permettent non seulement de mesurer les profils d'expression mais également le criblage de mutation, les réarrangements chromosomiques et les zones homozygotes communes entre patients malades.

Technique	Sonde	Cible	Application
CGH[a]	ADN génomique	ADN génomique (Fragment)	Gain ou perte chromosomique
CESH[b]	ADN génomique	ARNm (ADNc)	Expression relative fonction de la localisation chromosomique
Cartographie de zones d'homozygotie	SNP[c]	ADN génomique (Fragment)	Génotypage, polymorphisme
Profils d'expression	ADNc/oligonucléotide	ARNm (ADNc)	Expression relative des gènes
ChIP on chip[d]	ADN	Protéine	Site de fixation sur l'ADN

[a]CGH : Comparative Genomic Hybridisation
[b]CESH : Comparative Expressed Sequence Hybridisation
[c]SNP : Single Nucleotide Polymorphism
[d]ChIP on Chip : Chromatin ImmunoPrecipitation on chip

Tableau 7 **Les différentes applications des puces à ADN en biologie.**

4.3 Conception d'une puce à ADN

Basé sur un principe commun, il existe pourtant un nombre important de possibilités différentes pour concevoir une puce à ADN (Freeman *et al.* 2000; Holloway *et al.* 2002). Nous aborderons dans cette partie les éléments essentiels à la conception d'une puce à ADN. Chacun de ses composants sont des éléments cruciaux et contribue à la qualité de la puce à ADN et donc à la qualité des analyses effectuées.

4.3.1 Le support

Le support est généralement une lame de verre comme les lames de microscope et doit être parfaitement conditionné pour permettre une fixation optimale des sondes. Les qualités des lames de verre sont leur rigidité, leur surface plane et le faible bruit de fond associé à la fluorescence. Bien que les lames de verre offrent des qualités intrinsèques à même d'être utilisables pour les puces à ADN, il n'est est pas moins vrai que les sondes ne peuvent adhérer ou se fixer sur la lame sans prétraitement. Ce prétraitement chimique ou « coating » prend différentes formes selon la chimie utilisée (groupements amine, poly-Lysine, groupements époxy, groupements aldéhyde) (Microarray Handbook édité par Amersham

http://www5.amershambiosciences.com/aptrix/upp00919.nsf/content/list_all_handbooks et http://www.arrayit.com).

4.3.2 Les sondes

4.3.2.1 Les sondes spécifiques de gènes

Les sondes sont des séquences nucléiques capables de s'hybrider à d'autres séquences nucléiques. Indépendamment de leur utilisation, en dépôt ou directement synthétisées sur le support, elles peuvent être de différentes tailles en fonction de la problématique biologique, des contraintes expérimentales et des moyens disponibles dans le laboratoire. Les sondes sont un des éléments essentiels pour garantir la qualité de la puce à ADN. Une analyse bioinformatique est nécessaire pour optimiser les paramètres de choix des sondes comme la spécificité et les paramètres thermodynamiques (stabilité de l'hybride formé entre la sonde et la cible).

La société Affymetrix utilise une technologie spécifique faisant appel à des oligonucléotides courts dont la taille est de 25 nucléotides. Ces sondes sont caractérisées par une spécificité très forte mais une sensibilité faible. Affymetrix utilise le modèle du PM/MM (voir 4.2 Principe des puces à ADN et applications) et plusieurs sondes pour un même gène pour palier à ce problème. Ce type de sondes ne sera pas détaillé dans cette thèse.

Dans le cadre de l'étude du transcriptome, il existe 2 grandes familles de sondes utilisées pour la conception de puces à ADN : d'une part, les sondes obtenus par amplification d'ADNc par PCR qui ont des tailles comprises entre 200 et 2000 bases et une sensibilité maximale mais une spécificité réduite (Kothapalli et al. 2002; Zhu et al. 2005), et d'autre part, les oligonucléotides de synthèse dont la taille peut varier de 30 à 70 bases (oligonucléotides longs) et qui combinent à la fois une bonne spécificité et une bonne sensibilité.

D'un point de vue purement pratique la première approche requiert généralement la maintenance et la reproduciton à l'identique d'une banque de clones, tandis que dans la seconde approche implique la synthèse d'oligonucléotides. D'un point de vue qualitatif, si aucune différence n'est appréciable au niveau de la sensibilité entre les 2 technologies (pour des produits de PCR d'une taille <400 bases et des oligonucléotides de 50 bases) (Kane et al. 2000), au niveau de la spécificité l'utilisation d'oligonucléotide présentent plusieurs avantages (Hughes *et al.* 2001). Le premier concerne la possibilité de cibler une zone en particulier pour chacun des transcrits. Liées à la structure du gène et au phénomène d'épissage alternatif (voir 3.1.4 Structure et nombre de gènes) le choix de régions particulières pouvant détecter un ou plusieurs variants peut se révéler particulièrement utile

(Kane *et al.* 2000). Un second avantage est de pouvoir choisir des paramètres favorisant la réaction d'hybridation (Tm, GC). Ce choix peut être réalisé dans des intervalles proches, voir identiques, pour l'ensemble des sondes à définir. Ceci permet d'obtenir un comportement homogène au sein de la puce lors de l'étape d'hybridation, ce qui n'est pas le cas des sondes à base d'ADNc qui sont de tailles variables. De façon pratique, il a été montré que des oligonucléotides d'une taille de 60 bases combinent le meilleur rapport entre spécificité et sensibilité (Hughes *et al.* 2001)

4.3.2.2 Les contrôles

Un des avantages des puces à ADN est la possibilité de disposer sur le même support à la fois de sondes permettant de détecter un signal biologique et à la fois de sondes contrôles internes à l'expérience.

Les contrôles négatifs permettent de mesurer le bruit de fond d'une expérience (Selinger *et al.* 2000). Ils peuvent prendre la forme de spots vides ou de sondes issues d'un autre organisme. Ces sondes sont choisies pour ne détecter aucune des séquences de l'organisme en question, par exemple des séquences d'une plante comme *Arabidopsis thaliana* pour une puce humaine. Si un signal est détecté par ce type de sondess ceci indique généralement des conditions d'hybridation pas assez stringentes.

Les témoins positifs permettent de valider une expérience et correspondent souvent à des gènes exprimés constitutivement par tous les types cellulaires (gènes de ménage). Un manque de signal de la part de ces contrôles indique des problèmes lors de l'hybridation, une étape de lavage trop forte ou un problème de marquage. Enfin, des contrôles appelés « *Spike controls* » permettent de mesurer l'étendue dynamique de la puce. Les sondes utilisées sont souvent des séquences d'un autre organisme. Les séquences complémentaires aux sondes sont synthétisées puis ajoutées à des concentrations différentes pendant l'étape de marquage à l'échantillon biologique (Wang *et al.* 2003).

4.3.3 Le positionnement des sondes sur le support

Il existe 2 approches différentes pour effectuer la fixation des sondes sur la puce à ADN. On distingue ainsi les techniques par synthèse directement sur le support (synthèse *in situ*) et la fixation par dépôt de sondes déjà synthétisées. Dans les deux cas cependant, il est essentiel de déposer les sondes en excès par rapport à la concentration en cibles utilisée pour assurer une détection optimale et éviter toute saturation du signal.

4.3.3.1 La synthèse in situ

La méthode dit de « synthèse *in situ* » permet de synthétiser directement sur le support les sondes. Il existe 2 techniques différentes :

- La première technique mise en œuvre est un processus d'impression calqué du procédé utilisé par les imprimantes à jet d'encre. Elle repose sur la propulsion de très petites sphères de fluide dont le volume est inférieur au nanolitre (Hughes *et al.* 2001).
- La seconde technique appelée photolithographie (Lipshutz *et al.* 1999) a été développée par la société *Affymetrix*. Elle procède par dépôts de couches successives des quatre nucléotides sur le support. Le support et les nucléotides ajoutés sont couplés à des groupements photolabiles. Un masque, dont la configuration varie pour chaque couche déposée, assure l'activation par la lumière et la succession correcte des bases (Figure 40).

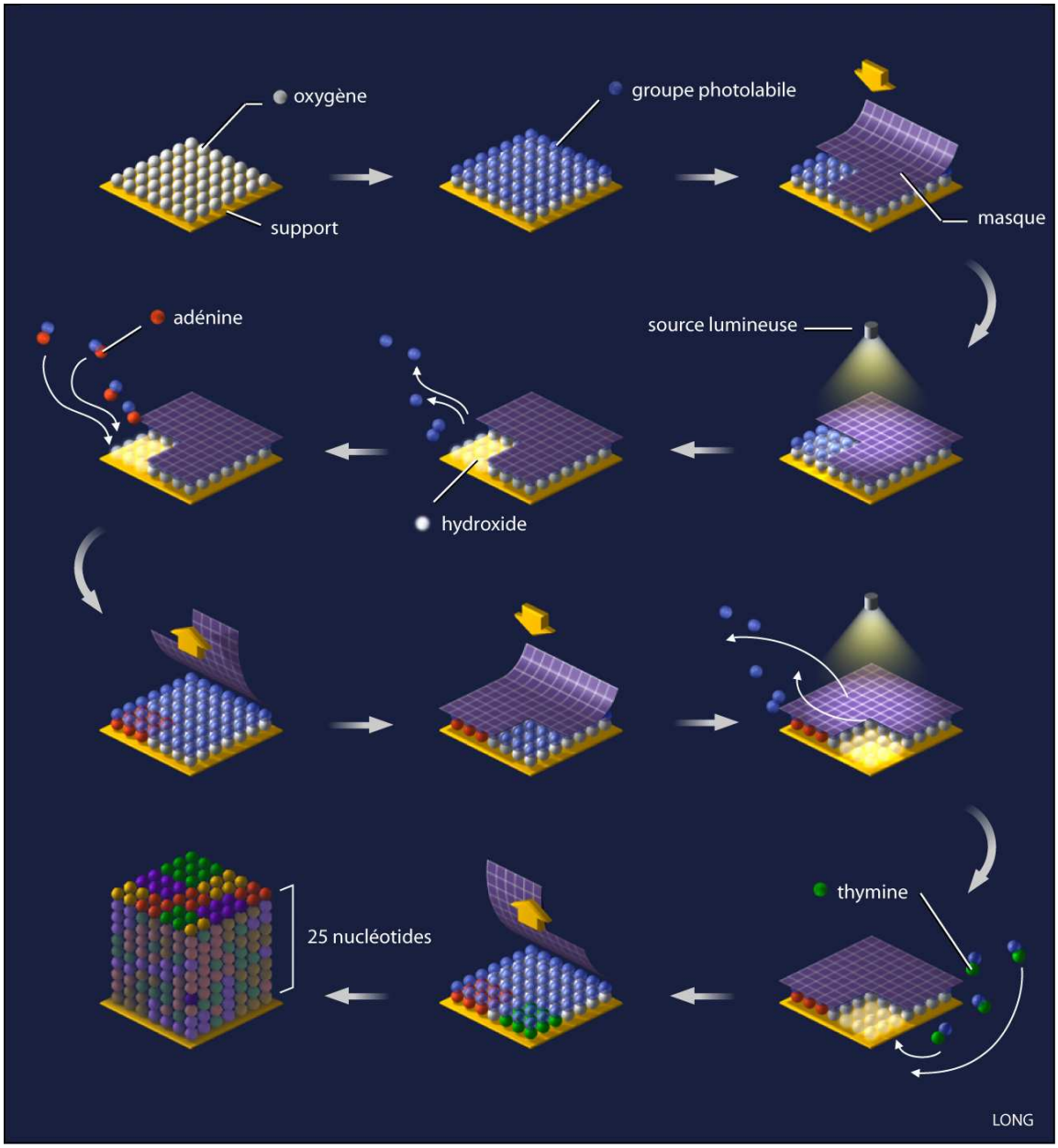


Figure 40 Schéma représentant le processus de photolithographie
La synthèse progressive des sondes est obtenue par l'application successive d'un masque, d'une base et d'une source lumineuse (source http://www.scq.ubc.ca/?page_id=247 par Jiane Long).

4.3.3.2 Le dépôt de sondes synthétisées

La méthode de fabrication des puces imprimées a été développée par l'équipe de P. Brown à l'université de Stanford aux Etats-Unis (DeRisi *et al.* 1996; DeRisi *et al.* 1997). Cette méthode bien établie aujourd'hui permet de déposer sur une lame de verre contenant des sondes oligonucléotidiques synthétiques (30 à 60 bases), des produits de PCR ou de banque d'ADNc. L'utilisation d'aiguilles permet de cibler des points particuliers du support (spot) et de définir un plan de dépôt précis. Bien que les plans originaux du modèle expérimental défini par J. DeRisi soient toujours disponibles sur internet

(http://cmgm.stanford.edu/pbrown/mguide/index.html), bon nombre de compagnies privées ont développé des modèles de robots de dépôt ou « spotter ».

4.3.4 Le marquage

L'étape de marquage est une étape importante qui permet d'incorporer aux sondes un composé radioactif ou fluorescent qui peut par la suite être détecté. En général, ce sont des composants fluorescents qui sont utilisés, comme les cyanines Cy3 et Cy5, qui émettent après excitation à une longueur d'onde spécifique, un signal fluorescent respectivement dans le vert ou dans le rouge.

Les protocoles classiques utilisent la séquence poly-A présente à l'extrémité 3' des ARNm pour ancrer une amorce poly-T permettant la synthèse d'ADNc par la reverse transcriptase. Cette première étape est essentielle pour l'étape de marquage mais constitue, pour le choix des sondes, une étape non moins importante. Effectivement, du fait du rendement des enzymes utilisées, la taille des ADNc ne dépassera que rarement les 3000 bases. Ceci restreint donc les zones de choix d'une sonde aux 3000 bases en 3' du transcrit initial.

L'étape de marquage à proprement dite peut avoir lieu en même temps que la reverse transcription, on parle alors de marquage direct avec incorporation de nucléotides portant les fluorochromes. Le marquage indirect quant à lui implique d'abord l'incorporation d'un nucléotide modifié puis une étape de fixation de la molécule fluorescente.

On peut également marquer directement les ARNm ou ARN total. Enfin, *Affymetrix* utilise le système d'affinité biotine-streptavidine, combinant un premier marquage des ADNc par la biotine puis le dépôt, après l'étape d'hybridation, de la streptavidine couplée à la fluorescéine (composé fluorescent), pour mesurer la fluorescence du signal (Figure 38).

4.4 De l'intérêt d'une puce dédiée

Selon la finalité scientifique, différentes approches sont envisageables pour l'obtention des données d'expression par la technique des puces à ADN. Elles reposent sur l'utilisation de différents types de puces : les puces pangénomiques (puce globale ou génome complet) ou les puces « dédiées » (puce thématiques). Sans rentrer dans les arguments et contre-arguments qui traversent la communauté scientifique à propos des avantages et inconvénients de chaque type de puces, on peut citer quelques éléments du débat.

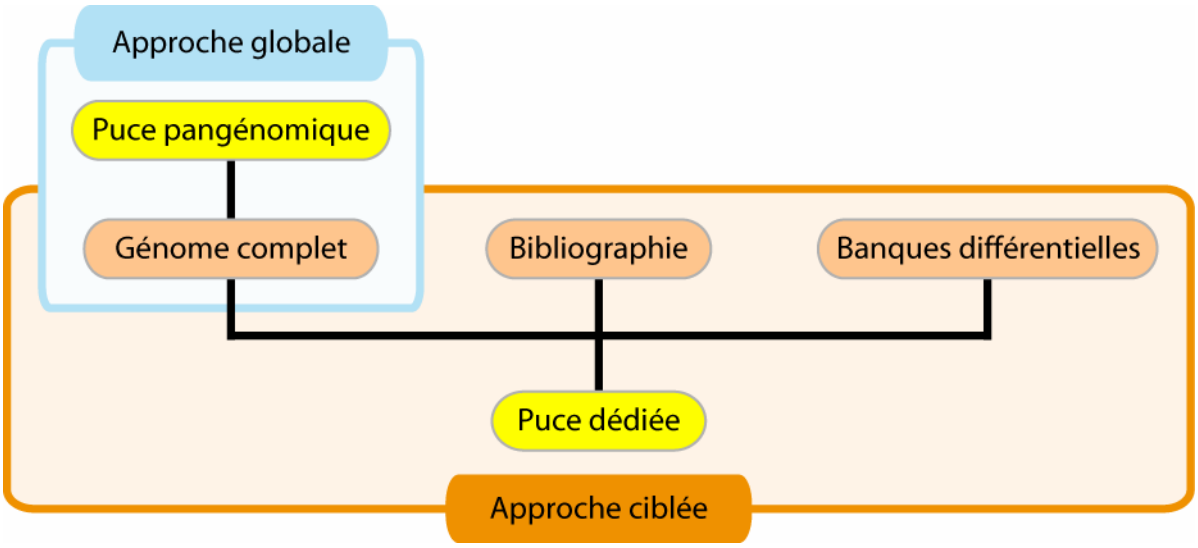


Figure 41 **Approche globale versus approche ciblée.**
Les puces pangénomiques proposent l'analyse d'un génome (ou de ces produits) complet. Les puces « dédiées » ciblent un sous-ensemble en particulier. Les gènes d'intérêt des puces « dédiées » ont été identifiés par l'analyse de la littérature, l'analyse des résultats de puces pangénomiques et/ou la construction de banques différentielles.

Les puces pangénomiques dites « généralistes » proposent des jeux de sondes oligonucléotidiques représentant l'ensemble des gènes d'un génome. Ainsi, il existe des puces à ADN contenant plus de 30000 sondes représentant les quelques 25000 gènes du génome humain. L'avantage des puces pangénomiques réside dans leur exhaustivité. Elles sont utiles pour une analyse globale des génomes et des classifications à haut débit par hybridations systématiques. En revanche, sur ce type de puces, chaque sonde n'est généralement déposée qu'une seule fois sur le support (limite essentiellement technologique). Aucune validation statistique objective des mesures ne peut donc être réalisée à l'intérieur de la puce, ce qui altère sensiblement la validité des résultats obtenus. La réalisation d'une puce dédiée permet par contre de choisir plus précisément plusieurs sondes spécifiques pour chaque gène et de modifier la composition à façon.

Les puces dédiées sont constituées d'une collection de gènes spécifiquement (voire exclusivement) liés à un tissu, une pathologie et/ou une thématique. Elles permettent de mieux cibler les transcrits pertinents pour l'étude envisagée. En effet, parmi les 10000 à 20000 transcrits potentiellement exprimés dans une cellule spécialisée, seuls 4000 à 6000 d'entre eux sont souvent caractéristiques de ce type cellulaire. Les transcrits d'intérêt peuvent être obtenus de différentes manières, souvent complémentaires :

- sélection expérimentale par criblage (screening) de puces à ADN pangénomiques.
- connaissances biologiques a priori, provenant de l'analyse de la littérature et/ou d'interrogations des bases de données publiques.

- constitution de banques différentielles par approches soustractives (Diatchenko *et al.* 1996), séquençage systématique et analyse de banques d'EST ou de banques SAGE.

Les premières générations de puces à ADN dédiées étaient issues des institutions académiques, les sociétés commerciales n'étant pas convaincues dans un premier temps, par une telle approche. Les sondes sont alors principalement des produits de PCR ou d'oligonucléotides longs ; à titre d'exemple on peut citer CardioChip (Barrans *et al.* 2001), MitoChip (Maitra *et al.* 2004) ou encore ApoChip (Magnusson *et al.* 2005).

Finalement, l'intérêt de ce type de puces réside dans la possibilité de proposer des puces à ADN de haute qualité avec un soin particulier pour le choix des gènes et des sondes déposés sur les lames. La taille réduite de ce type de puce simplifie leur analyse et permettent à terme une analyse automatique de leurs résultats. Ces avantages sont généralement compatibles avec des applications médicales, notamment pour le diagnostique de maladies. Néanmoins, ce type d'applications intégrées nécessite une connaissance approfondie des gènes choisis.

Matériel et méthodes

Chapitre 5 - Ressources informatiques et bioinformatiques

Ma thèse s'est partagée entre Strasbourg et Luxembourg et le soutien respectif de la plateforme de bioinformatique de Strasbourg (http://bips.u-strasbg.fr) et la plateforme de bioinformatique de Luxembourg (http://www.bioinformatics.lu).

5.1 Equipement informatique et réseau

L'équipement informatique et son bon fonctionnement ont une part importante dans le déroulement d'une thèse. Les 2 paragraphes suivants décrivent l'environnement informatique dans lequel s'est déroulé l'ensemble de mes travaux de recherche. Le nom des machines est indiqué en italique.

5.1.1 IGBMC

Nous disposons à l'IGBMC de plusieurs serveurs dédiés à la bioinformatique et aux calculs intensifs. Le serveur web est installé sur *titus* ; *beaufort* et *star* constituent les serveurs de calculs.

Titus est un serveur SUN E450 quadri-processeurs UltraSPARC II cadencés à 400 MHz de 1 Go de mémoire vive. L'architecture est de type 64 bits et dispose d'une installation Solaris 9.

Beaufort est un cluster de six machines Compaq quadri-processeurs Alpha ev67 cadencés à 667 MHz disposant de 16 Go (*ouragan*) ou de 4 Go (*blizzard*, *cyclone*, *tempête*, *tornade* et *trombe*) de mémoire vive reliée par Memory Channel. L'architecture est de type 64 bits et dispose d'une installation Tru64 UNIX en version 5.1.

Notre dernière « étoile », le serveur *star* est composé d'un cluster de 6 machines SUN V40Z quadri-processeurs Optéron cadencés à 2,6 GHz et dispose de 32 Go (*star1-2*) ou 16 Go (*star4-6*) de mémoire vive. L'architecture est de type 64 bits et dispose d'une installation Solaris 10 (*star1-2*) ou Linux (*star4-6*).

L'ensemble des machines dispose de 5 To d'espace disque pour stocker les données.

5.1.2 CRP-Santé

Nous disposons au CRP-Santé d'un serveur dédié à la bioinformatique (*CRPBIO*) et d'un serveur web en cours d'installation.

Le *CRPBIO* est un serveur SUN V880 quadri-processeur UltraSPARC III cadencés à 900 MHz de 4 Go de mémoire vive et de 712 Go d'espace disque pour le stockage des données. L'architecture est de type 64 bits et dispose d'une installation Solaris 8.

L'accès au *CRPBIO* de l'extérieur, par exemple de l'IGBMC, et l'exécution de programmes sont hautement optimisés par l'utilisation d'un serveur d'application citrix qui autorise un transfert optimal des fenêtres graphiques et une souplesse de travail très appréciable.

5.2 Les banques de données biologiques

Plusieurs banques, généralistes et à valeur ajoutée, sont installées et mises à jour de manière automatique sur les serveurs de l'IGBMC et du CRP-Santé. Ces banques sont disponibles au format GCG (voir ci-dessous), au format BLAST et sont également interrogeables grâce au système SRS (*Sequence Retrieval Software*) (voir 2.3.1 La recherche textuelle**)** qui permet des requêtes portant sur les champs indexés pour chaque entrée des différentes banques ainsi que des requêtes croisées exploitant les liens existant entre les différentes banques. Ces requêtes peuvent être faites de manière interactive (http://bips.u-strasbg.fr/srs) ou sous forme de commandes sous UNIX.

5.2.1 Les banques généralistes

Au cours de notre travail, nous avons utilisé de manière routinière les banques généralistes installées localement sur les serveurs du laboratoire :

- La banque de séquences nucléiques GenBank qui regroupe l'ensemble des séquences nucléiques déposées par la communauté scientifique internationale.
- La banque de séquences protéiques Swiss-Prot et son complément TrEMBL, récemment réunis en une banque unique UniProt. L'ensemble des séquences de ces banques constitue une banque protéique peu redondante.

5.2.2 Les banques à valeur ajoutée

Nous avons également utilisé certaines banques plus spécialisées telles :

- La banque de structures tridimensionnelles PDB (*Protein Data Bank*) (Kouranov *et al.* 2006), collectant les structures déterminées expérimentalement par rayons X ou RMN, est installée localement sur nos serveurs.
- La banque de séquences nucléiques et protéiques RefSeq qui disposent de l'ensemble non redondant des séquences annotées de référence pour chaque organisme. Nous avons notamment utilisé la banque contenant les ARNm humains pour le design de sondes pour la puce Actichip (Chapitre 14 - Actichip une puce dédiée au cytosquelette).
- La banque UniGene qui représente une vue organisée du transcriptome a été utilisée pour le design de sondes spécifiques pour la puce à ADN Actichip (Chapitre 14 - Actichip une puce dédiée au cytosquelette). Elle est constituée de 2 banques : UniSeq qui contient les séquences consensus identifiées et UniGene proprement dite qui contient les clusters calculés.

A ces banques, il faut également ajouter toutes les banques de séquences des génomes complets installées localement et mises à jour par la plateforme de bioinformatique de Strasbourg (http://bips.u-strasbg.fr)

Certaines analyses ont nécessité l'utilisation de banques telles, que la banque de connaissances Gene Ontology (http://www.geneontology.org) (Ashburner *et al.* 2000).

5.2.3 Interrogation des banques

La plupart des banques généralistes et à valeur ajoutée décrites ci-dessus sont installées et mises à jour sur les serveurs de l'IGBMC. Ces banques sont disponibles au format GCG (voir chapitre 5.3.1) et blast pour la plupart et sont également interrogeables grâce au système SRS (en version 7.1.3.2) qui permet (voir 2.3.1 La recherche textuelle) des requêtes portant sur les champs indexés pour chaque entrée des différentes banques ainsi que des requêtes croisées exploitant les liens existant entre les différentes banques. Ces requêtes peuvent être faites de manière interactive sur le serveur web de l'IGBMC (http://bips.u-strasbg.fr/srs) (Figure 42) ou sous forme de commandes sous UNIX.

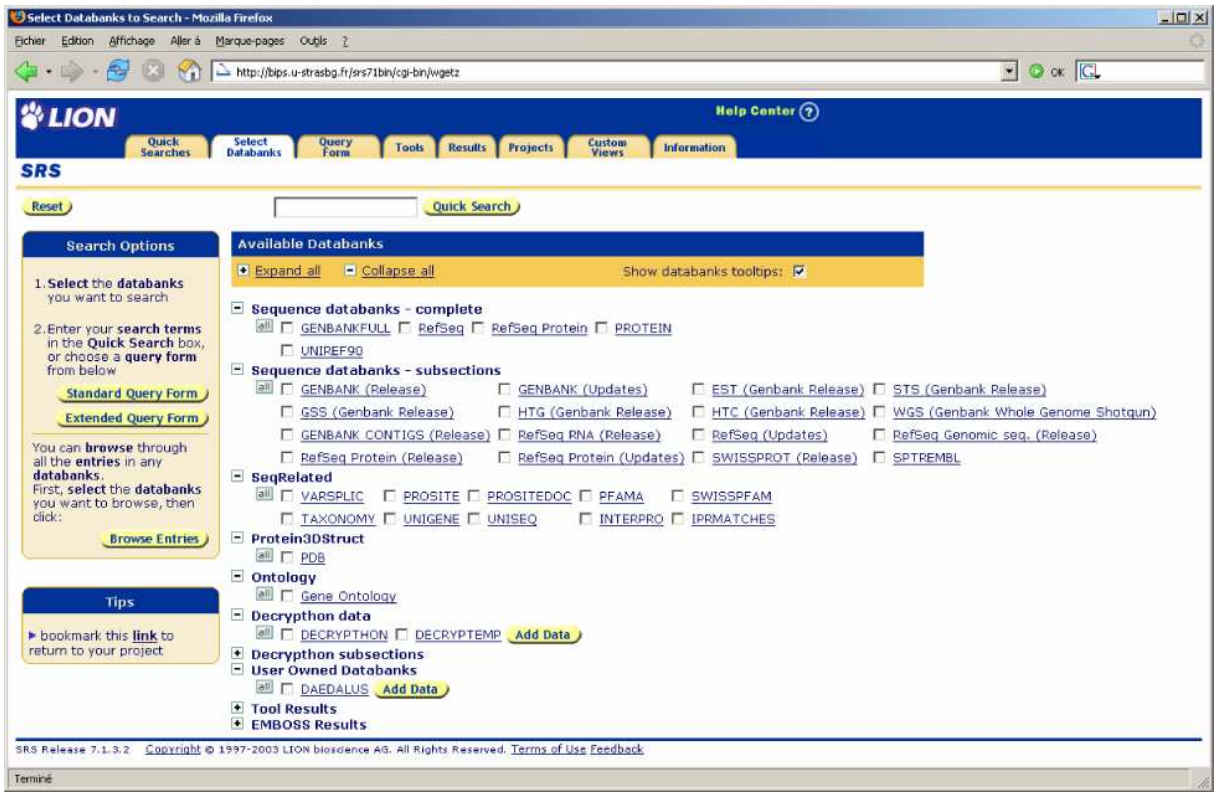


Figure 42 **Capture d'écran de la page d'accueil sur serveur web SRS de l'IGBMC.**

5.2.4 Les explorateurs de génomes ou « genome browser »

La recherche de protéines ou leur prédiction pour un organisme par la recherche de similarité aboutissent à des positions ou localisations sur leurs génomes. Afin de valider ce type de recherches, nous avons parfois besoin d'outils et d'information complémentaires à la simple séquence du ou des génomes. En effet, ce positionnement inscrit nos résultats de recherche de protéines dans un contexte, c'est-à-dire les éléments biologiques (gènes, …). Dans ce cadre, les grands centres de bioinformatique mettent à disposition pour un certain nombre de génomes des « genome browser » qui correspondent à des outils intégrés contenant, en plus de la séquence nucléique, le positionnement de nombreuses données ou résultats de prédictions (par exemple les cytobandes, la présence de mutations, les prédictions Genscan, les ARNm ou encore les ESTs). Nous avons régulièrement utilisé les « genome browser » de l'UCSC (*University of Santa Cruz*, http://genome.ucsc.edu), d'Ensembl (http://www.ensembl.org) et parfois, celui du NCBI (*National Center for Biotechnology Information*, http://www.ncbi.nlm.nih.gov) (Figure 43).

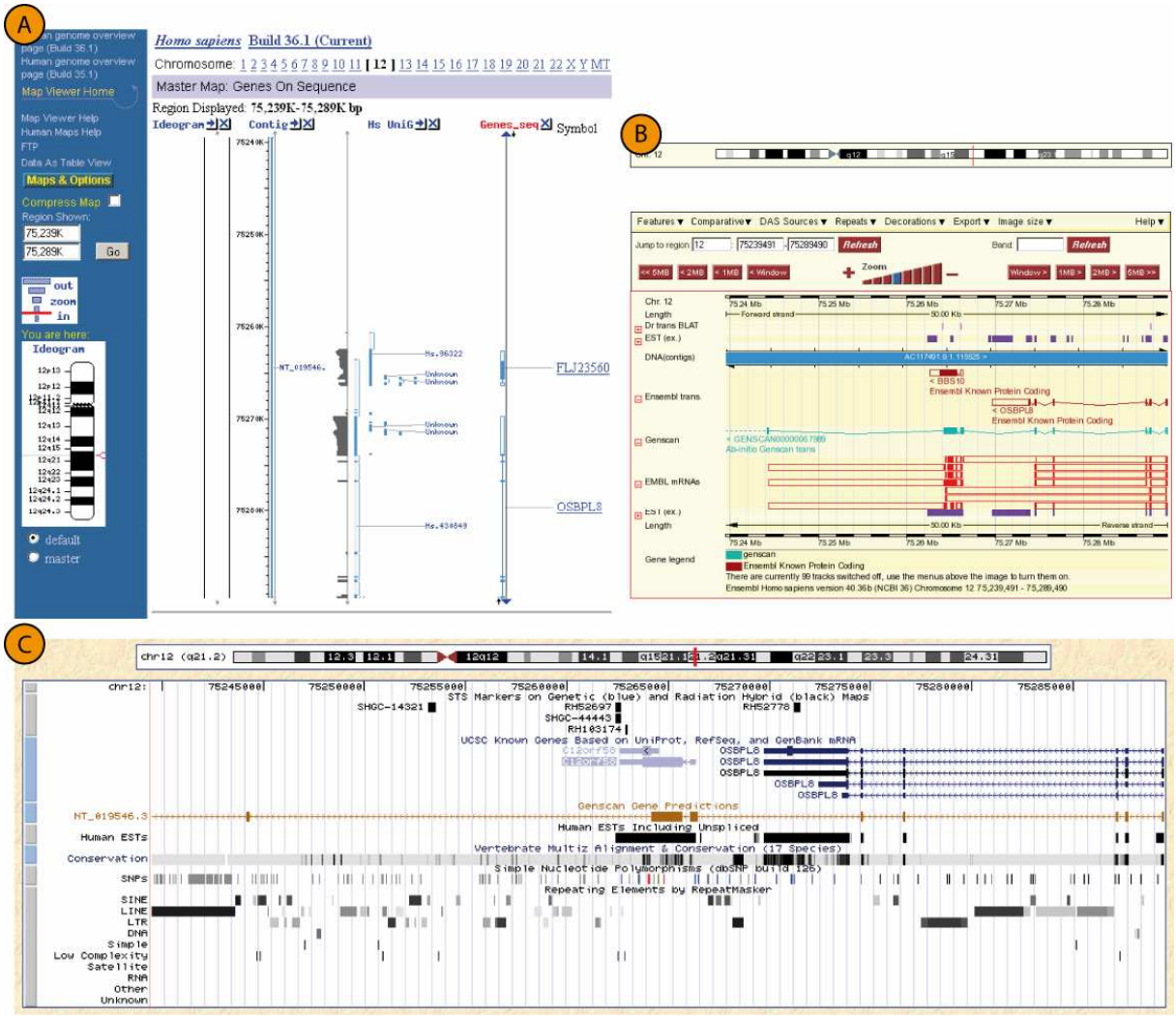


Figure 43 **Les « Genome browser ».**
Aperçu de la même région du chromosome 12 du génome humain dans les « genome browser » du NCBI (A), d'Ensembl (B) et de l'UCSC (C).

5.3 Les suites de programmes d'analyse de séquence

5.3.1 GCG

GCG Wisonsin package (Genetics Computer Group) est une suite de logiciels regroupant plus d'une centaine de programmes d'analyse de séquences commercialisée par Accelrys (http://www.accelrys.com/products/gcg). GCG est installée sur nos serveurs (la version 10 sur le serveur *titus* à l'IGBMC et la version 11 sur le serveur *CRPBIO* au CRP-Santé) et permet de manipuler, de visualiser, d'analyser et de comparer les séquences des banques installées localement et disponibles au format GCG.

5.3.2 EMBOSS

EMBOSS (European Molecular Biology Open Software Suite) est une suite de programmes non commerciale (http://emboss.sourceforge.net) dédiée à l'analyse de séquences et distribuée en Open Source. Il contient les outils majeurs d'analyse de séquence et permet une installation sur différents supports (Unix et Windows) ainsi qu'une interface web. EMBOSS peut être couplé à SRS et accède ainsi directement aux banques de séquences indexées par SRS. EMBOSS et l'interface web sont installés sur tous les serveurs à l'IGBMC et au CRP-Santé.

5.4 Le savoir-faire du laboratoire

5.4.1 GScope : l'ossature bioinformatique au service du laboratoire

L'annotation de *Pyrococcus abyssi* (Lecompte *et al.* 2001) étant le premier projet s'inscrivant dans le développement d'un axe de génomique au laboratoire, il a donné lieu au développement d'une plateforme logicielle de manipulation, de visualisation et d'analyse de données massives, GScope (Ripp *et al*., manuscrit en préparation) (Figure 44).

GScope est une plateforme de génomique dédiée à l'étude de données massives qu'il s'agisse de génomes complets ou d'un ensemble de gènes ou de protéines. Elle est développée au laboratoire en grande partie par Raymond Ripp, qui maintient la cohérence générale de cet outil, et agrémentée par tous les programmes individuels de chaque utilisateur ou développeur de GScope. La plateforme est écrite en Tcl/Tk et gère des fichiers ascii. Une interface Web est également utilisable.

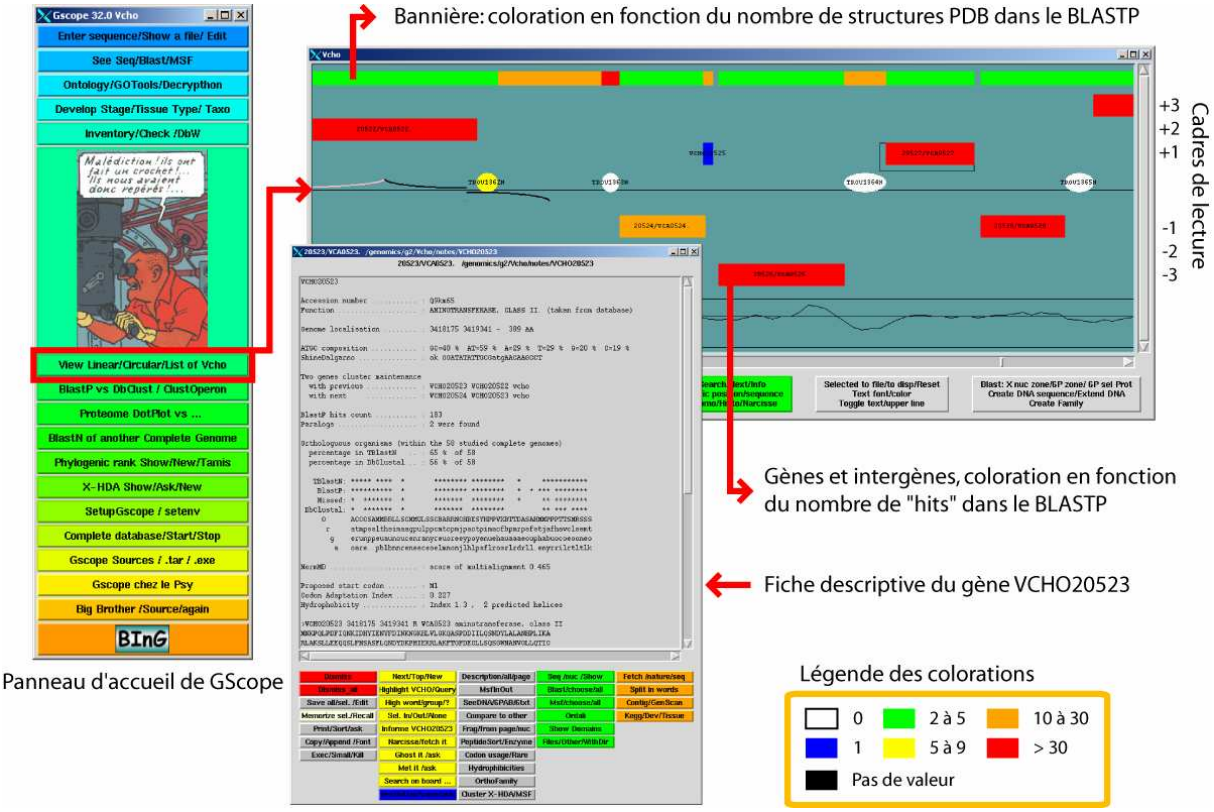


Figure 44 **Interface graphique de GScope.**
Illustration d'une région du 2ème chromosome de ***V. cholerae.*** Les gènes sont représentés sous la forme de rectangle dans leurs cadres de lecture respectifs (-1, -2, -3, +1, +2, +3). Les gènes et intergènes (formes ovales) sont coloriables en fonction de différents résultats. Une coloration additionnelle est possible pour les gènes dans la bannière supérieure. Une fiche descriptive contenant les informations calculées ou exportées est accessible pour chaque élément du génome.

Le gène et son produit étant au cœur de nombreux projets biologiques, GScope effectue des traitements bioinformatiques universels basés sur des programmes bioinformatiques « classiques » applicables à tous types de projets (tels que les traitements liés aux séquences) et sur des outils originaux développés dans le laboratoire. Son architecture modulaire permet l'utilisation de nouveaux modules dédiés à des activités d'analyse spécifiques en intégrant des programmes externes ou les nouveaux développements réalisés au sein du groupe.

Compte tenu de l'étendue des possibilités offertes par GScope, nous ne citerons que quelques exemples des modules disponibles dans le cadre de l'étude des génomes dans GScope :

- La prédiction de gènes dans les séquences génomiques par l'utilisation des programmes GLIMMER 1.0 et 2.0 (Salzberg *et al.* 1998; Delcher *et al.* 1999) pour les génomes procaryotiques.
- L'extraction des informations et la classification des gènes par la recherche de similarité dans les banques de séquences par l'utilisation des programmes BLAST et la réalisation d'alignements multiples de séquences complètes (PipeAlign).

- La validation des codons initiateurs (Lecompte *et al* manuscrit en préparation).
- L'étude de la distribution phylogénétique des gènes dans les génomes complets.

Le point commun de tous ces traitements est l'aspect qualitatif associé à chaque information récoltée par l'utilisation de validations croisées et/ou statistiques.

L'objectif de GScope, rappelons-le, est de prendre en charge l'analyse automatique de nombreuses données à haut-débit quelque soit leur origine (données liées à un génome, provenant d'expériences de protéomique, de transcriptomique, liées à une maladie …). Initialement dédié à la génomique comparative procaryotique, GScope a été adapté et diversifié afin d'étendre son champ d'action et de gérer tous les divers types de projets de génomique fonctionnelle.

GScope est également une fantastique plateforme de développement à laquelle toutes les personnes du laboratoire peuvent contribuer par l'ajout de fonctions ou de modules. C'est dans cette optique que RetScope, la plateforme d'analyse de séquences eucaryotiques s'est intégré au côté de GScope. Ce même cadre s'est appliqué à l'ensemble de mes développements qui se sont intégrés à GScope pour servir l'ensemble du laboratoire mais également par l'extraction de certains programmes dédiés et permettre ainsi leur distribution.

J'ai donc pu bénéficier de la longue expérience du laboratoire dans le développement d'outils bioinformatiques intégrés dans GScope afin de contribuer à son expansion.

5.4.2 RetScope : la plateforme d'analyse de séquences biologiques eucaryotes

Dans le cadre de la construction de la banque Actinome ainsi que de manière générale dans tous les développements informatiques, nous avons toujours été en étroite relation avec la plateforme d'analyse RetScope développée par Frédéric Chalmel au sein du laboratoire. Ceci a permis d'une part, de tester les programmes et d'autre part, de fournir des cas tests pour le développement de nouvelles applications. Le protocole de RetScope représente l'ensemble des modules dédiés à l'analyse et à l'annotation de séquences biologiques eucaryotes.

Je présenterai brièvement le protocole de RetScope et BlastPanel, un des outils majeurs utilisés lors de ma thèse.

5.4.2.1 Le protocole

L'aspect analyse de séquences nucléiques et protéiques est constitué d'un ensemble de protocoles reliés les uns aux les autres. Dans chacun d'eux, l'objectif est d'extraire des informations diversifiées des banques de données afin de caractériser un gène ou une protéine de la manière la plus complète possible. La plateforme RetScope a été développée dans le but d'assurer la qualité des séquences à analyser et de proposer différentes annotations (domaines, GO). A partir de séquences nucléiques, Retscope va déterminer parmi l'ensemble des ARNm disponibles pour un gène dans la banque GenBank, celui qui sera le plus complet. Ceci se traduit par une partie codante complète et par des régions non transcrites (UTR) les plus longues en 3' et en 5'. A partir de ses séquences nucléiques, RetScope va également attribuer une séquence protéique. Une validation croisée permet d'assurer que la séquence protéique et la séquence nucléique sont bien attribuées.

5.4.2.2 BlastPanel

Un nombre important de développements au sein de cette thèse est lié à l'utilisation de la recherche de similarité dans les banques de séquences tant au niveau nucléique que protéique et ceci notamment, par l'utilisation du programme BLAST (voir chapitre 2.3.2.2 BLAST p60). Les résultats d'une recherche au moyen du programme BLAST génére une liste de séquences détectées ainsi que différents scores (E-value, pourcentage d'identité) et des alignements.

Il apparaît ainsi fondamental de disposer d'un outil pouvant manipuler ces données. Ceci est réalisé par BlastPanel qui est un programme écrit en Tcl/Tk contenant un analyseur syntaxique ou « parser » du format de sortie du BLAST, des fonctions d'analyses avancées et une interface graphique avancée de visualisation des résultats. Son code source a été intégré et utilisé abondamment dans toutes les applications développées.

Parmi les fonctions les plus utilisées on peut citer, outre le « parser », les calculs du « Pourcentage d'Identité Globale » et des « Pourcentages de Recouvrement » qui ont été intégrées dans ARPAnno, CADO4MI et ComIcs. Le calcul de ces différentes valeurs sont décrites dans la Figure 45 :

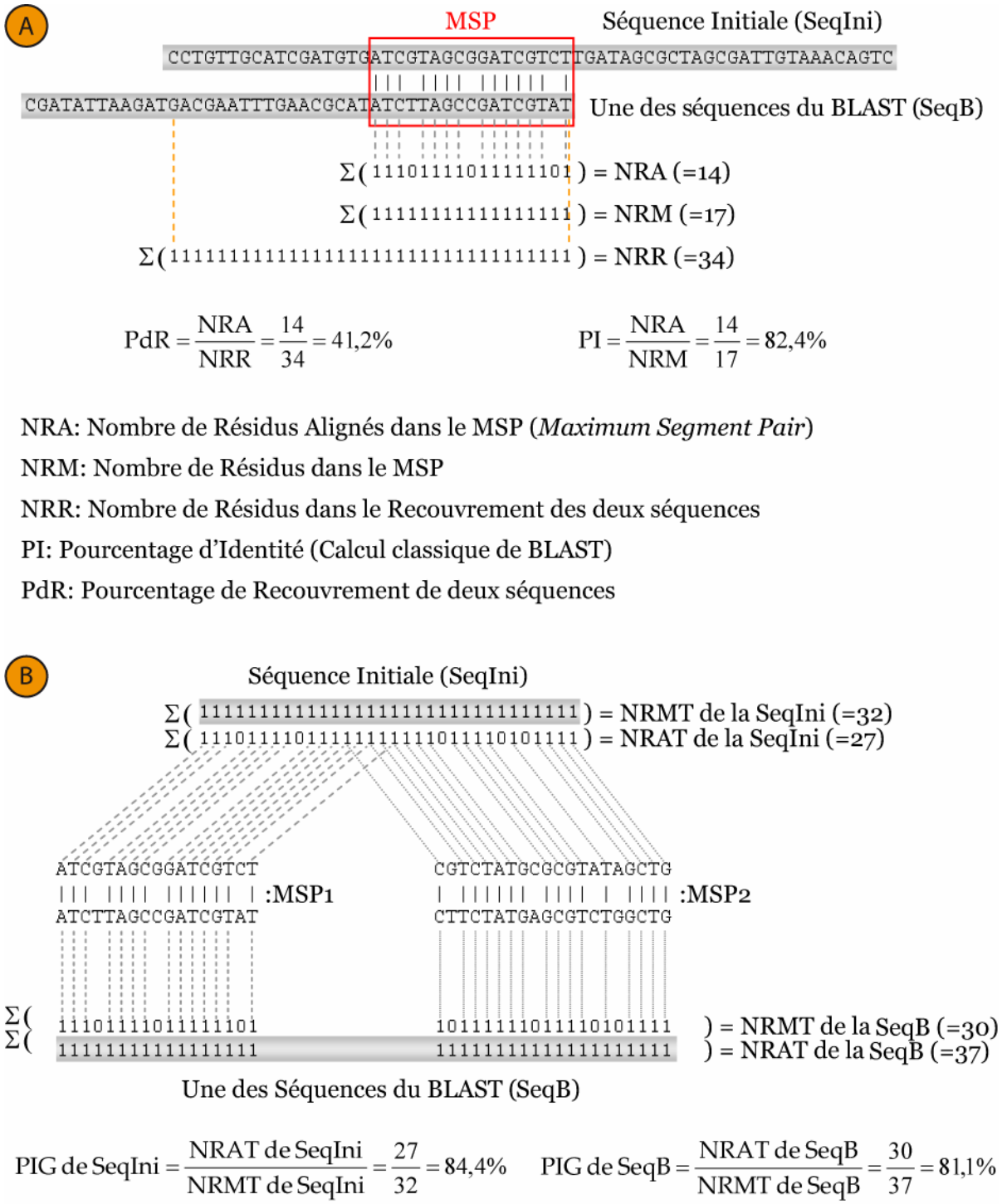


Figure 45 **Schéma illustrant les calculs du « Pourcentage d'Identité Globale » et des « Pourcentages de Recouvrement ».**
Figure issue de la thèse de Frédéric Chalmel 2004.

- Le Pourcentage de Recouvrement (PdR) entre deux séquences qui se définit comme le rapport entre le Nombre de Résidus Alignés (NRA) et le Nombre de Résidus dans

la région de Recouvrement entre les deux séquences (NRR). Ce calcul est notamment capital pour le module de validation de la présence de la séquence dans CADO4MI (9.9.2 Validation de la séquence appât) et constitue un critère de sélection dans ARPAnno (8.1.2 Protocole du serveur ARPAnno).

- Le Pourcentage d'Identité Globale de SeqIni (ou SeqB) comme étant le ratio entre NRAT (l'ensemble des résidus de SeqIni alignés correctement dans au moins un MSP avec SeqB) de SeqIni et NRMT (Nombre de Résidus dans les MSP Total) de SeqIni. L'avantage de ce calcul est que contrairement au calcul PI du BLAST, celui-ci tient compte de l'ensemble des segments (MSP). Il permet donc de calculer un pourcentage d'identité entre une SeqA et une SeqB sur toute la surface de SeqA et non pas uniquement sur la surface d'un segment du Blast.

5.5 Autres programmes utilisés

5.5.1 Recherche de motifs

Les programmes Findpatterns de GCG ou Fuzznuc et Fuzzpro d'EMBOSS permettent de rechercher spécifiquement un motif protéique ou nucléique dans une banque de séquences à partir d'une séquence consensus. Le motif peut être défini de façon stricte ou ambiguë. La recherche peut être elle-même assouplie par l'utilisation du paramètre « MISmatch » qui définit le nombre de mésappariements autorisés entre le motif et les séquences de la banque. Ce type de recherche a été en particulier utilisé pour vérifier et affiner la définition des motifs caractéristiques des familles d'« *Actin-Related Protein* » ou ARP et pour rechercher les protéines appartenant à ces familles.

5.5.2 La recherche par la méthode des profils

Pour rechercher des homologues éloignés de la protéine BBS10, nous avons utilisé la méthode du profil qui permet de détecter des similarités protéiques plus faibles que la comparaison deux à deux des séquences. Le profil est en effet une matrice de scores position-spécifique qui reflète la probabilité pour chaque acide aminé de se trouver en une position donnée d'un motif. Elle est construite à partir des fréquences d'occurrence des acides aminés à chaque position d'un alignement multiple et des probabilités de mutation des acides aminés suivant une matrice de substitution.

Cette méthode est utilisée dans le programme PSI-BLAST (Altschul *et al.* 1997) qui permet de comparer une séquence protéique à une banque de séquences protéiques. La première étape

est une recherche BLASTP « classique ». Les séquences atteignant une valeur d'Expect seuil sont utilisées pour construire un profil qui sera à son tour comparé aux séquences de la banque, et ainsi de suite de manière itérative.

5.5.3 Edition et mise en forme des alignements multiples

La correction manuelle des alignements multiples, en se basant sur la conservation de la structure secondaire par exemple, a été réalisée dans l'interface graphique Seqlab du package GCG qui permet de visualiser, de manipuler et d'éditer les alignements multiples de manière conviviale (Figure 46). Pour la mise en forme des alignements multiples en vue de publication ou de présentation des résultats, nous avons eu recours au programme Genedoc (http://www.psc.edu/biomed/genedoc) et au programme Jalview (http://www.jalview.org) (Clamp *et al.* 2004).

Figure 46 **Capture d'écran de 2 éditeurs d'alignements multiples.**
Les 2 éditeurs présentés sont Seqlab (A) et Jalview (B).

5.5.4 Edition et mise en forme d'arbres phylogénétiques

L'étude approfondie de certains arbres phylogénétiques a été réalisée grâce au programme PHYLO_WIN (Galtier *et al.* 1996) qui offre une interface graphique conviviale. Il permet de définir facilement les positions d'un alignement multiple retenues pour la construction d'un arbre et offre la possibilité d'utiliser plusieurs méthodes de construction ainsi que de tester la robustesse de l'arbre par la méthode du Bootstrap. Nous avons essentiellement utilisé PHYLO_WIN avec la méthode de « Neighbour Joining » en « global gap removing » et la robustesse des arbres construits a été testée par 500 « bootstraps ». Pour la mise en forme des arbres phylogénétiques, nous avons utilisé principalement 2 programmes TreeView (Page 1996) pour les arbres sans racine et ATV (Zmasek *et al.* 2001) pour les arbres avec racine.

5.5.5 Visualisation et mise en forme des structures tridimensionnelles

Nous avons utilisé le logiciel PyMOL (http://pymol.sourceforge.net) pour représenter et déplacer les molécules dans l'espace. Des fonctions permettent de visualiser et d'explorer les différentes parties ou propriétés d'une molécule. PyMOL est capable d'importer tous les formats de modèles moléculaires disponibles et peut être contrôlé au moyen de scripts. La qualité et la diversité des représentations disponibles ainsi que les possibilités offertes à un utilisateur averti font de ce logiciel un outil indispensable à l'analyse des structures.

5.6 Ecriture de programmes

5.6.1 Le Tcl/Tk

Le Tcl/Tk (http://www.tcl.tk) est un langage de programmation développé par John Ousterhout, de l'université de Californie, Berkeley en 1988. Il est composé de 2 parties ; le Tcl (*Tool command language*) qui est un langage de commandes et Tk (*Toolkit*) qui est une extension graphique de haut niveau. Le Tcl/Tk est, comme le Perl ou le Python, un langage de script, ou encore de programmation dynamique qui autorise un développement rapide.

Tcl/Tk a permis le développement de l'ensemble des applications qui sont présentées dans cette thèse. Ce langage informatique possède des nombreux avantages :

- Le Tcl/Tk est un langage de script interprété, il ne nécessite donc aucune compilation.
- C'est un langage multi-plateforme (Windows, MacOS et tout système Unix ou Linux) qui ne nécessite aucun changement de code source pour être exécuté.

- Le langage et sa documentation sont open source et donc distribués gratuitement.
- Il est très adapté pour la lecture et la manipulation des fichiers car en Tcl/Tk, toutes les variables sont des chaînes de caractères (langage non typé).
- Le Tk est une extension graphique avancée et extrêmement complète pour développer des interfaces graphiques indispensables pour les biologistes.
- Enfin, le Tcl/Tk est un langage dont la syntaxe riche et simple ajoute à sa souplesse des qualités pour un développement rapide.

5.6.2 La philosophie

L'ensemble des programmes développés au cours de cette thèse s'inscrit dans la biologie à haut débit et donc, dans le traitement d'informations nombreuses et hétérogènes. Autour de l'architecture classique des programmes écrits de manière séquentielle (Figure 47 A) qui comprennent un module central ou cœur du programme et une interface graphique (*Graphical User Interface* ou GUI), nous avons associés des contrôleurs à chaque type de donnée à utiliser. Le fonctionnement des contrôleurs réside dans le fait que ce sont eux seuls qui détiennent l'information et qui la transmettent au moment choisi (Figure 47 B). Ces fonctions sont généralement appelées « *Interroge*… » ou « *Ask*... ». Ainsi *InterrogeLesInfosDeLOligo* permet d'interroger l'intégralité des informations stockées dans un fichier « .oligo » généré par CADO4MI. Pour cela, lors du premier accès à une information par l'utilisation d'une procédure *Interroge*, cette dernière appelle une fonction « *Charge*… » ou « *Load*… » qui permet de lire l'intégralité des informations du fichier et de les indexer dans un tableau conservé en mémoire (variable globale en Tcl). Les fonctions *Charge* sont des « parser » ou analyseurs syntaxiques. Une fois les informations indexées et laissées en mémoire, *Interroge* peut être utilisée indéfiniment sans jamais avoir besoin de relire le fichier pour répondre très rapidement aux interrogations de l'utilisateur si ces dernières correspondent à une indexation dans le tableau. Une fois l'utilisation du fichier terminée, *Interroge* permet également d'effacer toutes les informations stockées dans la mémoire. Selon l'utilisation, il est possible de manipuler plusieurs fichiers du même type simultanément.

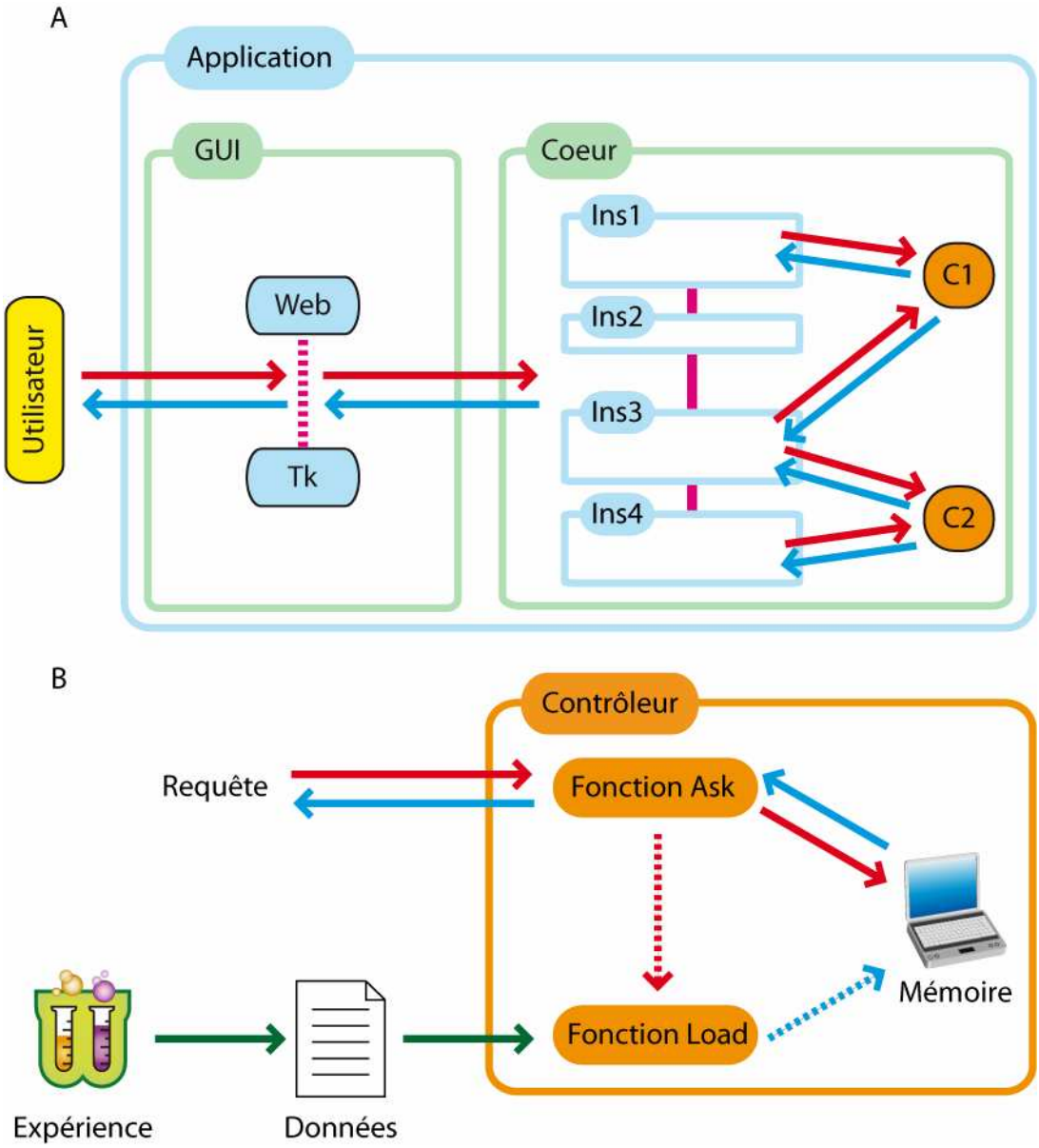


Figure 47 Organisation générale des applications développées ainsi que l'utilisation des contrôleurs.

Les flèches rouges indiquent des requêtes alors que les flèches bleues indiquent des réponses. (A) L'interface graphique (GUI) transmet les données soumises par l'utilisateur au programme (cœur) qui exécute les instructions (Ins1-4), puis retourne un résultat à l'utilisateur. Les instructions (Ins) font appel aux données au travers de contrôleurs (C1 et C2). (B) Un contrôleur reçoit une requête qu'il traite la première fois (flèches en pointillée) par le chargement (Fonction Load) des données et retourne la réponse. Toute nouvelle requête ne nécessitera plus cette étape de chargement et accèdera directement aux informations stockées en mémoire par le contrôleur.

Cette architecture a plusieurs avantages, d'une part, un accès contrôlé par un seul point d'entrée et d'autre part, une optimisation du temps de requête pour accéder à l'information. Cette manière de procéder peut également s'apparenter à de la programmation objet puisqu'on accède aux données d'un objet au travers d'une fonction.

Chapitre 6 - Actinome

6.1 Une collection de séquences

Actinome est une banque de données bioinformatiques créée par GScope et RetScope, construite autour du cytosquelette. Elle contient des séquences nucléiques et protéiques, les résultats de recherche de similarité dans les banques de séquences généralistes, l'alignement multiple pour chaque famille de protéines et les annotations de la Gene Ontology. Cette banque de séquences est à la base de l'ensemble des données générées au cours de cette thèse, notamment pour le design de sondes pour Actichip et pour l'étude des profils phylogénétiques.

6.1.1 Historique

Actinome a été élaborée à partir de plusieurs stratégies successives combinant recherche bibliographique et bioinformatique. Chacune des étapes a permis d'enrichir la banque en nombre de séquences liées aux cytosquelettes. Son nom est lié à la première version de la banque qui était centrée sur le cytosquelette d'actine.

Ainsi, la première version de la banque est issue d'une recherche bibliographique et des connaissances des biologistes spécialistes du cytosquelette au LBMAGM. Ils ont dans un premier temps focalisé la recherche sur les gènes impliqués dans le cytosquelette d'actine et certains composants majeurs des 2 autres cytosquelettes (microtubules et filaments intermédiaires). Cette recherche a permis de collecter 1164 séquences d'ARNm issues de la banque nucléique GenBank. Le protocole RetScope a ensuite été appliqué à ces 1164 séquences et a permis d'attribuer une séquence protéique et une séquence nucléique validées. RetScope a également permis par exemple, la création d'alignements multiples pour chaque entrée de la banque et de déterminer l'annotation de ses protéines par l'attribution de termes GO. La localisation chromosomique et la carte exonique de chacun des gènes ont également été déterminées.

La seconde version de la banque correspond à l'extension d'Actinome à l'ensemble des protéines du cytosquelette présentes dans les banques de protéines. En effet, à partir de l'analyse des annotations GO calculées par RetScope, nous avons déterminé un ensemble de termes ou access GO qui nous ont permis par la suite d'identifier les protéines

correspondantes. Ainsi, cette recherche a conduit à la détection de 1513 protéines. La répartition des protéines détectées pour les 3 cytosquelettes est variable : 48% pour le cytosquelette d'actine, 44% pour les microtubules et seulement 8% pour les filaments intermédiaires.

Enfin, une mise à jour supplémentaire a permis à partir d'un travail bibliographique supplémentaire d'ajouter des protéines marqueurs de certains tissus et des protéines des voies de signalisation comme les petites Rho GTPases.

La banque complète en version 3 compte ainsi 3061 entrées initiales (Figure 48).

Figure 48 **Les différentes versions de la banque Actinome.**
Les chiffres au-dessus et en-dessous des versions indiquent respectivement le nombre d'entrées soumises et le nombre d'entrées retenues après traitement de la redondance. En gris sont indiquées les limites de numéros d'accès utilisés.

Au travers des différentes analyses de RetScope, il est apparu que la première version de la banque dénombrait des doublons. En conséquence, nous avons décidé de traiter la redondance pour cette version et pour toutes les mises à jour suivantes. Ainsi, à chaque nouvelle mise à jour les nouvelles séquences sont comparées avec les entrées de la version précédente de la banque. Le protocole de traitement des doublons est basé sur la comparaison des séquences entre elles à la fois au niveau nucléique et au niveau protéique. En conséquence, si la version initiale comptait 3061 entrées, la version 3 traitée pour la redondance ne compte plus que 1742 entrées. A titre d'exemple, la version 2.0 contenait 241 séquences déjà présentes dans la première version.

D'un point de vue pratique, chaque entrée se voit attribuer un numéro d'accès de la forme ACTXXXX, X étant un nombre (par exemple ACT0004, correspond à l'isoforme d'actine appélée « alpha skeletal muscle »). L'accès à la banque se fait par l'interface commune aux projets GScope ou par des programmes écrits en Tcl/Tk.

6.1.2 Des catégories de gènes

Le première version de la banque contient un maximum de séquences liées au cytosquelette d'actine, cependant les différentes mise à jour ont permis de construire une banque dédiée à l'ensemble du cytosquelette et contient par exemple :

- Les acteurs majeurs des 3 cytosquelettes ; les isoformes d'actine, les tubulines et les protéines responsables des filaments intermédiaires (kératines, vimentine, lamine…)
- Les protéines de nucléation (complexe Arp2/3, VASP) qui initient la polymérisation de l'actine près de la membrane plasmique.
- Les protéines de « coiffe » (p.ex. gelsoline, adséverine) qui, en s'associant à l'extrémité « plus » du filament, inhibent l'addition des monomères d'actine et restreignent ainsi la polymérisation de l'actine à des sites précis.
- Les protéines qui, en déstabilisant les filaments d'actine (p. ex. cofiline, ADF) ou en séquestrant les monomères d'actine (p. ex. thymosine, profiline), assurent la continuité du cycle de polymérisation.
- Les protéines motrices comme les myosines, les kinésines, la dynactine.
- Les protéines de structuration (p. ex. alpha-actinine, filamine, plastine) qui, en organisant les filaments d'actine en faisceaux ou réseaux, participent à la structuration du cortex cellulaire et ainsi, au modelage de la forme cellulaire, au mouvement et à l'assemblage de noyaux de signalisation.
- Les protéines d'« échafaudage » (p. ex. zyxine, ezrine, diaphanous) qui sont capables d'interagir avec plusieurs composants du cytosquelette et intègrent les signaux motogènes.
- Les petites GTPases de la famille rho, leurs facteurs régulateurs et leurs cibles. Situées en aval des récepteurs kinases des voies motogènes (EGF, HGF), elles contrôlent non seulement l'organisation du cytosquelette et la polarité cellulaire (réponse aux signaux chémotactiques), mais également la transcription.
- Les protéines de lien entre les cytosquelettes.
- Les récepteurs des jonctions adhérentes (cadhérines, caténines, taline et vinculine).
- Les marqueurs des cellules épithéliales ou mésenchymateuses (kératines, vimentine)

Chapitre 7 - ComIcs

7.1 Une plateforme de génomique comparative

ComIcs, l'abréviation de « **Com**parative Genom**Ics** », est une plateforme d'analyse dédiée à la génomique comparative. Elle a pour but de rassembler des outils de calcul, d'analyse et de visualisation des résultats ayant trait à la génomique comparative. ComIcs a été développé en Tcl/Tk et utilise les programmes BLAST, getz et Fastacmd ainsi que les banques de protéines généralistes (au format BLAST pour les recherches de similarité et au format SRS pour les requêtes visant les informations attachées aux séquences) et les banques de séquence des génomes eucaryotes complets. Le programme est focalisé sur la définition des profils phylogénétiques, c'est-à-dire la distribution des gènes ou protéines au sein d'un groupe d'organismes.

7.1.1 Philosophie générale du programme

La philosophie de ComIcs s'articule autour de la notion de projet. Un projet d'analyse doit être paramétré, sauvé, exécuté et ses résultats étudiés. Ceci se traduit au niveau de l'interface centrale du programme (Figure 49) qui permet de définir un projet ComIcs, d'effectuer les actions choisies ou d'accéder à la visualisation des résultats. Le panneau central est également accompagné d'un panneau de log qui permet de visualiser les paramètres stockés et définis par l'utilisateur, mais également de lui fournir certaines indications.

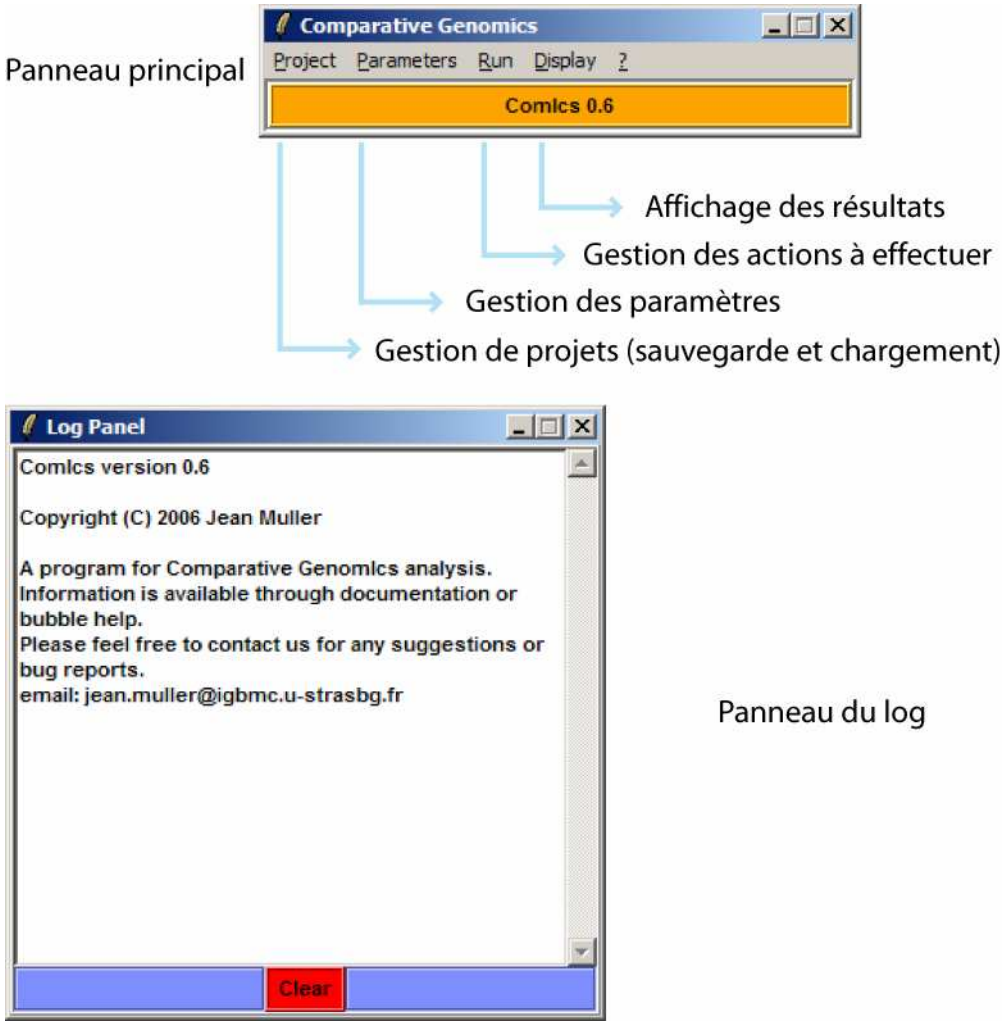


Figure 49 **Panneau central de ComIcs.**

Un projet ComIcs est défini à partir de séquences « appâts », des critères de recherche de similarité et d'analyse de ces recherches. Les paramètres peuvent être sauvés dans un fichier « .comics ». Ce fichier peut être chargé ultérieurement et permet par exemple de scinder l'exécution des taches dans le temps ou l'analyse ultérieure des données. L'ensemble des paramètres est accessible au travers du panneau principal (Figure 50).

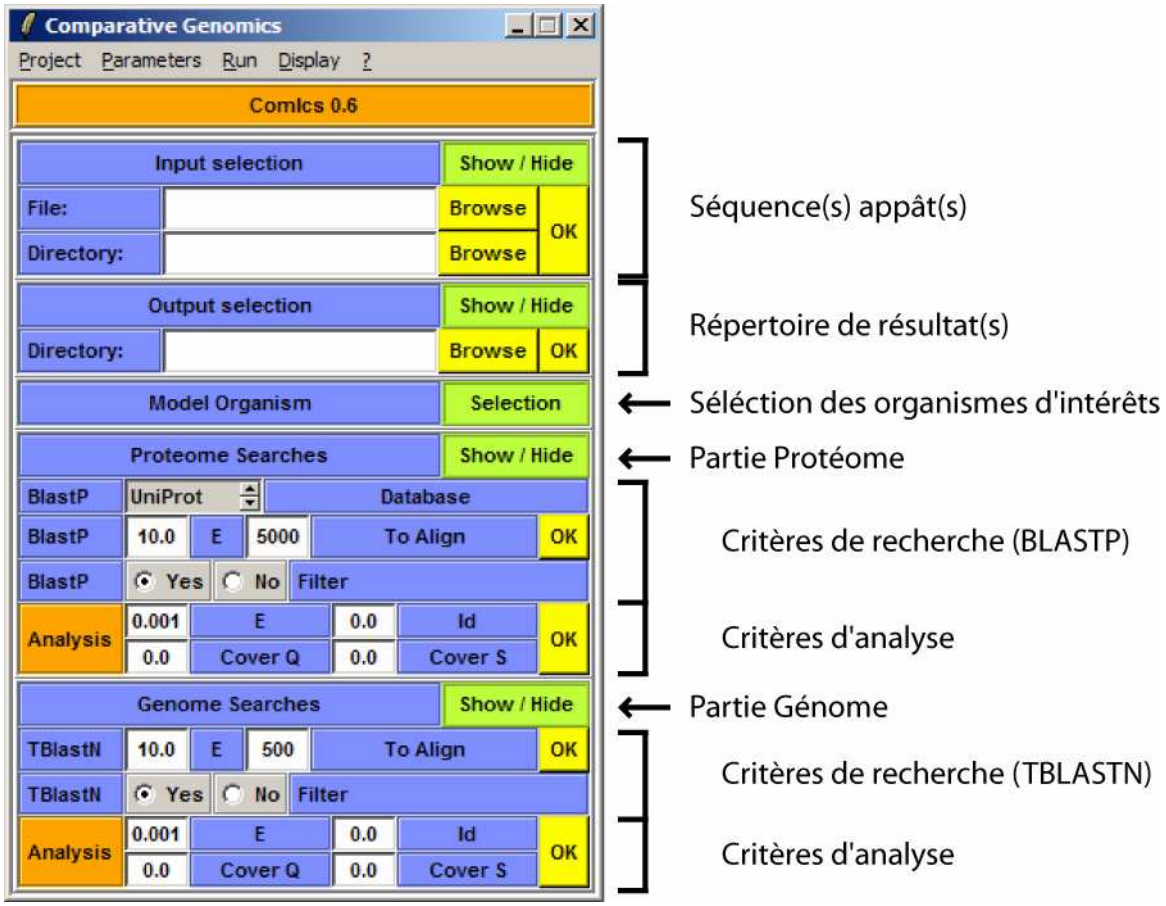


Figure 50 **Panneau de configuration de ComIcs**
Le panneau central de ComIcs est représenté avec l'affichage des tous les volets de sélection des paramètres de recherche. Ceci constitue le point de départ des recherches dans les protéomes et dans les génomes.

7.1.2 Nomenclature

Dans le but de clarifier l'emploi de certains termes dans l'interface de ComIcs, voici quelques définitions ; Query ou SeqID désigne la séquence appât, BID ou Subject désigne une séquence ou un identifiant détecté dans une banque au moyen de BLAST, PID le pourcentage d'identité, CoverQ pour le pourcentage de recouvrement de la Query vis-à-vis d'un Subject et CoverS le contraire.

7.1.3 Les fichiers d'entrée et de sortie

Les fichiers d'entrée sont des fichiers de séquences au format FASTA (voir Annexe 4). L'utilisateur peut soumettre directement un ou plusieurs fichiers ou un répertoire contenant l'ensemble de ses séquences au format FASTA (Figure 51).

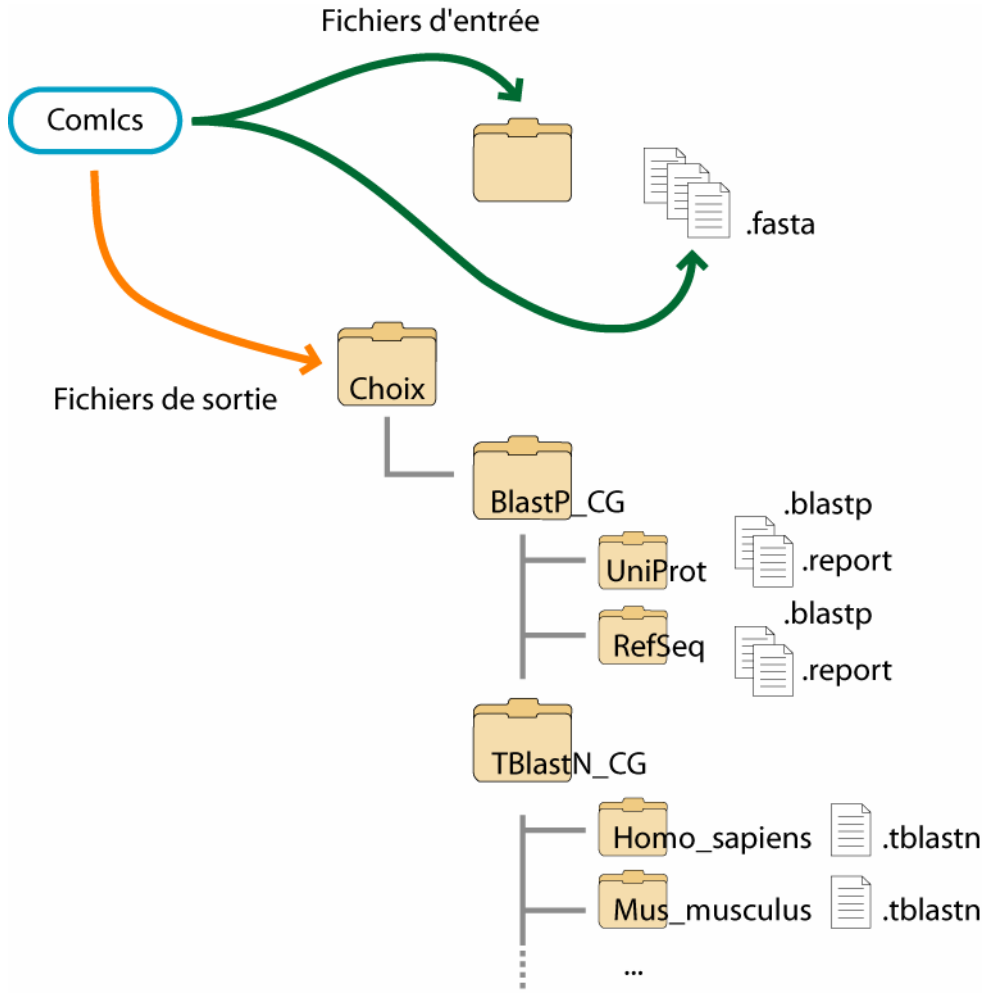


Figure 51 **Organisation des fichiers d'entrée et de sortie dans ComIcs.**
Pour réaliser une analyse avec ComIcs, l'utilisateur définit les séquences appâts en choisissant un répertoire contenant les fichiers des séquences au format FASTA ou directement un ou plusieurs fichiers séparés. Les fichiers de sortie et leur arborescence sont gérés entièrement par ComIcs à partir du choix de l'utilisateur.

La gestion des fichiers de sortie se fait à partir d'un point de l'arborescence choisi par l'utilisateur. ComIcs organise alors les fichiers de sortie en fonction des actions demandées (Figure 51). Le répertoire « BlastP_CG » contient les résultats des recherches de similarité dans les banques de protéines et le répertoire « TBlastN_CG » contient les résultats de recherche de similarité dans les génomes. Les fichiers générés sont des fichiers au format BLAST, « .blastp » ou « .tblastn » (recherches BLASTP et TBLASTN) et des fichiers d'analyse des BLASTs « .report ». Le fichier « .report » est un fichier séparé par des tabulations qui contient une ligne par organisme et l'ensemble des identifiants retenus ainsi que les valeurs calculées lors de l'analyse des résultats de la recherche de similarité (cf Figure 53).

Les fichiers de résultats sont des fichiers à plat qui peuvent éventuellement être consultés par l'utilisateur, mais une interface de visualisation permet de rendre leur analyse totalement transparente.

7.1.4 La liste des organismes et leur gestion

La définition des profils phylogénétiques est réalisée à partir d'une liste d'organismes dont la séquence géomique complète est disponible. Nous avons ainsi répertorié 41 génomes d'organismes eucaryotes : *Homo sapiens, Mus musculus, Rattus norvegicus, Gallus gallus, Xenopus tropicalis, Tetraodon nigroviridis, Brachydanio rerio, Takifugu rubripes, Ciona intestinalis, Strongylocentrotus purpuratus, Drosophila melanogaster, Anopheles gambiae, Apis mellifera, Caenorhabditis elegans, Caenorhabditis briggsae, Schizosaccharomyces pombe, Saccharomyces cerevisiae, Candida albicans, Candida glabrata, Kluyveromyces lactis, Ashbya gossypii, Debaryomyces hansenii, Yarrowia lipolytica, Aspergillus fumigatus, Neurospora crassa, Cryptococcus neoformans, Encephalitozoon cuniculi, Dictyostelium discoideum, Entamoeba histolytica, Thalassiosira pseudonana, Trypanosoma cruzi, Leishmania major, Plasmodium falciparum, Cryptosporidium parvum, Cryptosporidium hominis, Theileria parva, Tetrahymena thermophila, Giardia lamblia, Arabidopsis thaliana, Oryza sativa* et *Cyanidioschyzon merolae.* La séquence génomique de chacun de ces génomes est rendue disponible et mise à jour quotidiennement sous la forme de banques au format BLAST par la plateforme de bioinformatique de Strasbourg. Les séquences sont téléchargées à partir de différents sites web généralistes (NCBI, JGI, UCSC, Ensembl…) ou spécialisés (Dictybase, CryptoDB…).

La cohérence des différents noms d'organismes et leur utilisation sont totalement transparentes au sein de ComIcs. Nous utilisons à la fois le nom scientifique, le nom abrégé (utilisé par UniProt) et le TaxId (identifiant de la taxonomie établie par le NCBI). La liste de tous les abrégés (http://www.expasy.org/cgi-bin/speclist) ainsi que l'ensemble des identifiants de la taxonomie (ftp://ftp.ncbi.nih.gov/pub/taxonomy) sont intégrés dans ComIcs et permettent par exemple, d'identifier le génome humain à la fois par *Homo sapiens* (nom scientifique), HUMAN (nom abrégé) et 9606 ou 63221 (TaxId).

7.1.5 Définition des profils phylogénétiques

La stratégie employée par ComIcs est basée sur la recherche de similarité dans les banques protéiques pour une ou plusieurs séquences appâts (Figure 52). Les séquences requêtes sont alignées au moyen du programme BLASTP et les résultats sont analysés en fonction de plusieurs critères (voir ci-dessous) et les profils phylogénétiques sont ainsi déterminés pour chaque organisme.

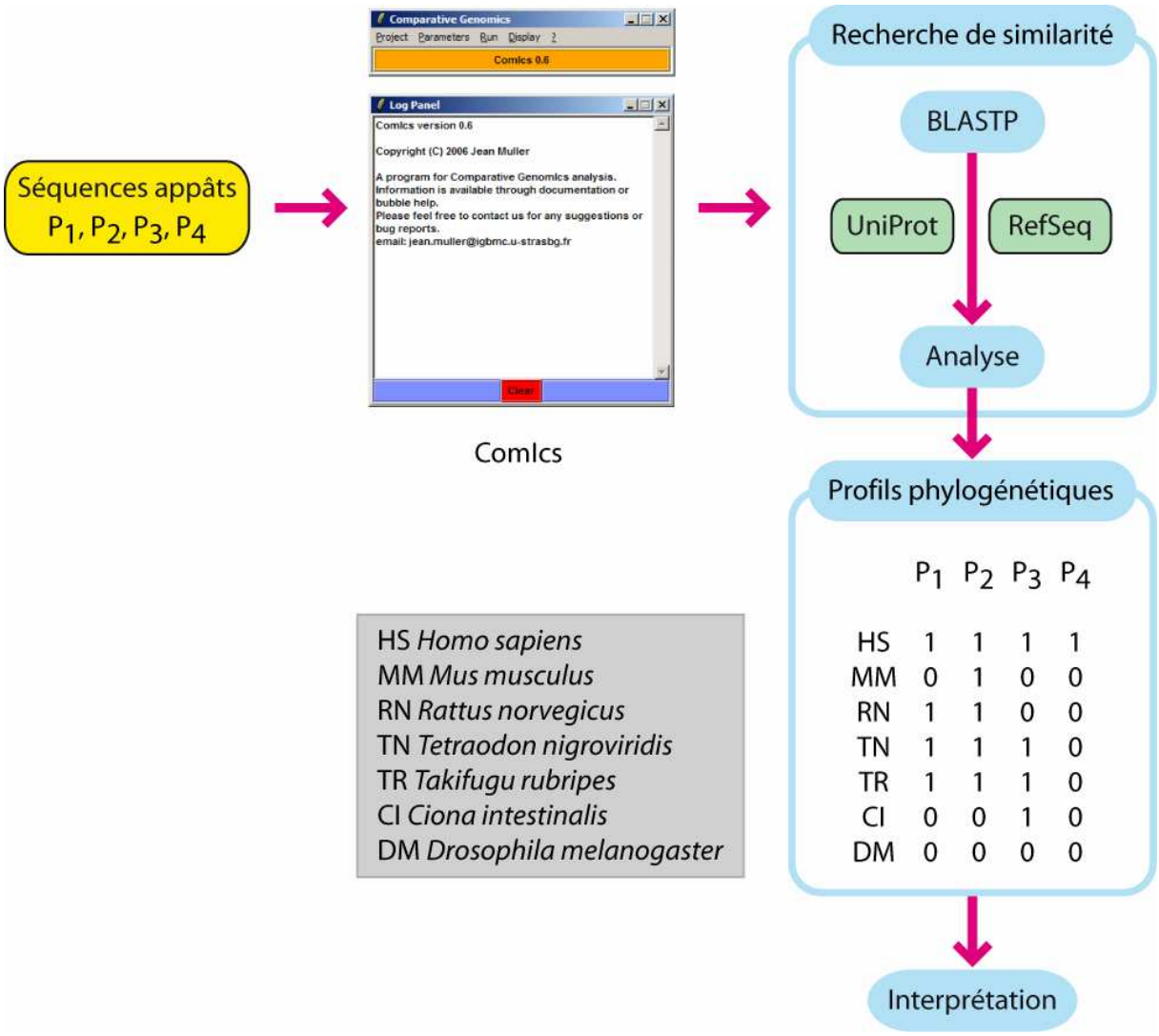


Figure 52 **Stratégie utilisée pour la détermination des profils phylogénétiques.**

7.1.5.1 BLASTP

La détection des séquences protéiques pour chacun des organismes est réalisée par le programme BLASTP appliqué aux banques de protéines généralistes UniProt et RefSeq (voir 2.2.1.2 Les banques de protéines). Les paramètres principaux sont accessibles au travers de l'interface (Figure 50) permettant à l'utilisateur de spécifier par exemple, une limite d'expect et/ou de nombre de séquences. Il faut noter qu'il n'existe pas de banque dédiée à chacun des protéomes des organismes utilisés et qu'il faudra donc extraire l'organisme attaché à la séquence traitée (voir ci-dessous).

7.1.5.2 Analyse des résultats

L'analyse des résultats de la recherche de similarité fait appel aux programmes développés par Frédéric Chalmel (voir 5.4.2.2 BlastPanel) et consiste à trier les séquences détectées en

fonction de 2 critères : proximité entre la séquence appât et une séquence « subject » et appartenance ou non de la séquence « subject » à un organisme de notre liste.

Pour ce faire, nous utilisons :

- d'une part l'analyseur syntaxique pour extraire les identifiants des séquences qui remplissent les critères choisis par l'utilisateur. Ces critères sont la valeur d'expect, le pourcentage d'identité (PID) et les 2 pourcentages de recouvrement (CoverQ et CoverS) (voir 5.4.2.2 BlastPanel).
- d'autre part, l'extraction de l'organisme d'origine de chaque séquence. Cette extraction dépend de la banque utilisée. Pour UniProt, ce sont les Noms Abrégés de l'organisme qui sont utilisés puisque depuis 2005 l'ensemble des identifiants (ID) UniProt sont de la forme XXX_Abrégé, par exemple l'actine béta humaine est référencée sous le nom ACTB_HUMAN. Dans le cas de RefSeq, nous utilisons le TaxId. Le TaxId est extrait des entrées de la banque RefSeq en utilisant des requêtes SRS et des expressions régulières ou bien au moyen d'un script ICARUS qui permet d'extraire directement le TaxId.

On obtient alors pour une Query donnée une liste d'identifiants de séquences (BID) et leurs valeurs calculées qui seront sauvegardées dans le fichier « .report » (Figure 53).

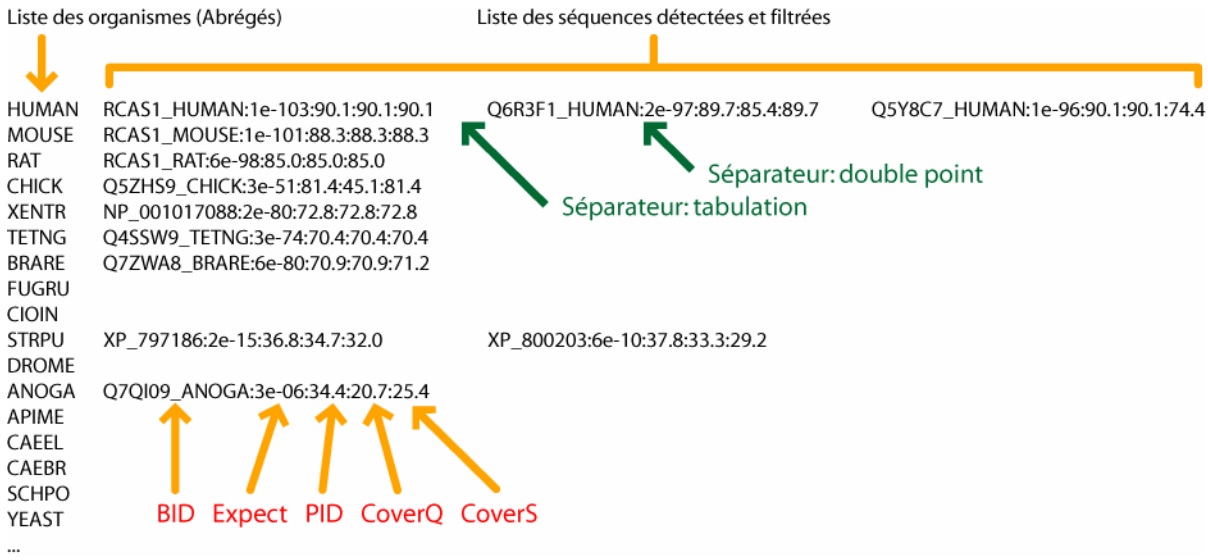


Figure 53 **Exemple de fichier au format « .report ».**

7.2 Interface d'analyse des données

7.2.1 Bilan de présence/absence

Les profils phylogénétiques ou bilan de présence/absence sont affichés sous la forme d'un tableau dont les entêtes de lignes correspondent aux noms des organismes et les entêtes de colonnes correspondent aux identifiants des séquences appâts (« SeqID ») (Figure 54). L'intersection d'une ligne et d'une colonne permet d'afficher un rectangle de couleur correspondant à la présence (rectangle vert) ou à l'absence (rectangle rouge) d'une séquence similaire à la séquence appât dans un organisme donné. On obtient ainsi par colonne, le profil phylogénétique d'une séquence dans les 41 organismes de référence. La présence ou l'absence est déterminée à partir des critères définis par l'utilisateur au début du projet. Cependant, l'interface graphique permet de changer dynamiquement ces critères pour observer leurs effets sur les résultats affichés.

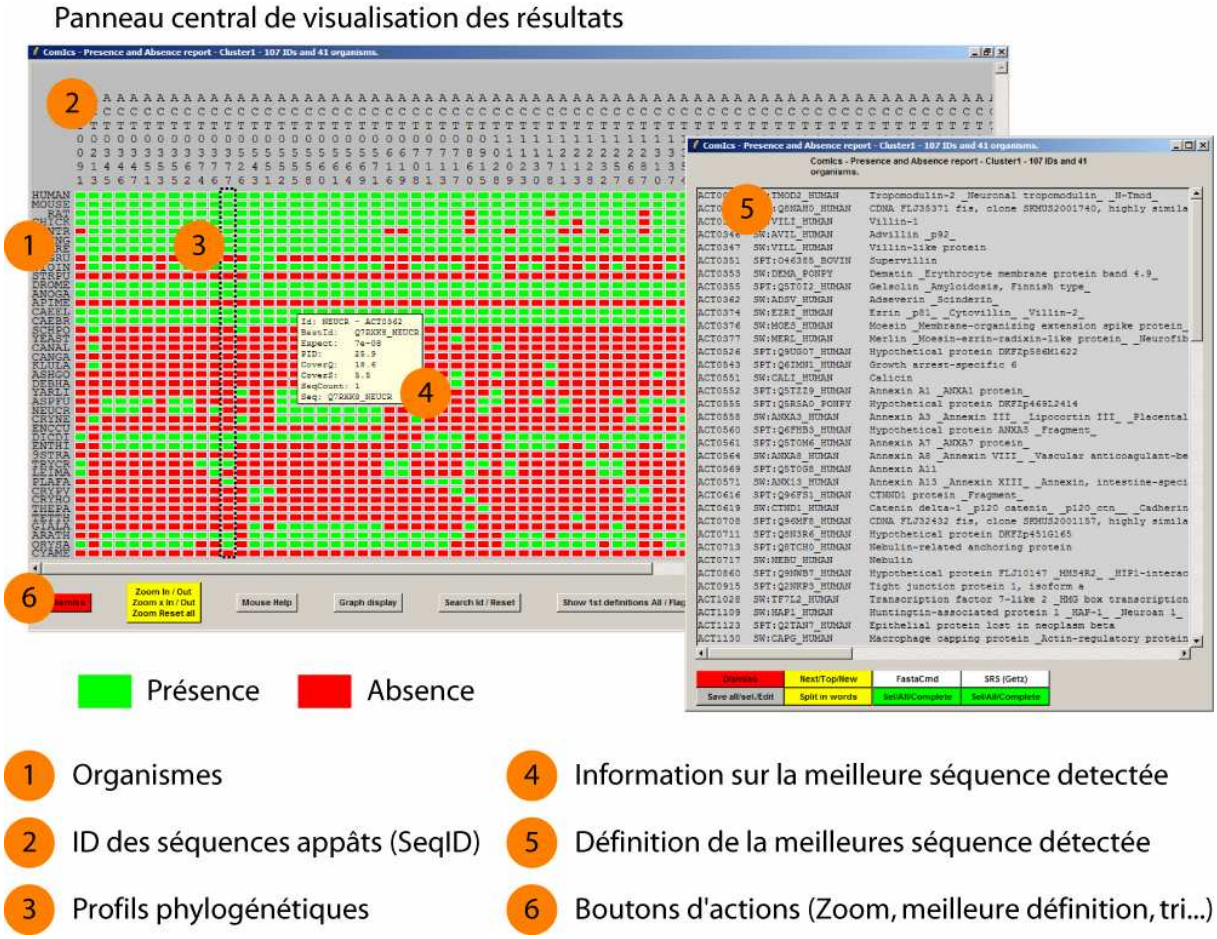


Figure 54 **Interface de visualisation et d'analyse des profils phylogénétiques.**

Différentes informations et outils sont accessibles à l'utilisateur afin d'analyser les données générées. L'utilisateur peut ainsi aisément déplacer des séquences (colonnes) ou des organismes (lignes) entre eux pour mieux organiser les résultats. Les changements peuvent être annulés, sauvés et rechargés.

A chaque rectangle est associée une série de données calculée lors de l'analyse du fichier BLAST (pourcentage d'identité, Expect et les pourcentages de recouvrement de la meilleure séquence détectée ainsi que le nombre de séquences détectées pour cet organisme et cette séquence) qui peuvent toutes être visualisées (voir l'infobulle Figure 54). Outre la possibilité de réarrangement manuel, l'utilisateur peut trier automatiquement et indépendamment les profils en fonction de chacune des valeurs décrites plus haut. On peut ainsi trier par ligne en cliquant sur un organisme ou par colonne en cliquant sur un SeqID. Enfin, afin de comparer directement un profil à tous les autres, une fonction de tri basée sur la distance euclidienne est implémentée. La distance euclidienne mesure la distance qui sépare 2 vecteurs (2 profils). Ainsi, plus cette distance est faible plus les vecteurs sont proches. Pour 2 « Query » X et Y, la distance euclidienne entre leurs 2 profils (DE_{XY}) est calculée de la manière suivante :

$$DE_{XY} = \sqrt{\sum_i \left(X_i - Y_i\right)^2} \text{ avec i l'organisme}$$

Les profils sont ensuite triés du plus proche au plus éloigné par rapport au profil initialement choisi (Figure 55).

Figure 55 **Exemple d'un tri de profils phylogénétiques.**
6 SeqIDs ont été triés en fonction de la distance euclidienne. Cette méthode permet de regrouper les profils phylogénétiques similaires.

Toutes les fonctions de tri sont accessibles au moyen de la souris en cliquant sur l'un des objets actifs (« SeqID » ou « Organisme »). Une interface permet de configurer l'ensemble des possibilités pour chaque bouton et pour chaque objet (Figure 56). L'utilisation des trois boutons en combinaison ou non avec des modificateurs tel que « Shift » ou « Control » permet de disposer potentiellement de 9 actions pour un seul objet.

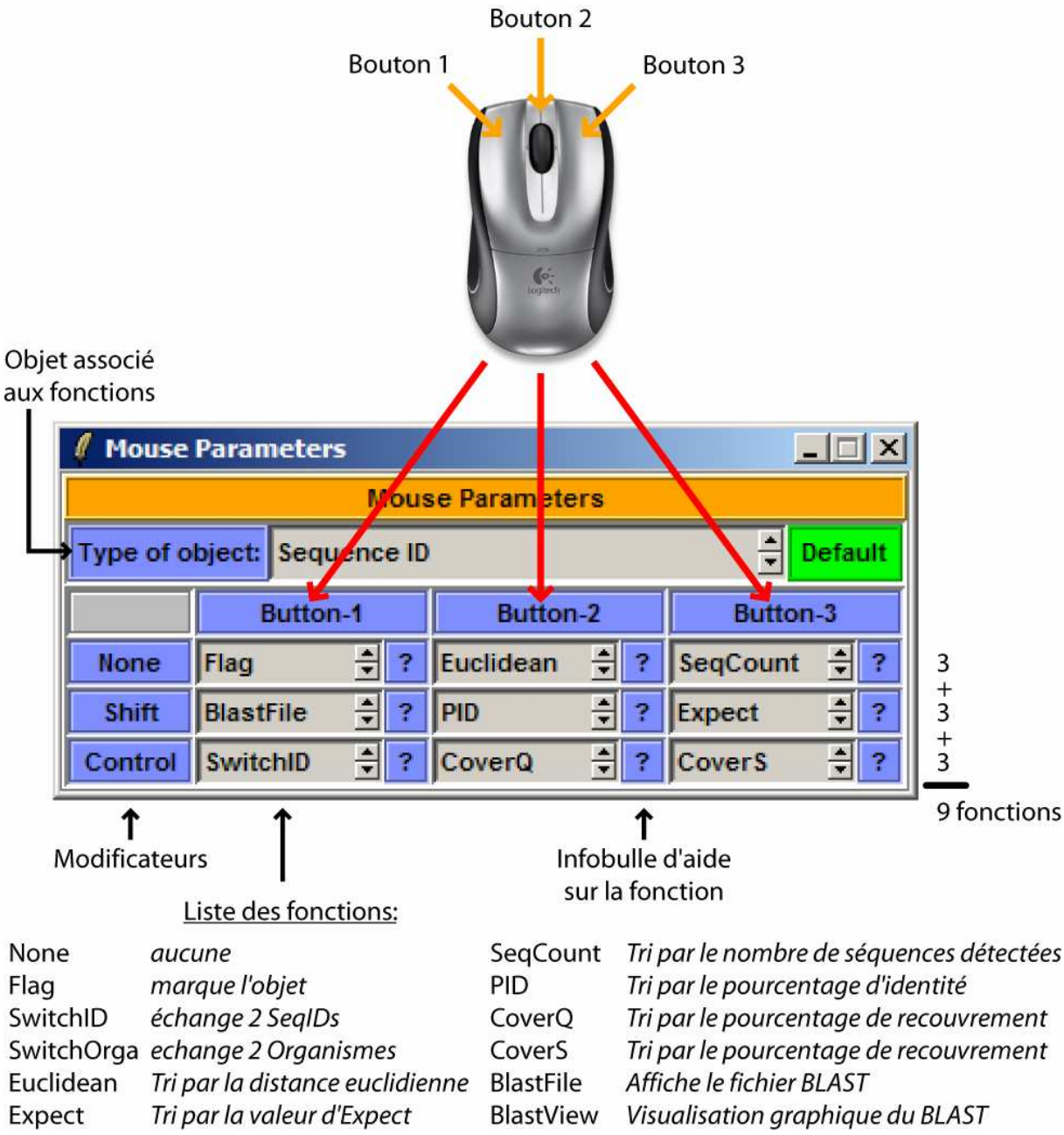


Figure 56 **Interface de gestion des actions disponibles.**

Un soin particulier a été apporté à l'interface de gestion des profils et sa capacité à soutenir une charge importante d'information notamment lors des tris et des mises à jour. En effet, un contrôleur permettant de charger toutes les informations pour chaque séquence appât et une gestion maitrisée des moments de mise à jour réduisent grandement le temps d'affichage. Nous avons également opté pour une immobilité des rectangles qui mettent automatiquement à jour leurs informations (couleur, et données à afficher) en s'adressant directement au contrôleur en fonction de leurs entêtes de ligne et de colonne. Ainsi par exemple, l'optimisation du temps de chargement permet d'un part de charger 1742 Query et les résultats associés sur un PC récent (processeur centrino 2GHz) en 1 minute environ et

d'autre part de calculer, classer et mettre à jour les 1742 profils en fonction de la distance euclidienne (par exemple) en 15s. Cependant, une telle quantité de données nécessite une quantité de mémoire vive non négligeable.

7.2.2 Informations liées à la recherche par BLAST

Afin de faciliter l'interprétationdes résultats de la recherche de similarité dans les banques de séquences, nous avons implémenté une interface de visualisation sous la forme de graphiques. L'utilisateur peut ainsi afficher pour un gène (« Query ») ou un groupe de gènes plusieurs valeurs (moyennées ou non) sur la meilleure séquence détectée dans le BLAST ou sur toutes en fonction du nom de la séquence (« SeqID ») ou des organismes (Figure 57).

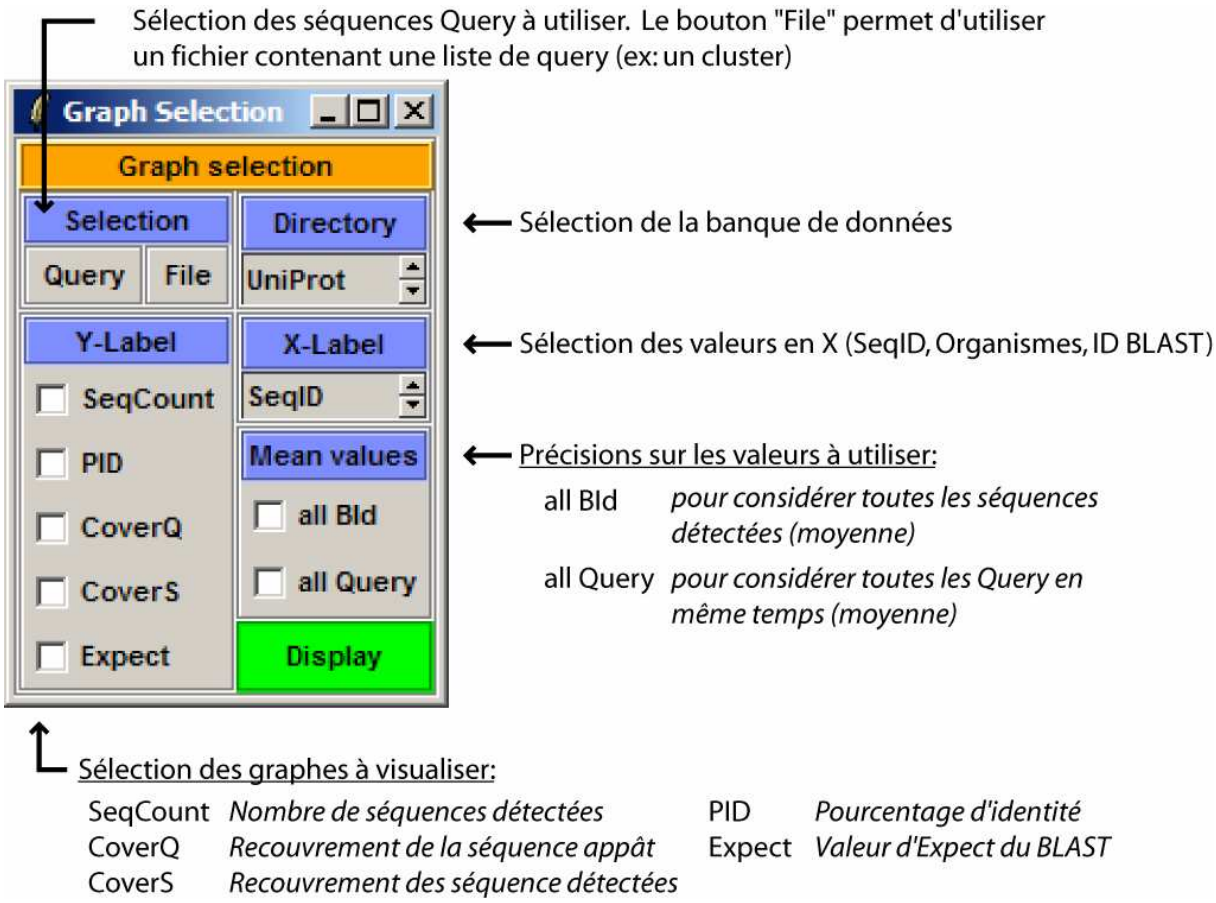
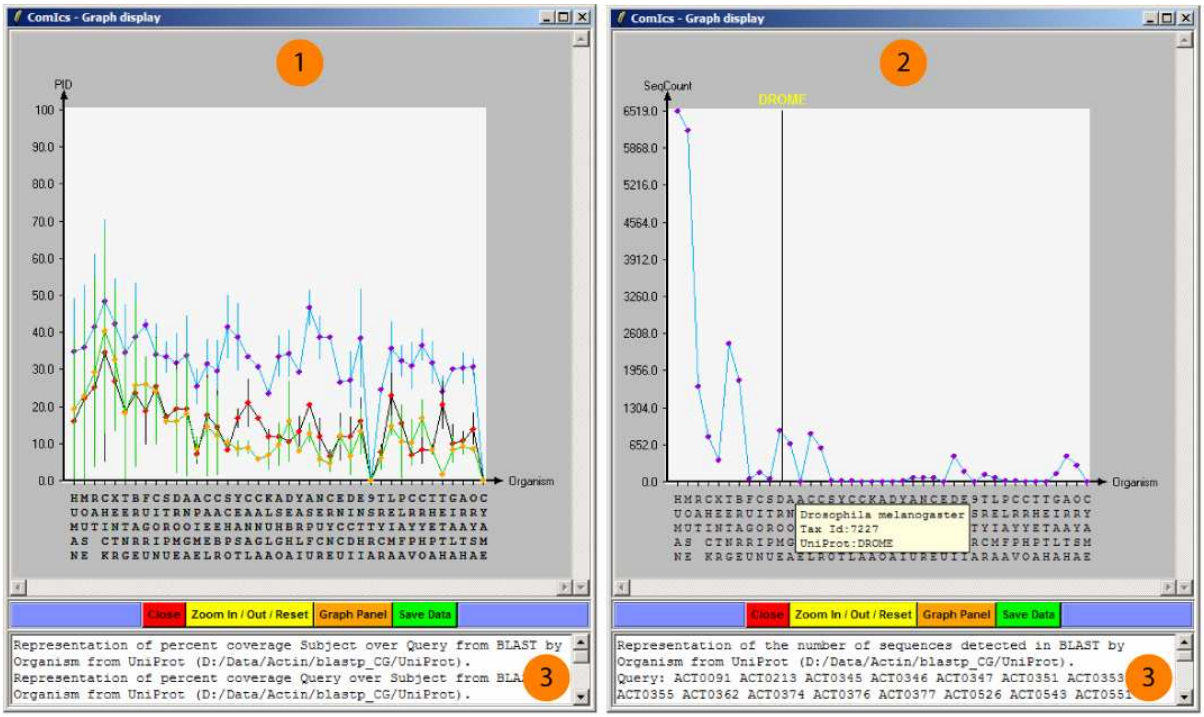


Figure 57 **Interface de sélection des graphiques à visualiser.**

Les graphiques sont ensuite affichés selon les modes choisis, par exemple les graphiques de la Figure 58 représentent la moyenne des valeurs pour un cluster de 50 séquences Query en fonction des organismes et tiennent compte pour chaque séquence Query de toutes les séquences détectées dans le BLASTP.



1 Graphique représentant le pourcentage d'identité (PID) moyen et les pourcentages de recouvrement (CoverQ et CoverS) moyens d'un ensemble de séquences détectées par organisme.

2 Graphique représentant le nombre de séquences détectées (SeqCount) par organisme

3 Fenêtre indiquant les informations sur les séquences et les méthodes utilisées (moyenne ou non, nom des séquences, quel graphe...)

Figure 58 **Interface de visualisation des données liées au BLAST.**

Dans un premier temps, cette interface ne permet que de visualiser les données, mais à terme elle permettra d'utiliser les graphiques pour sélectionner des seuils et des séquences afin, par exemple, de les inclure dans les profils ou dans un alignement multiple.

7.3 Perspectives :

7.3.1 Validation au niveau génomique

La définition des profils phylogénétiques à partir des banques de protéines est grandement dépendante de l'effort consenti pour la caractérisation et la validation des séquences provenant d'un organisme donné. Certains organismes, comme l'homme ou la souris, sont mieux représentés que l'abeille par exemple. La disponibilité des séquences génomiques est une opportunité importante pour assurer sans ambigüité la présence d'une protéine homologue dans un organisme. Néanmoins, ceci nécessite de mettre en place un protocole particulier pour la détection et la caractérisation automatique des séquences. La première partie de ce protocole, c'est-à-dire la gestion de la recherche dans les banques de séquences (1

par génome) en automatique, est déjà intégrée. Elle utilise le programme TBLASTN et les critères principaux de recherche sont disponibles (voir Figure 50 p128). L'utilisateur peut ainsi effectuer des recherches, mais il doit valider lui-même les alignements détectés.

7.3.2 Validation par l'alignement multiple

Le profil phylogénétique permet de définir les membres d'une famille de protéines présents dans chacun des organismes et constitue de ce fait une base idéale pour définir la liste des protéines à intégrer dans un alignement multiple des séquences complètes représentatif de sa famille. ComIcs sera de ce fait connecté avec l'ensemble des outils développés au laboratoire et permettra de basculer directement à l'alignement multiple. Le bénéfice sera alors réciproque car l'alignemant multiple permettra également de valider les différents homologues d'une famille.

Chapitre 8 - ARPAnno

ARPAnno est une application directe des résultats que nous avons obtenus pour mieux caractériser les Actin Related Proteins (ARPs) (Muller *et al.* 2005). Notre analyse se base sur une connaissance approfondie de l'alignement multiple des séquences complètes des homologues des ARPs et des actines. Cet alignement multiple nommé ARP-MACS est disponible pour consultation ou téléchargement via une page web (http://bips.u-strasbg.fr/ARPAnno/ARPMACS.html). Dans le but de rendre nos résultats disponibles pour la communauté scientifique et de caractériser de nouvelles séquences, nous avons développé un service web d'annotation automatique des sous-familles d'ARP appelé ARPAnno (http://bips.u-strasbg.fr/ARPAnno).

8.1 Le serveur ARPAnno

ARPAnno est un serveur web (Figure 59) qui permet à un utilisateur de tester l'appartenance de sa séquence à une des sous-familles d'ARPs. Il utilise les informations collectées et analysées au cours de l'étude approfondie de cette famille de protéines, notamment la diversité des séquences extraites des banques de protéines, la présence de résidus discriminants et la présence d'insertion ou de délétions par rapport à la séquence d'actine conventionnelle, notre référence.

Figure 59 **Capture d'écran de la page d'accueil d'ARPanno.**
Le serveur web ARPAnno est disponible à l'adresse http://bips.u-strasbg.fr/ARPAnno.

8.1.1 Fonctionnement général et conception modulaire

ARPAnno est constitué de 2 modules principaux (Figure 60) ; une partie dite « client » qui représente toute la partie visible (la partie web) et utilisable par l'utilisateur et une partie « serveur » qui contient les fonctions d'analyse du programme. Cette conception modulaire permet d'utiliser la partie serveur comme un programme à part entière et autorise donc son intégration dans des scripts pour une utilisation à haut débit.

La partie « client » est constituée d'une page html et d'un script cgi. Le cœur fonctionnel d'ARPAnno est écrit en script Tcl/Tk et fait appel à 2 programmes externes BLASTP et ClustalW. Il utilise également 3 fonctions développées par Julie Thompson (aln_pos, aln_res et aln_insert) qui remplissent la fonction d'analyseur syntaxique ou « parser » des alignements multiples.

ARPAnno fonctionne de la manière suivante (Figure 60) ; un utilisateur soumet une séquence protéique au serveur web installé sur *titus* (partie « client ») (formulaire ARPAnno.html), le script cgi (ARPAnno.cgi) exécute le programme d'analyse (voir 8.1.2 Protocole du serveur ARPAnno) sur le serveur de calcul (*Beaufort*). Le résultat est ensuite

retourné au script cgi qui génère à la volée une page html (Results.html) contenant le résultat des prédictions ainsi que des liens vers les éventuels BLASTP et alignements multiples.

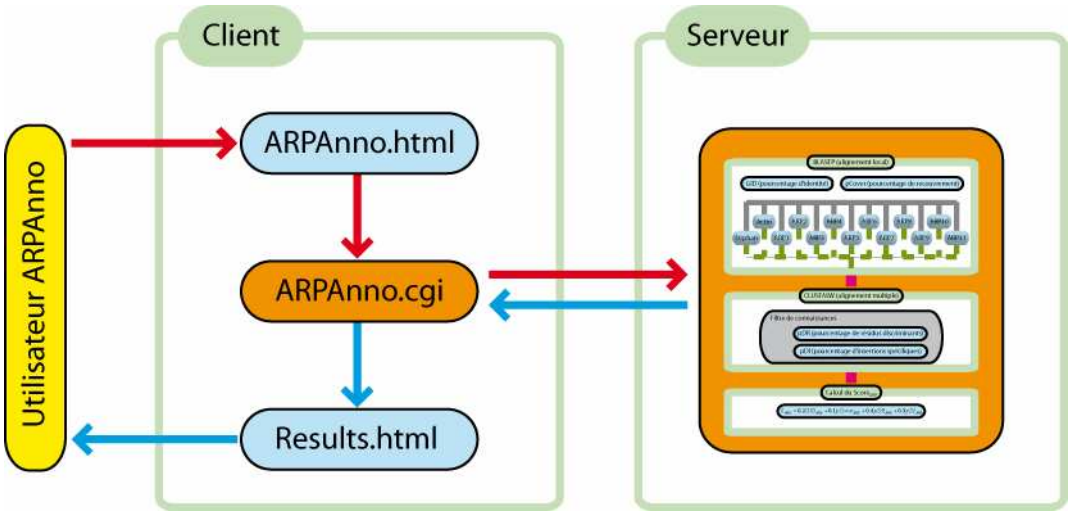


Figure 60 **Schéma décrivant l'organisation du serveur ARPAnno.**
Le détail de la partie « Serveur » est indiqué dans la Figure 61. Les parties exécutées sont indiquées en orange, la requête soumise par l'utilisateur est indiquée par des flèches rouges et la réponse générée par ARPAnno par des flèches bleues.

8.1.2 Protocole du serveur ARPAnno

ARPAnno est un programme informatique développé en script Tcl/Tk et en langage C ANSI pour certaines fonctionnalités. Il utilise également les programmes BLAST et ClustalW. La stratégie utilisée dans ARPAnno est divisée en 3 étapes (Figure 61) :

- Premièrement, ARPAnno aligne la séquence requête contre chacune des sous-familles présentes dans ARP-MACS (actine, 11 sous-familles d'ARPs et les orphelins) au moyen du programme BLASTP (banques BLAST dédiées). ARPAnno utilise 2 critères pour filter les sous-familles et sélectionner celles qui passeront à la seconde étape. Le premier critère est le pourcentage d'identité global (GID) d'une séquence contre une autre, qui est calculé comme le ratio entre le nombre total de résidus identiques et le nombre total de résidus dans tous les HSPs (*High Scoring Pairs* voir 2.3.2.2 BLAST) de la séquence requête. Le deuxième critère est le pourcentage de recouvrement (pCover) qui est défini par le ratio entre le nombre de résidus identiques et le nombre de résidus qui auraient pu être alignés entre les 2 séquences.

- Deuxièmement, la séquence requête est alignée contre les sous-familles choisies et présentes dans ARP-MACS avec ClustalW en mode profile. L'alignement multiple obtenu est ainsi filtré en utilisant les critères discriminants définis lors de l'analyse approfondie de l'alignement complet des ARPs (résidus discriminants et les insertions délétions) (Muller *et al.* 2005). Pour chaque sous-famille choisie, 2 nouveaux scores sont ainsi calculés : le pourcentage d'insertions discriminantes détectées (pDI) et le pourcentage de résidus discriminants détectés (pDR) pour la sous-famille.
- Un score final compris entre 0 et 100 est calculé pour chaque sous-famille en combinant l'ensemble des scores calculés dans les étapes précédentes.

Les poids attribués pour chaque score sont définis empiriquement afin d'optimiser la séparation des sous-familles d'ARP définies dans ARP-MACS. L'alignement de la séquence requête contre chaque sous-famille choisie est disponible au format XML et/ou MSF. Chaque élément discriminant est colorié de façon différente.

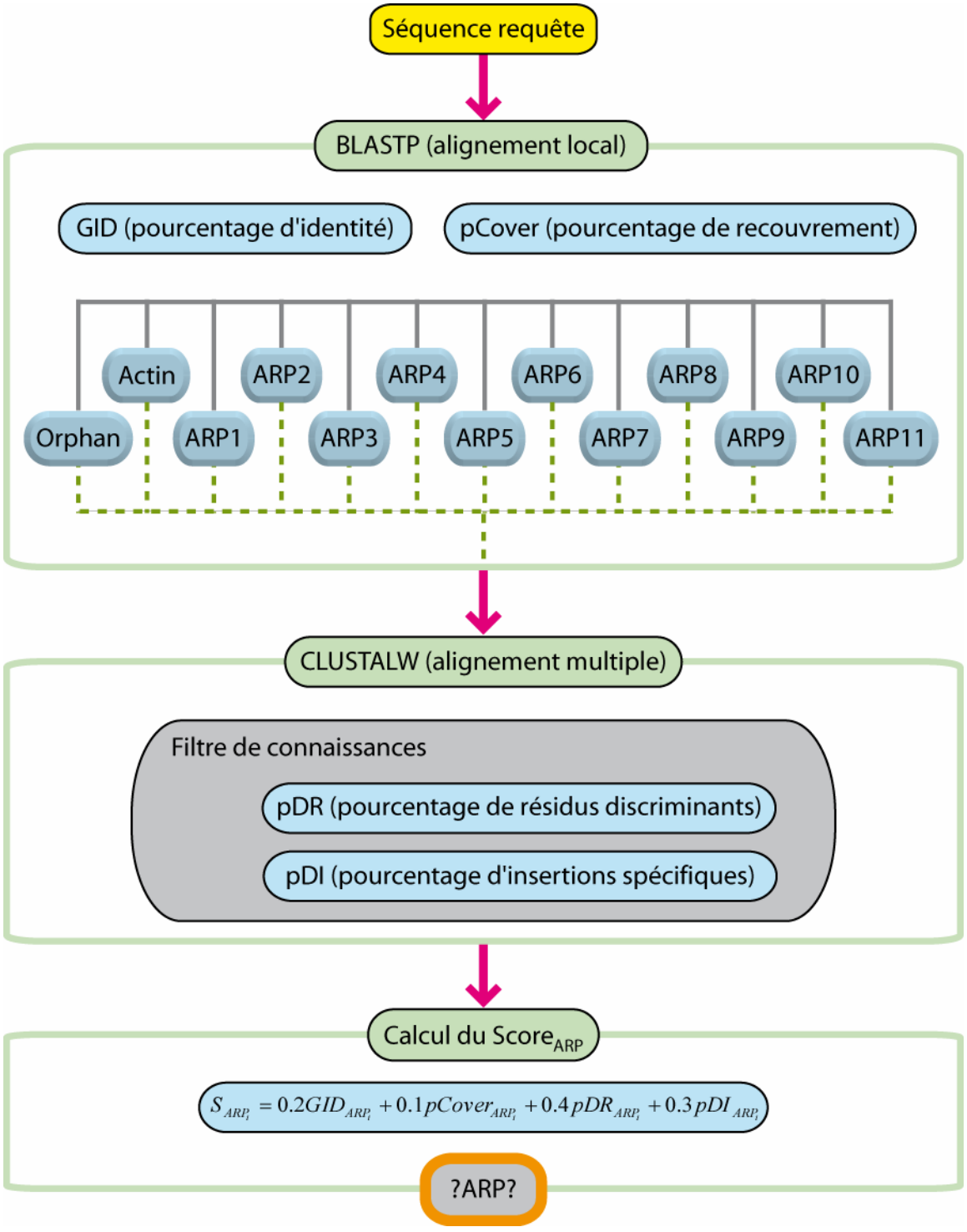


Figure 61 **Schéma représentant le protocole du serveur ARPAnno.**
Les 3 grandes étapes sont représentées ; l'alignement local et le calcul des 2 premiers critères de sélection sur chaque sous-famille testée, l'alignement global sur les meilleures sous-familles est filtré en fonction des critères de connaissance, enfin le calcul du score final pour chaque sous-famille.

Chapitre 9 - CADO4MI

Dans la conception d'une puce à ADN, il existe un nombre important d'étapes essentielles. Le choix de sondes ou oligonucléotides spécifiques représente une telle étape puisqu'elle détermine la spécificité de détection de cette puce. Le design de sondes est également une des étapes de conception faisant appel à la bioinformatique. Il consiste à définir des régions complémentaires de la séquence des ARNm que l'on veut détecter afin qu'elles possèdent des caractéristiques propres à une hybridation spécifique et homogène pour l'ensemble de la puce.

CADO4MI pour « *Computer Assisted Design of Oligonucleotide For Microarray* » est un programme écrit en Tcl/Tk pour réaliser le design de sondes pour puces à ADN. Ce développement a été rendu nécessaire par la présence de gènes dans le cytosquelette partageant une identité de séquences très forte, susceptible de générer des hybridations croisées.

Nous verrons dans un premier temps, les différents éléments qui permettent de définir une sonde pour puce à ADN et dans un deuxième temps, le programme et son utilisation.

9.1 La sonde et le transcrit

Le principe des puces à ADN repose sur l'hybridation entre 2 acides nucléiques, l'un étant défini comme la cible ou « *target* », l'autre étant défini comme la sonde ou « *probe* » (4.3.2.1 Les sondes spécifiques de gènes).

Une sonde est une séquence d'acide nucléique de taille variable (25 à 2000 bases) en fonction du type de puce à ADN utilisée. Elle doit être complémentaire d'une partie de la séquence cible à détecter. Dans notre cas, nous nous somme focalisés sur des sondes de tailles courtes ou moyennes (entre 25 et 60 bases).

La cible est en général l'ARNm d'un gène que l'on cherche à détecter, après transformation de l'ARNm en ADNc marqué.

9.2 La spécificité

La spécificité d'un oligonucléotide pour un gène X est définie par la capacité de l'oligonucléotide à s'hybrider seulement au gène X. L'hybridation croisée correspond aux hybridations établies entre l'oligonucléotide et d'autres séquences que celle du gène X. Selon Kane *et al.* (Kane *et al.* 2000), ce type de signal peut être évité, ou du moins réduit au bruit de fond, en choisissant des oligonucléotides qui ne possèdent pas une similarité trop importante avec d'autres séquences. Les limites de similarité ont été déterminées expérimentalement pour des sondes de 50 bases de long par Kane *et al*. (Kane *et al.* 2000) et de manière similaire pour les 60mer par Hughes *et al.* (Hughes *et al.* 2001). Elles sont ainsi exprimées selon 2 règles dites des « règles de Kane » :

- Pourcentage d'identité entre la sonde et une autre séquence inférieur à 75-80%
- Nombre de nucléotides successifs identiques avec une séquence autre que la séquence cible inférieur à 15 bases

9.3 Température de fusion

La température de fusion ou Tm (*Temperature of melting* ou plus exactement *Temperature of mid-transition*) est une mesure directe de la stabilité de l'association de deux brins d'acides nucléiques en solution (Figure 62). Il s'agit de la température à laquelle la moitié des acides nucléiques est sous forme double brin. En l'absence d'agents déstabilisants, comme la formamide ou l'urée, le Tm dépend principalement de la nature des acides nucléiques, du taux de GC, de la concentration en acides nucléiques et de la concentration en sels. Ces trois facteurs augmentent en effet le Tm en stabilisant l'hybride formé.

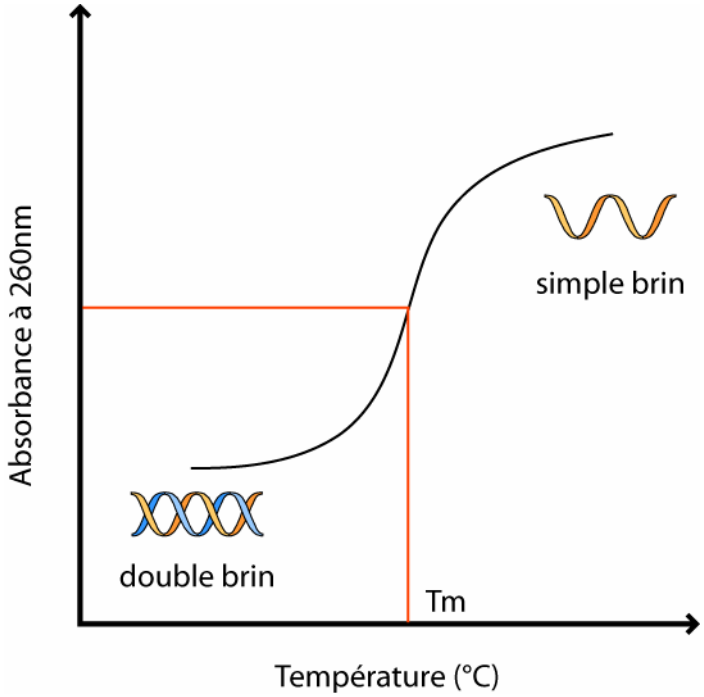


Figure 62 **Exemple de la courbe de fusion obtenue pour un oligonucléotide.**
La dénaturation d'une molécule d'ADN s'accompagne d'une augmentation de l'absorption lumineuse à 260 nm appelée aussi effet hyperchrome. Cette dénaturation s'effectue généralement dans une zone de température restreinte dont le point médian correspond à la température de fusion ou Tm.

Le calcul du Tm est un critère important pour le design de sondes pour puces à ADN. En fonction de la taille de la molécule il existe plusieurs méthodes pour calculer le Tm. Il est important de noter que l'ensemble de ces formules sont développées pour des acides nucléiques en solution or dans notre cas (puces à ADN) l'un des partenaires (la sonde) est fixé sur un support. En dépit de cette approximation et du manque d'un modèle spécifique aux acides nucléiques fixés sur support et dédiés aux puces à ADN, il est important d'utiliser une seule méthode et de choisir un Tm optimal pour un set d'oligonucléotides. Un set d'oligonucléotides ayant une distribution de Tm proche disposera de propriétés d'hybridation similaires et garantira une hybridation homogène pour l'ensemble de la puce. L'ensemble des méthodes décrites ci-dessous sont implémentées dans CADO4MI.

9.3.1 Modèle thermodynamique du plus proche voisin

Actuellement, le modèle le plus complet pour étudier les rendements d'hybridation des sondes sur les cibles est le modèle thermodynamique dit du plus proche voisin (*Nearest Neighbor Model*). Au sein d'une molécule d'ADN double brin, ce modèle définit des interactions de voisinage entre deux nucléotides successifs qui permettent de prendre en compte à la fois la nature et la place des nucléotides. Ces paires de nucléotides sont associées à des valeurs distinctes d'enthalpie, d'entropie et d'énergie libre pour l'association des deux

brins d'ADN (Doktycz *et al.* 1995). En 1998, SantaLucia (SantaLucia 1998) a proposé des valeurs unifiées de ces énergies (Tableau 8) calculées d'après l'ensemble des études menées auparavant (Gotoh *et al.* 1981; Vologodskii *et al.* 1984; Breslauer *et al.* 1986; Delcourt *et al.* 1991; Doktycz *et al.* 1992; SantaLucia *et al.* 1996; Sugimoto *et al.* 1996; Allawi *et al.* 1997). Des études complémentaires ont montré que le marquage des nucléotides ne modifie pas la valeur de ces énergies. Ce modèle thermodynamique peut donc s'appliquer à l'analyse des hybrides formés entre sondes et cibles marquées lors d'expérience de puces à ADN (Griffin *et al.* 1998). Ainsi le Tm est calculé par la formule suivante :

$$T_m = \frac{1000 \times \Delta H}{\Delta S + R \times \ln\left(\frac{[DNA]}{4}\right)} - 273.5$$

R représente la constante des gaz parfaits (R=1,987 cal/K.mol), [DNA] la concentration en ADN (défaut à $1e^{-6}$ M), ΔH est la somme des enthalpies et ΔS est la somme des entropies.

En accord avec le modèle du plus proche voisin, le calcul des ΔH et ΔS peuvent être décomposés en plusieurs termes (Breslauer *et al.* 1986 ; Sugimoto *et al.* 1996; SantaLucia 1998) :

$$\Delta H_{total} = \sum_i \Delta H_i + \Delta H_{initialisation5'} + \Delta H_{initialisation3'} + \Delta H_{symétrie}$$

$$\Delta S_{total} = \sum_i \Delta S_i + \Delta S_{initialisation5'} + \Delta S_{initialisation3'} + \Delta S_{symétrie}$$

Séquence	ΔH kcal/mol	ΔS kcal/mol
AA/TT	-7.9	-22.2
AT/TA	-7.2	-20.4
TA/AT	-7.2	-21.3
CA/GT	-8.5	-22.7
GT/CA	-8.4	-22.4
CT/GA	-7.8	-21.0
GA/CT	-8.2	-22.2
CG/GC	-10.6	-27.2
GC/CG	-9.8	-24.4
GG/CC	-8.0	-19.9
Initialisation GC	0.1	-2.8
Initialisation AT	2.3	4.1
Symétrie	0	-1.4

Tableau 8 Paramètres thermodynamiques unifiés décrits par SantaLucia.
Ces paramètres sont utilisés dans le calcul du Tm avec modèle du plus proche voisin pour 1M de NaCl (SantaLucia 1998).

La table des paramètres thermodynamiques est chargée en mémoire par un contrôleur et le programme CADO4MI y accède de la manière suivante :

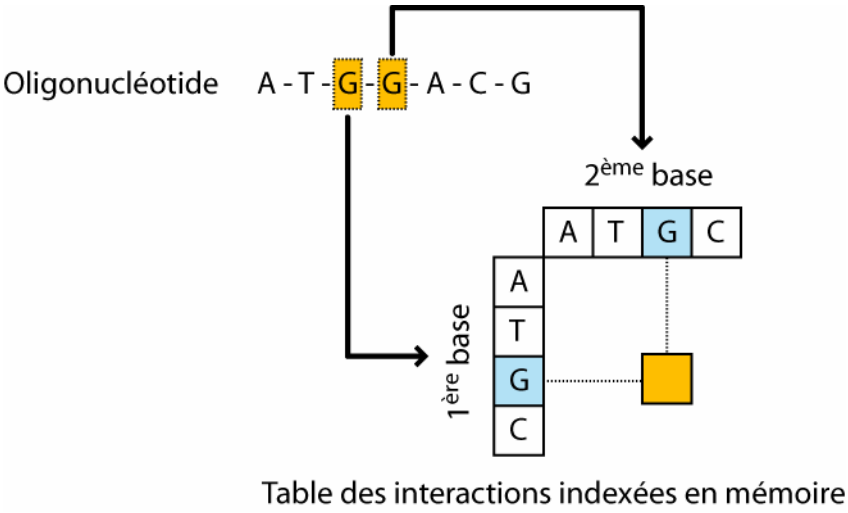


Figure 63 Méthode de lecture d'un couple de bases dans la séquence d'un oligonucléotide dans la table contenant les données thermodynamiques d'énergie d'interaction.
Sur la séquence, pour chaque couple de bases contiguës, la première base est lue en ligne et la deuxième base en colonne.

Cette méthode permet de calculer des valeurs pour le Tm de manière efficace et rapide.

9.3.2 La règle de « Wallace »

Parmi les méthodes de calcul du Tm qui sont les plus connues, la méthode de calcul du Tm selon Wallace est certainement l'une des plus utilisées par les biologistes (Wallace *et al.* 1979). Elle est cependant limitée aux sondes d'une longueur inférieure à 30 bases et dans le cas contraire, surprédit la valeur de Tm.

$$T_m = 2\times(A+T)+4\times(G+C)$$

(A+T) représente la somme du nombre de bases A ou T dans la séquence et (G+C) la somme des bases G ou C.

9.3.3 Autres méthodes

D'autres méthodes ont également été décrites qui sont plus adaptées aux oligonucléotides de grande taille (>70mer) et tiennent compte du pourcentage de GC, de la concentration en acides nucléiques, de la longueur de la séquence et du pourcentage de formamide (Bolton *et al.* 1962; Casey *et al.* 1977; Howley *et al.* 1979; Meinkoth *et al.* 1984; Bodkin *et al.* 1985). Elles sont différentes en fonction du type de molécules en présence :

- ADN-ADN

$$T_m = 81.5 + 16.6 \times (\log M) + 41 \times (\%GC) - 0.62(\%F) - 500/S$$

- ADN-ARN

$$T_m = 79.8 + 18.5 \times (\log M) + 58.4 \times (\%GC) + 11.8 \times (\%GC)^2 - 0.50(\%F) - 820/S$$

- ARN-ARN

$$T_m = 79.8 + 18.5 \times (\log M) + 58.4 \times (\%GC) + 11.8 \times (\%GC)^2 - 0.35(\%F) - 820/S$$

M représente la concentration molaire de sel (défaut 1 M), GC le pourcentage de bases GC dans la séquence, S la longueur de la séquence et F le pourcentage de formamide.

9.4 Autres critères d'intérêt pour le choix des sondes

D'autres paramètres, moins importants que la valeur du Tm, peuvent influençer la stabilité de l'hybridation.

9.4.1 Distance de l'extrémité 3' du transcrit

La distance à l'extrémité 3' du transcrit est un critère simple, mais extrêmement important, dicté par le protocole de fabrication des ADNc. En effet, la production des ADNc marqués et déposés plus tard sur la puce nécessite souvent une étape de transcription inverse à partir d'amorces poly-T. La taille des transcrits ainsi générés est dépendante du rendement de l'enzyme utilisée. Par exemple, si les ADNc marqués ne dépassent pas la taille de 2000 bases et que, pour l'un des gènes (d'une taille supérieure à 2000 bases) présent sur la puce, une sonde a été choisie au-delà de cette limite, la sonde n'aura aucune chance de détecter sa cible pourtant bien présente.

9.4.2 Taux de GC

La stabilité d'un ADN double brin dépend du nombre de liaisons hydrogène formées. Les hybridations impliquant des séquences riches en GC sont donc plus stables (von Ahsen *et al.* 1999). En pratique, un taux de GC compris entre 40 et 65 % semble assurer un rendement d'hybridation optimal et limite les risques d'hybridation aspécifique (Luebke *et al.* 2003; Talla *et al.* 2003).

9.4.3 Séquence prohibées

Les séquences prohibées sont des séquences qui ne doivent pas être trouvées dans les sondes designées comme, par exemple, des sites de restriction ou une succession de bases identiques (par exemple GGGGGGG) qui peuvent poser problème notamment lors de sa synthèse.

9.5 Le protocole de CADO4MI

Le protocole de design de sondes est séparé en 4 étapes décrites dans les paragraphes suivants et illustrés par la Figure 64.

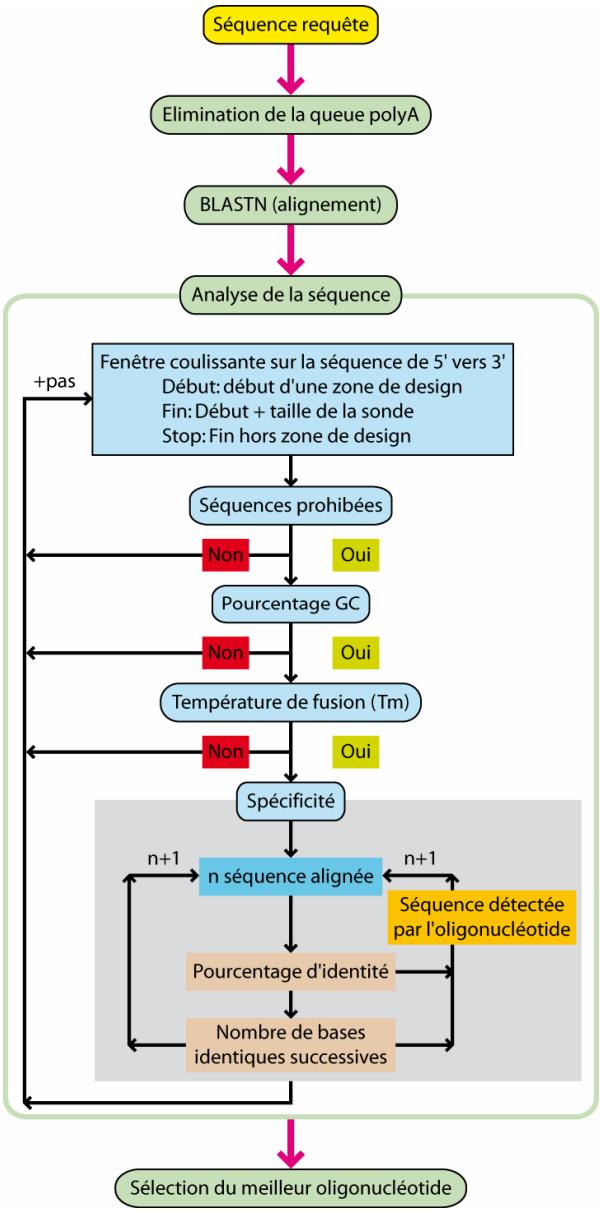


Figure 64 **Schéma du protocole du programme CADO4MI pour le design de sondes.**
Les 4 grandes étapes sont indiquées dans les cartouches vertes.

9.5.1 L'élimination de la queue poly-A

Le masquage des séquences est une étape qui consiste à remplacer une ou plusieurs bases (ou acides aminés) de la séquence par un caractère (« N » ou « X » en général) pour qu'elles ne soient pas traitées par les programmes, dans notre cas BLAST. Pour le design de sondes, nous utilisons des ARNm qui possèdent le plus souvent des queues poly-A en 3' de la séquence. Dans le but d'utiliser le maximum des séquences cibles pour choisir des sondes, les résultats de BLAST ne sont pas filtrés. Dans ce cas, un ARNm poly-A est potentiellement capable de détecter tous les autres ARNm poly-A de la banque de donnée. Ces régions nécessitent un traitement particulier. Nous avons ainsi remplacé toutes les queues poly-A d'une longueur supérieure à 15 bases par des « N ».

9.5.2 Recherche de similarité : BLASTN

La recherche de similarité entre la séquence appât et les séquences stockées dans les banques de séquences nucléiques est réalisée au moyen du programme BLASTN. La recherche de similarité nous permet d'obtenir des alignements entre les séquences appâts et les séquences tests qui sont analysés pour caractériser la spécificité des oligonucléotides potentiels (voir ci-dessous). Les paramètres usuels de BLASTN comme la valeur d'Expect, le nombre de séquences que l'on veut aligner au maximum, la taille des mots et l'option de filtrage sont paramétrables dans l'interface de CADO4MI.

9.5.3 Analyse de séquence

L'analyse de la séquence appât complète est effectuée de manière séquentielle à partir d'une « fenêtre coulissante » déplacée le long de la séquence. La fenêtre possède une longueur correspondant à la taille des oligonucléotides designés (dans notre cas 60 bases) et elle est déplacée selon un pas de 10 bases (valeur par défaut).

Le programme évalue dans un premier temps, les zones de design (s'il y a des zones spécifiques choisies par l'utilisateur), puis procède à l'analyse en 4 étapes :

- Premièrement, la séquence contenue dans la fenêtre est testée pour la présence de séquences prohibées (séquence choisie par l'utilisateur comme ne devant pas se trouver dans les oligonucléotides, i.e. GGGGGG). Dans le cas de la présence de telles séquences, l'oligonucléotide est rejeté et la fenêtre coulissante est déplacée.

- Deuxièmement, le pourcentage de bases GC de la séquence contenue dans la fenêtre est comparé à l'intervalle spécifié par l'utilisateur. Si l'intervalle n'est pas respecté, l'oligonucléotide est rejeté et la fenêtre coulissante est déplacée.
- Troisièmement, le Tm de la séquence contenue dans la fenêtre est calculé et comparé à l'intervalle des températures spécifié par l'utilisateur. Si l'intervalle n'est pas respecté, l'oligonucléotide est rejeté et la fenêtre coulissante est déplacée.
- Quatrièmement, la spécificité de chaque séquence de la banque détectée par la recherche BLASTN est calculée et filtrée au moyen des 2 règles de Kane. Ainsi pour chacune des séquences alignées le pourcentage d'identité et le nombre de bases successives identiques sont calculés. Pour chaque oligonucléotide potentiel, on obtient une liste filtrée ne contenant que les séquences capables de potentiellement s'hybrider à cette sonde sur la puce (0, 1 ou plusieurs séquences).

9.5.4 Sélection du meilleur oligonucléotide

Parmi la liste des oligonucléotides générés par les étapes précédentes, CADO4MI propose une solution basée sur le nombre de séquences potentiellement détectées par la future sonde et sa distance à l'extrémité 3' du transcrit. Il choisit ainsi l'oligonucléotide ayant le moins de séquences détectées et le plus proche du 3'. Une autre sélection appellée « Selection croisée » permet de combiner la selection en parallèle sur 2 résultats de banques.

9.6 Les fichiers d'entrée et de sortie

9.6.1 Les fichiers d'entrée

D'une part, CADO4MI requiert des séquences appâts qui correspondent à un ou plusieurs fichiers contenant des séquences au format FASTA. Ces fichiers contiennent une ou plusieurs séquences correspondant aux gènes d'intérêt, c'est-à-dire les séquences que l'utilisateur souhaite représenter par des sondes sur la puce.

D'autre part, l'évaluation de la spécificité des oligonucléotides par CADO4MI nécessite le choix d'un programme de recherche de similarité et de l'utilisation de banques au format approprié. Nous avons fait le choix de BLASTN, car ce programme est rapide et parce que nous disposons des programmes de BlastPanel pour analyser et décomposer les résultats obtenus. Son utilisation requiert des banques au format BLAST. Les banques utilisées pour tester la spécificité des oligonucléotides doivent contenir l'ensemble des séquences

exprimées par les cellules d'un organisme. Nous avons dans notre cas utilisé les banques RefSeq et UniGene pour l'homme.

9.6.2 Fichier de sortie

CADO4MI génère plusieurs fichiers de sortie. Ces fichiers correspondent à chacune des étapes du design et permettent de contrôler les résultats obtenus (Tableau 9).

Etape ou fonction	Extension	Format
1- Elimination queue poly-A	.masked	FASTA
2- Recherche de similarité	.blastn	BLAST
3- Analyse de séquence	.log	Texte
	.oligo	FASTA
4- Selection de la sonde	.selection	FASTA
Temporaire	.working	Texte
-	.parameters	Texte séparé par des tabulations
Résulat global	-	Texte séparé par des tabulations
Sélection croisée	.croisee	FASTA

Tableau 9 Les différents types de fichiers générés par CADO4MI.

Le fichier de la séquence appât est sauvegardé après l'élimination de la queue poly-A et correspond au fichier « .masked ». Le fichier « .blastn » contient le résultat de la recherche de similarité et peut être réutilisé plusieurs fois (cas de redesign). Le fichier « .log » contient le suivi et le résultat des analyses (Tm, %GC, spécificité…). Les séquences des oligonucléotides sont stockées dans le fichier « .oligo ». Le fichier « .selection » contient la séquence de l'oligonucléotide retenue. Le fichier « .working » est un fichier temporaire qui permet de savoir si le design pour une séquence appât donnée est toujours en cours. Tous les paramètres du design sont stockés dans un fichier « .parameters ».

De plus, l'utilisateur peut générer 2 autres types de fichiers de résultats. Un premier type de fichier correspond à une sélection de sondes combinant le résultat de 2 designs en même temps (« .croisee »). Cette sélection particulière est appelée « Selection croisée ». Elle a par exemple été utilisée pour le design des sondes de la puce Actichip en combinant les résultats des banques RefSeq et Unigene (14.2.2 Les paramètres du design). Le second type est un fichier de résultat global qui contient les informations d'un design pour un ensemble de séquences (un set d'oligonucléotide). Ce dernier est un fichier dont le contenu est séparé par des tabulations (ou TSV pour *Tab Separated Value*), qui peut être lu dans un tableur.

9.7 Une interface graphique

9.7.1 Philosophie du programme

A l'image de ComIcs, CADO4MI est également organisé autour de la notion de projet. Ainsi, l'interface centrale (Figure 65) permet de définir un projet de design, d'effectuer les actions choisies ou d'accéder à la visualisation des résultats. Le panneau central est également accompagné d'un panneau de log qui permet de visualiser les paramètres stockés et définis par l'utilisateur, mais également, de lui afficher certaines informations.

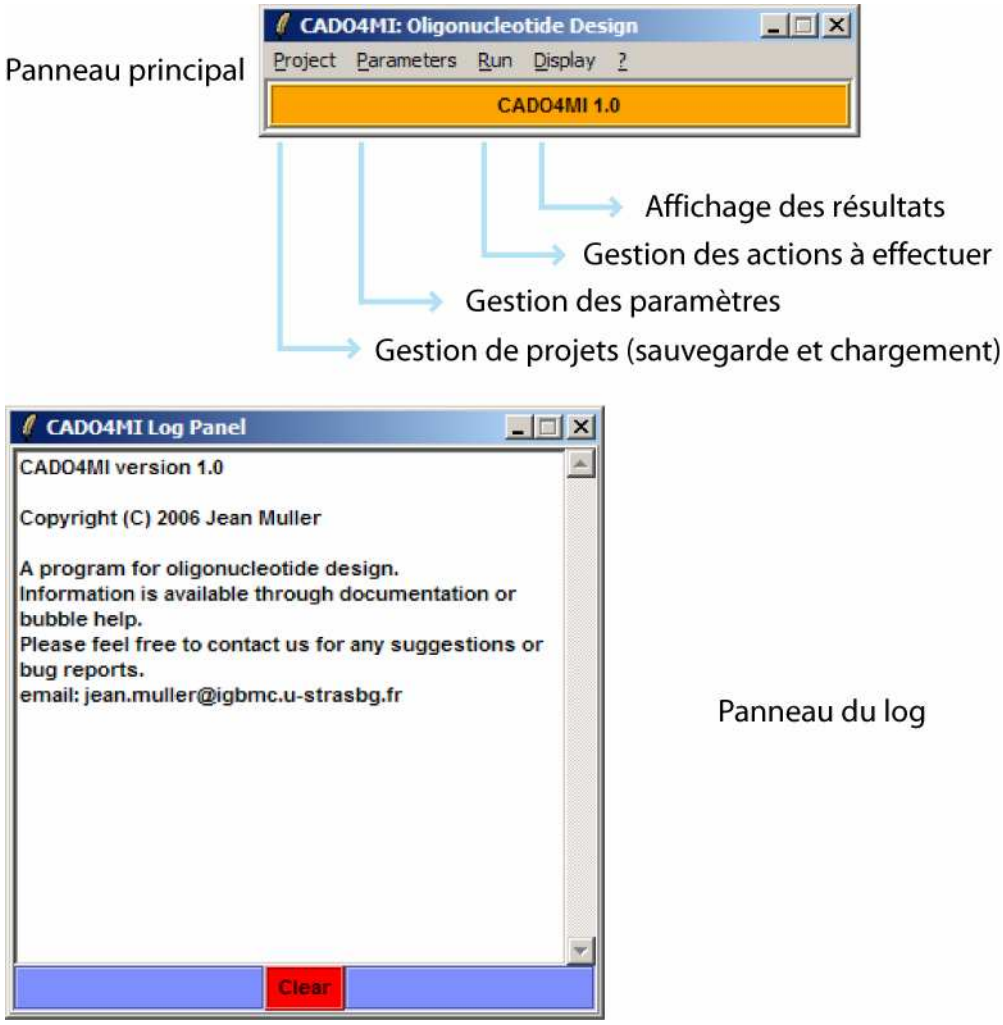


Figure 65 **Panneau central de CADO4MI.**

Un projet CADO4MI est défini à partir de séquences appâts et des critères de design choisis par l'utilisateur. Les paramètres peuvent être sauvés dans un fichier « .parameters ». Ce fichier peut être chargé ultérieurement et permet, par exemple, de scinder l'exécution des taches dans le temps, pour une analyse ultérieure des données ou encore pour refaire un

design (en changeant par exemple l'un ou l'autre des paramètres). L'ensemble des paramètres est accessible au travers du panneau principal (Figure 66).

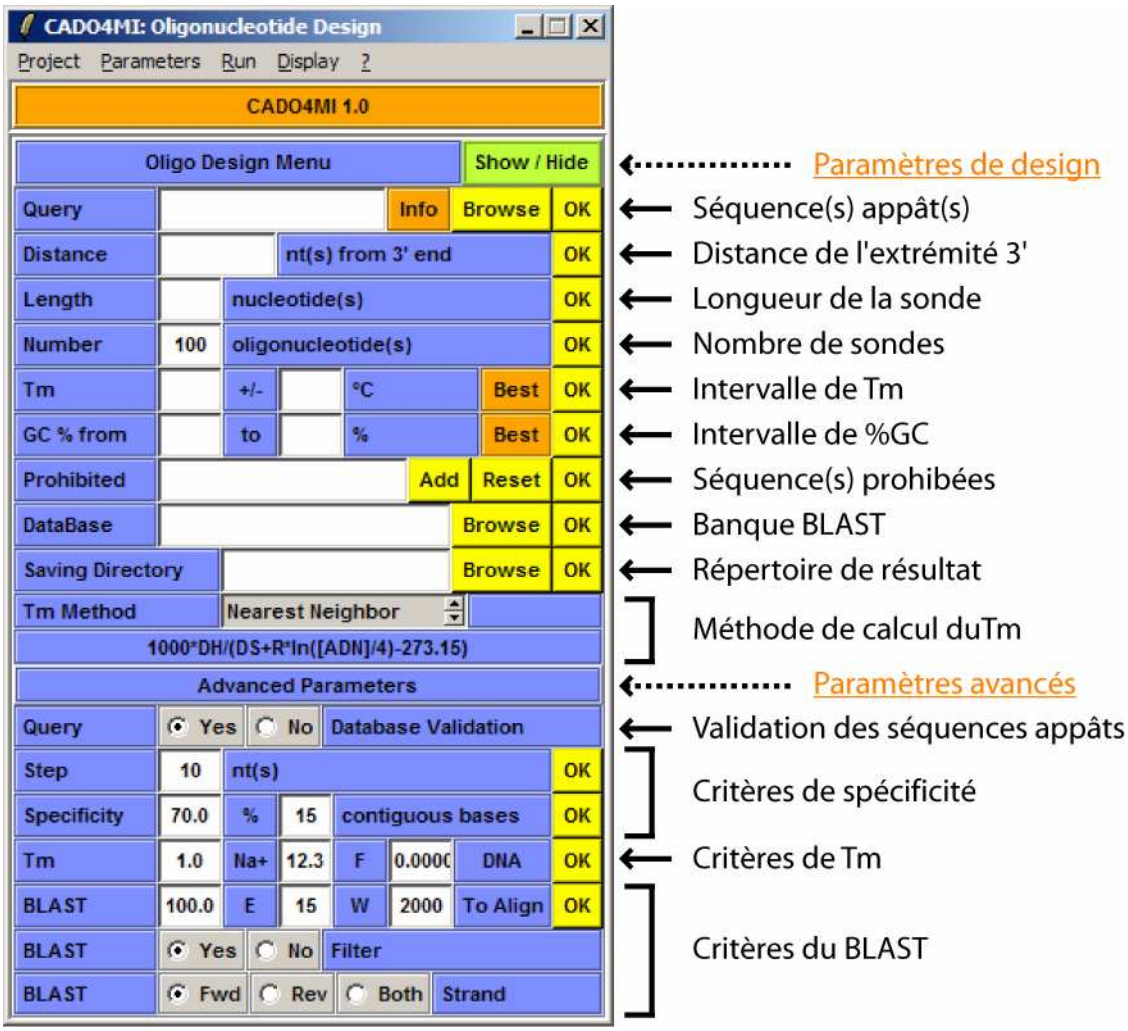


Figure 66 **Panneau de configuration de CADO4MI.**
Le panneau central de CADO4MI est représenté avec l'affichage des tous les volets de sélection des paramètres de design.

Afin de mieux appréhender la ou les séquences pour un design, chaque séquence peut être visualisée dans une fenêtre d'information qui permet de vérifier les critères de base de la séquence comme sa composition en nucléotides (A, T, G, C, pyrimidine et purine) et son Tm global (Figure 67). Cette fenêtre permet également de définir graphiquement, et donc simplement, des zones de design restreintes.

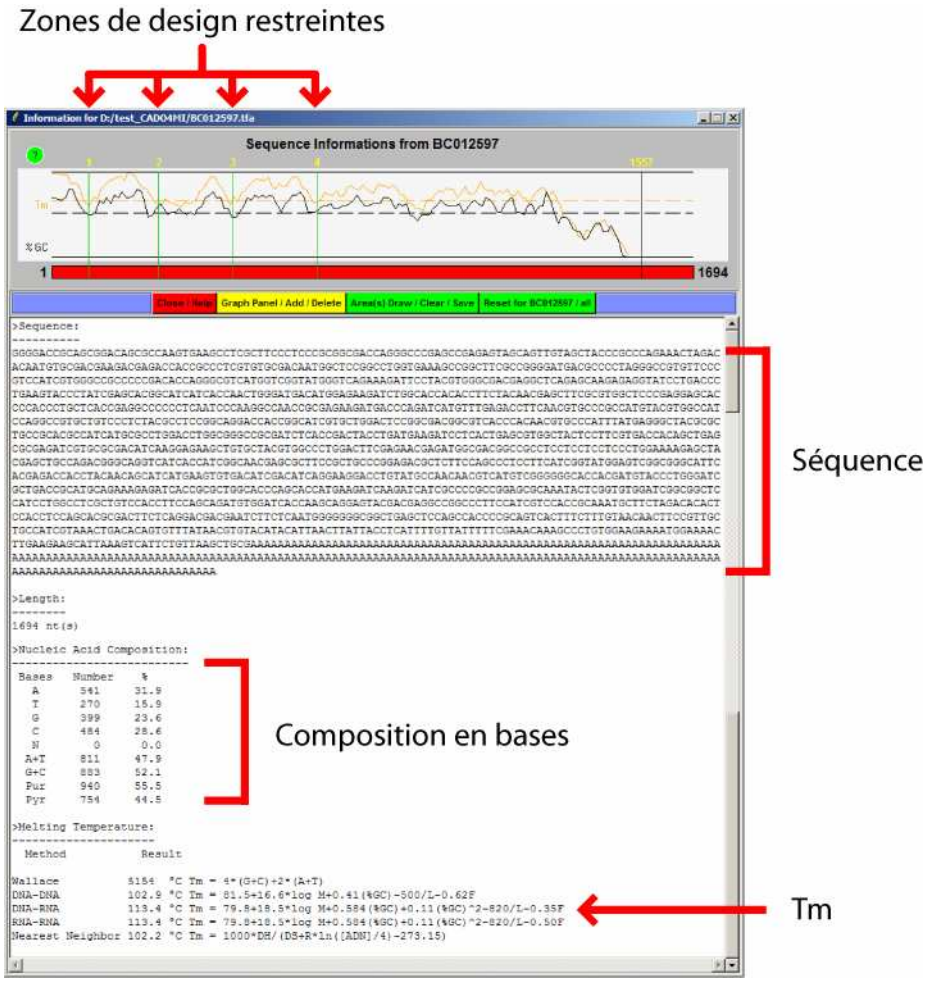


Figure 67 Fenêtre d'information d'une séquence soumise à CADO4MI.
Elle permet de visualiser la séquence, sa composition en bases, les Tm sur la taille totale ainsi que les profils associés à une certaine fenêtre. Des zones de design restreintes peuvent être définies par l'ajout et la sauvegarde de point.

9.7.2 Fenêtre de résultats

J'ai implémenté dans CADO4MI une interface graphique de visualisation des résultats de design afin de permettre une meilleure compréhension des oligonucléotides trouvés et faciliter le traitement des cas problématiques (Figure 68). Cette interface est dynamique et permet d'afficher l'ensemble des sondes calculées et d'accéder à des fonctions ou aux détails du design.

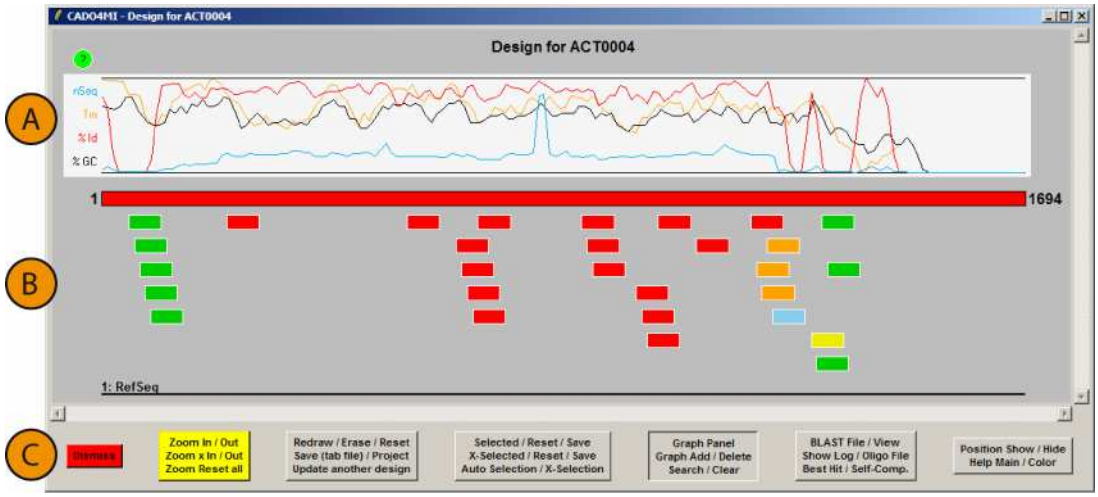


Figure 68 **Interface de visualisation des résultats de design.**
La fenêtre de résutat est composée de 3 parties centrées autour de la séquence appât (rectangle rouge avec bordure noire) ; une partie (A) pour l'affichage des profils (dans cet exemple le Tm est représenté en orange, le %GC en noir, le %ID en rouge et le nombre de séquences détectées en bleu), une partie (B) pour la séquence requête et le placement des oligonucléotides potentiels et une partie (C) regroupant les boutons pour les différentes actions possibles.

La fenêtre de résutat est organisée en 3 parties (Figure 68) :

- Une partie pour afficher des profils de données le long de la séquence appât. On peut ainsi visualiser les Tm (calculés selon plusieurs méthodes), les %GC, le pourcentage d'identité et le nombre de séquences détectées par BLASTN.
- Une partie pour visualiser les sondes par rapport à la séquence appât. Les sondes sont coloriées selon un code couleur décrit dans la Figure 69. Les positions en nombre de bases par rapport à la séquence appât sont visualisables ainsi q'un résumé des informations pour un oligonucléotide.
- Une partie regroupant les boutons pour toutes les actions disponibles comme par exemple, zoomer, afficher la sonde choisie, afficher le résultat du BLASTN…

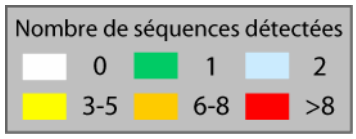


Figure 69 **Schéma regroupant les colorations des sondes.**
La coloration est faite en fonction du nombre de séquences retenues dans le BLASTN.

CADO4MI permet également de visualiser les résulats de plusieurs designs effectués en parallèle pour une même séquence appât. Cette stratégie permet de réaliser un design de sondes plus précis en tenant compte de plusieurs banques de séquences aux propriétés différentes. La Figure 70 montre un exemple de visualisation de 2 designs pour une même séquence (l'un avec la banque RefSeq et l'autre avec la banque Unigene) et illustre l'intérêt

de ce type de stratégie. En effet, l'utilisation de la banque RefSeq seule aurait conduit au choix de loligonucléotide en position 1331 (par rapport à la séquence appât) alors que l'utilisation combinée des 2 banques a permis de choisir l'oligonucléotide en position 71. La Figure 70 permet également de visualiser un certain nombre d'autres fonctions disponibles au travers de l'interface.

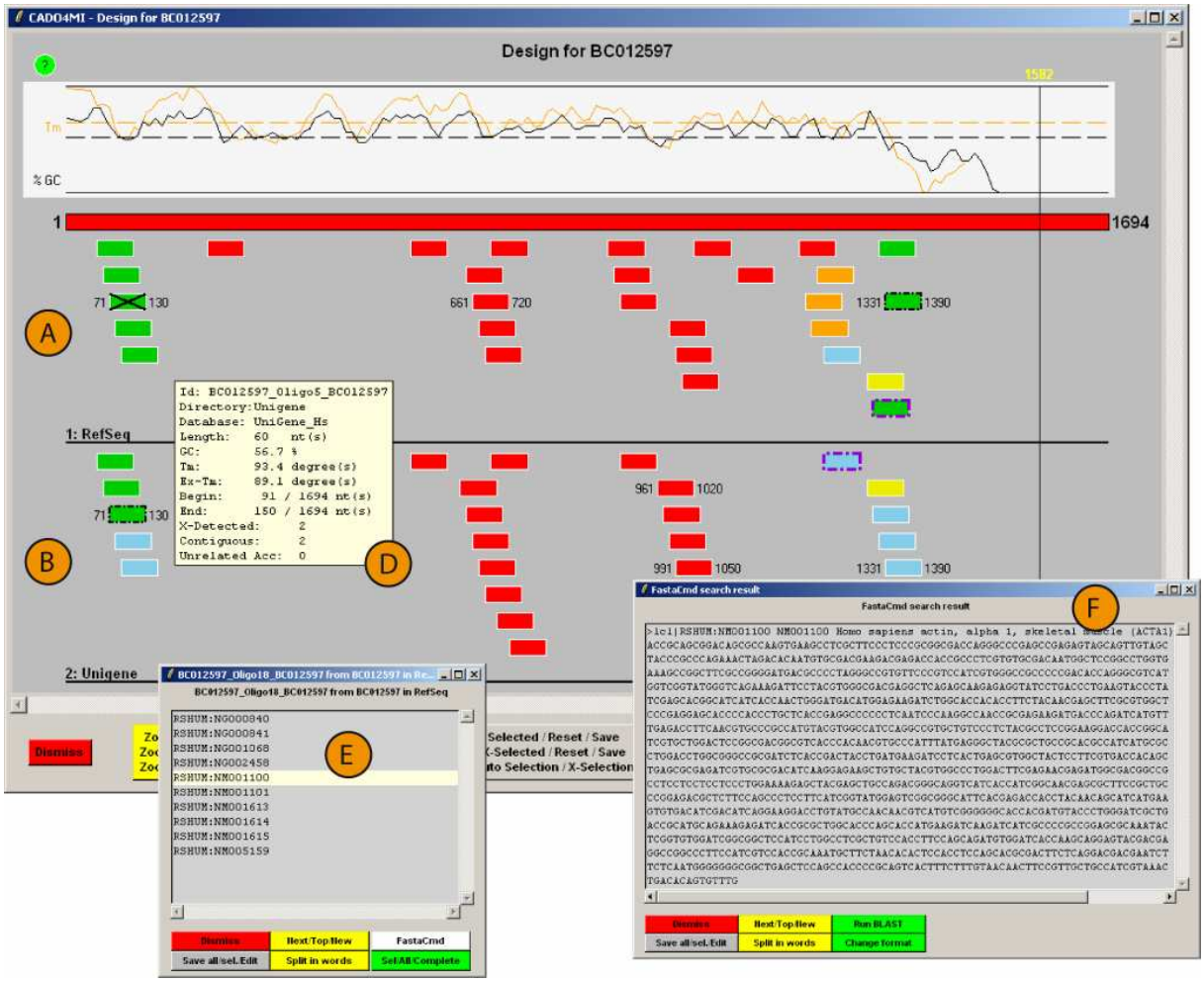


Figure 70 Exemple de visualisation d'un design pour une même séquence dans 2 banques de séquence différentes.

(A) Résultat pour la banque RefSeq, (B) résultat pour la banque UniGene. Le meilleur oligonucléotide pour chaque banque est indiqué avec une bordure en pointillés noir alors que le meilleur oligonucléotide des 2 banques combinées (« Selection croisée ») est indiqué par une croix noire. A partir d'un oligonuléotide, on peut afficher la liste de ses propriétés (D) comme par exemple le Tm ou son %GC, obtenir la liste des séquences détectées (E) et une séquence en particulier (F) pour effectuer d'autres actions (BLAST par exemple). Une fonction de recherche textuelle permet également de rechercher une sonde par son identifiant (oligonucléotides avec une bordure pointillée mauve).

9.8 La ligne de commande

Afin de rendre CADO4MI encore plus flexible à l'utilisation et favoriser son incorporation dans des scripts d'automatisation, j'ai implémenté un interpréteur de commandes permettant d'utiliser des arguments pour piloter et paramétrer le programme.

Argument	Action ou paramètre
-q (Query)	Séquence appât au format fasta [Texte]
-l (Length)	Longueur de l'oligonucléotide
-de (Distance)	Distance de la fin 3' (zone de design restreint)
-ar (Area)	Liste de couples de positions (zone de design restreint) ex: 1 300 500 800 will allow design only within the two ranges
-nb (nboligo)	Nombre d'oligonucléotide à conserver en fin de design (défaut 100)
-t (Tm)	Seuil pour la valeur de Tm [Flottant]
-r (TmRange)	Ecart de Tm (ex: 5) [Flottant]
-m (TmMethod)	Méthode de calcul du Tm [Texte] NearestNeighbor (défaut), Wallace, DNADNA, DNARNA ou RNARNA
-gl (GCLower)	Le % minimum de GC accepté (ex: 35.0) [Flottant]
-gu (GCUpper)	Le % maximum de GC accepté (ex: 70.0) [Flottant]
-p (Prohib)	Séquence prohibée (ex: AAAA TTTTT)
-b (Database)	Banque BLAST (ex: .nsq) [Texte]
-w (WorkingDir)	Répertoire de sauvegarde des résultats [Texte]
-qc (QueryCheck)	Activation du module de validation (1 or 0) [Entier]
-st (Step)	Valeur de déplacement de la fenêtre d'analyse (défaut 10)
-sp (Specificity)	Spécificité 1, pourcentage d'identité (défaut 70.0) [Flottant]
-nc (Contiguous)	Spécificité 2, nombre de bases successives identiques (défaut 15) [Entier]
-ts (Salt)	Concentration en sel (défaut 1.0) [Flottant]
-tf (Formamide)	Formamide (défaut 12.3 %) [Flottant]
-tc (DNA)	Concentration en acides nucléiques (défaut 0.000001) [Flottant]
-ev (BlastExpect)	Seuil de valeur d'Expect du BLAST (défaut is 100) [Flottant]
-bw (BlastWord)	Taille de mot pour BLAST (défaut 15) [Entier]
-ba (BlastNbAligned)	Nombre de séquences à aligner dans BLAST (défaut 2000) [Entier]
-bs (BlastStrand)	Blast option forward (1, défaut), reverse (2) or both (3) strands
-bf (BlastFilter)	Filtre de BLAST, T pour activer (défaut), F pour désactiver [Texte]

Tableau 10 Liste des arguments disponibles pour paramétrer CADO4MI en ligne de commande.

9.9 Autres modules de CADO4MI

9.9.1 Estimation de paramètres (Tm et GC)

Le design de sondes utilise certains critères de stabilité comme le Tm ou le %GC qui peuvent être très différents d'une séquence à l'autre. Ces 2 critères sont très importants pour garantir une hybridation homogène de la cible sur sa sonde. A l'échelle d'une puce entière, la valeur optimale du Tm d'une séquence peut être différente de celle de l'ensemble du set auquel elle appartient. Ainsi, afin d'obtenir une hybridation homogène pour l'ensemble des sondes présentes sur la puce, il est primordial d'estimer leur valeurs de Tm et de %GC optimaux. Ceci est possible en déterminant l'intervalle de valeurs permettant de choisir le maximum de sondes.

Nous avons ainsi implémenté une fonction permettant le calcul et la comptabilisation de l'ensemble des valeurs de Tm ou de %GC pour une liste de séquences soumises à CADO4MI. La visualisation et l'interprétation de ses valeurs permet ensuite de choisir l'intervalle optimal (Figure 71). L'utilisateur peut également paramétrer la taille de la fenêtre de calcul pour correspondre avec la taille de ses sondes.

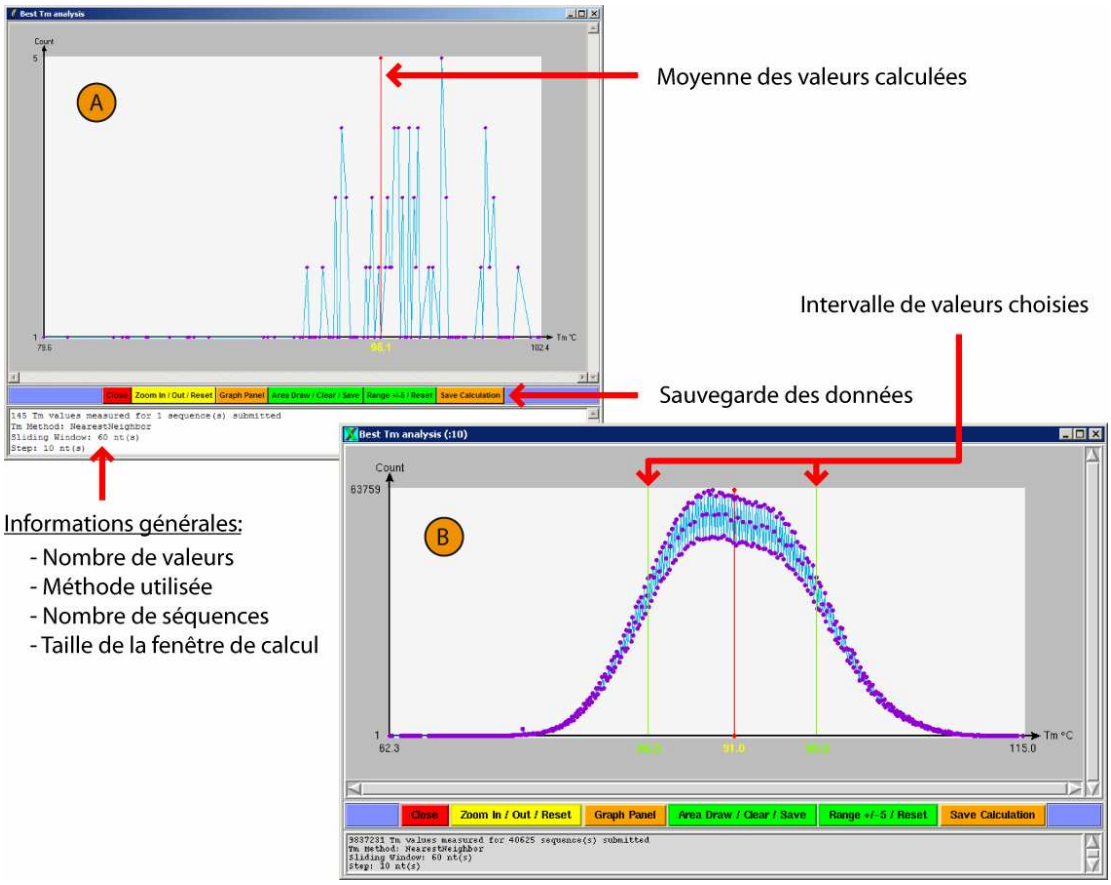


Figure 71 **Exemples de résultats de l'estimation de paramètres.**
L'interface permet de visualiser un graphique indiquant le nombre de valeurs en fonction de la valeur calculée. (A) Estimation des valeurs de Tm avec la méthode du Nearest Neighbor pour une fenêtre de taille 60 déplacée toutes les 10 bases pour une seule séquence (NM_001100). (B) Estimation dans les mêmes conditions que pour (A) en utilisant la totalité des séquences de la banque RefSeq (release 16) (>40000 séquences).

Le graphe visualise les valeurs (axe des abscisses) et leur décompte dans la population testée (axe des ordonnées). La moyenne est indiquée et permet, par exemple, de choisir un intervalle de plus ou moins 5°C dans le cas du calcul du Tm. Le graphe étant interactif l'utilisateur peut définir des intervalles différents et les sauver. Les résultats de ces calculs peuvent également être sauvés dans un fichier TSV (*Tab Separated Value* ou fichier délimité par des tabulations) afin de comparer différents paramètres au moyen d'un tableur.

9.9.2 Validation de la séquence appât

Dans le cas idéal, un gène correspond à un transcrit dans les banques de séquences et à une sonde choisie. Cependant dans la réalité, nous devons faire face à l'hétérogénité des gènes auxquels peuvent correspondre plusieurs transcrits et à l'hétérogénité des banques de séquences qui peuvent contenir des versions différentes des mêmes transcrits. Ce type de problème a été rencontré lors du design de sondes avec CADO4MI. Cette différence se traduit par plusieurs cas distincts révélés lors de l'étape de recherche de similarité, comme

l'absence totale ou partielle de la séquence appât dans la banque testée (plusieurs cas possibles voir Figure 72). Ces situations sont responsables d'un design erroné par un mauvais positionnement des sondes dont les conséquences sont des sondes trop éloignées de l'extrémité 3' (cas « D » ou « E » Figure 72) ou des sondes limitées à des zones de spécificité délicates (plusieurs séquences détectées).

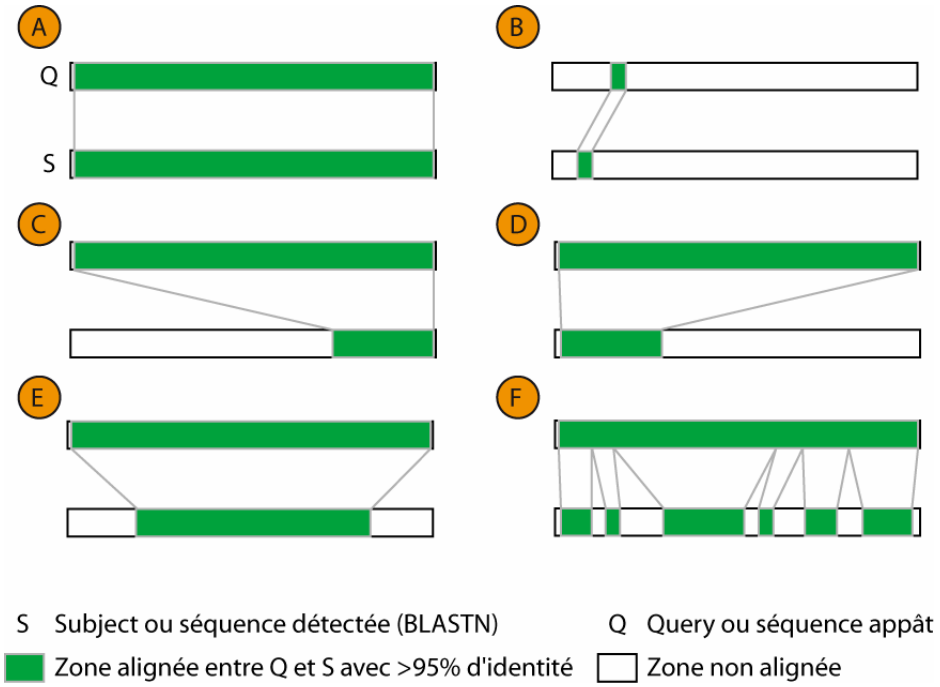


Figure 72 Schématisation des différents cas de figure rencontrés en comparant une séquence appât à une banque de séquences.
Le résultat de l'alignement, entre la séquence appât ou « Q » et la meilleure séquence ou « S » dans la banque de données, est représenté. Les zones vertes correspondent aux zones réellement alignées et les zones blanches correspondent aux zones non alignées. (A) représente une séquence détectée par BLAST parfaitement équivalente à la séquence appât. (B) représente le cas opposé de (A) avec une séquence absente de la banque (seule une petite portion résiduelle est alignée). (C) et (D) montrent l'existence dans la banque d'une séquence plus longue respectivement en 5' et en 3'. (E) illustre le fait que Q peut être incluse dans une séquence de la banque testée. (F) représente le cas particulier de Q segmentée dans le S.

Pour minimiser les conséquences de ces différences, nous avons développé un module permettant de vérifier si la séquence appât est présente dans la banque de séquence utilisée pour le design des sondes. Ce module est basé sur des critères de séquence, comme le pourcentage d'identité et le pourcentage de recouvrement (5.4.2.2 BlastPanel). Le seuil pour le pourcentage d'identité global entre les séquences détectées et la séquence appât est positionné à 95% et celui pour le pourcentage de recouvrement de la query par rapport aux séquences detectées est positionné à 70%. Ces informations sont également disponibles dans le fichier de résultat global généré par l'utilisateur.

Chapitre 10 - Autres outils développés

Dans le contexte du développement de la plateforme de transcriptomique du Luxembourg et pour la mise en place d'Actichip, divers outils et protocoles bioinformatiques ont été développés pour remplir des taches dédiées. Ces outils permettent par exemple de comparer les séquences de différentes puces entre elles, de reformater un fichier FASTA en modifiant son entête ou encore de gérer les fichiers issus de programmes comme GenePix.

10.1 DbFastER

DbFastER « DataBase Fasta HeadER modification » est un outil de modification des entêtes de fichiers FASTA fournis par le NCBI. Il permet ainsi de simplifier les entêtes et la lecture des résultats d'une recherche par le programme BLAST. Cet outil est développé en Tcl/Tk et comporte une interface minimaliste (Figure 73).

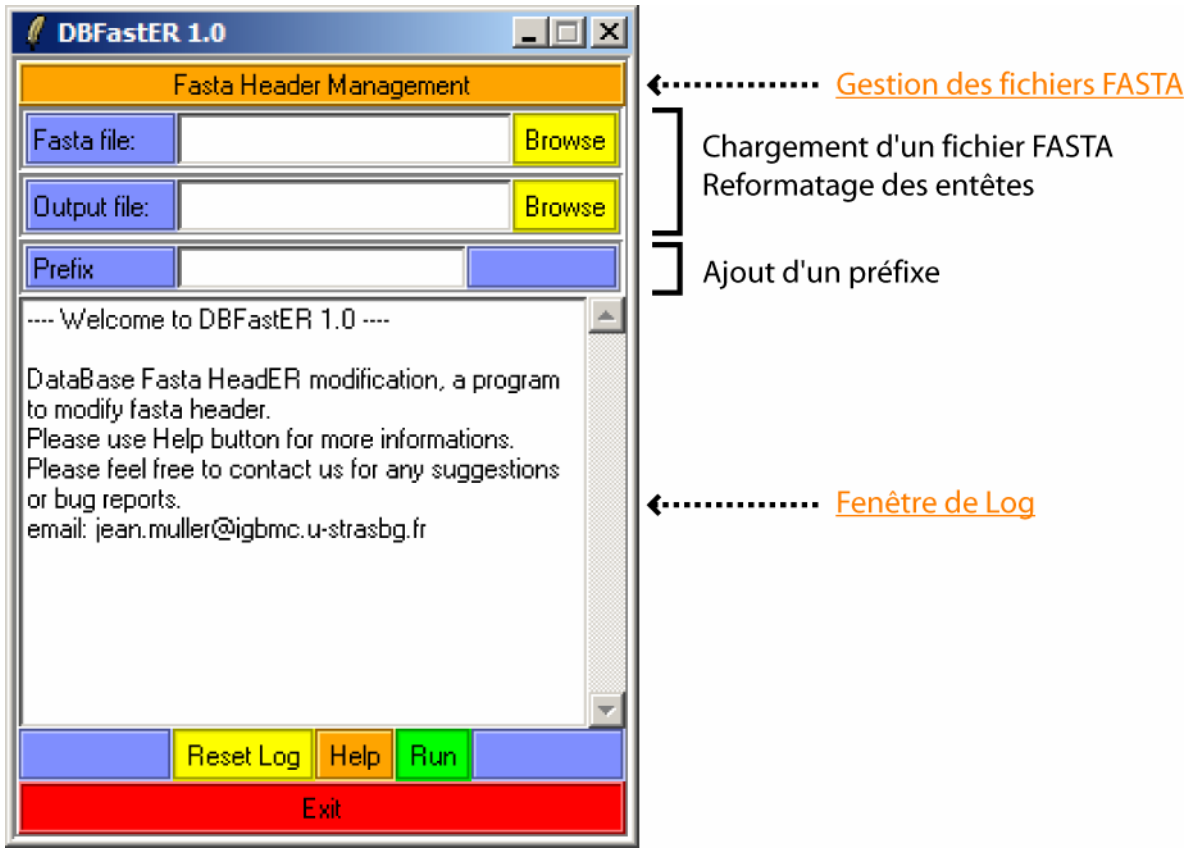


Figure 73 **Interface du programme DbFastER.**

L'utilisation de BLAST pour la recherche de similarité nécessite l'utilisation de banques de de séquences au format adéquat, c'est-à-dire au format BLAST. La création de banques de données au format BLAST est basée sur la conversion d'un fichier de séquences au format FASTA au moyen du programme formatdb également distribué par le NCBI. Au cours de ce processus, l'entête de chaque séquence FASTA est utilisée pour décrire cette séquence dans un résultat d'une recherche de similarité. Pour une utilisation simplifiée et une analyse plus rapide des résultats de BLAST, les entêtes doivent être simples et dans l'idéal contenir uniquement l'identifiant de la séquence et sa définition, et éventuellement un préfixe pour préciser la banque de séquences utilisée (par exemple RS pour RefSeq ou UG pour UniGene…). Cependant, les fichiers de séquences des banques de séquences comme RefSeq (ftp://ftp.ncbi.nih.gov/refseq/H_sapiens/mRNA_Prot) ou UniGene (ftp://ftp.ncbi.nih.gov/repository/UniGene/Homo_sapiens) proposés par le NCBI possèdent des entêtes particulières, qui ne sont pas directement fonctionnelles (Figure 74). J'ai développé un petit outil qui permet de convertir de façon rapide et efficace ces fichiers et éventuellement, d'ajouter un préfixe. Un exemple de ce type de conversion est illustré par la Figure 74.

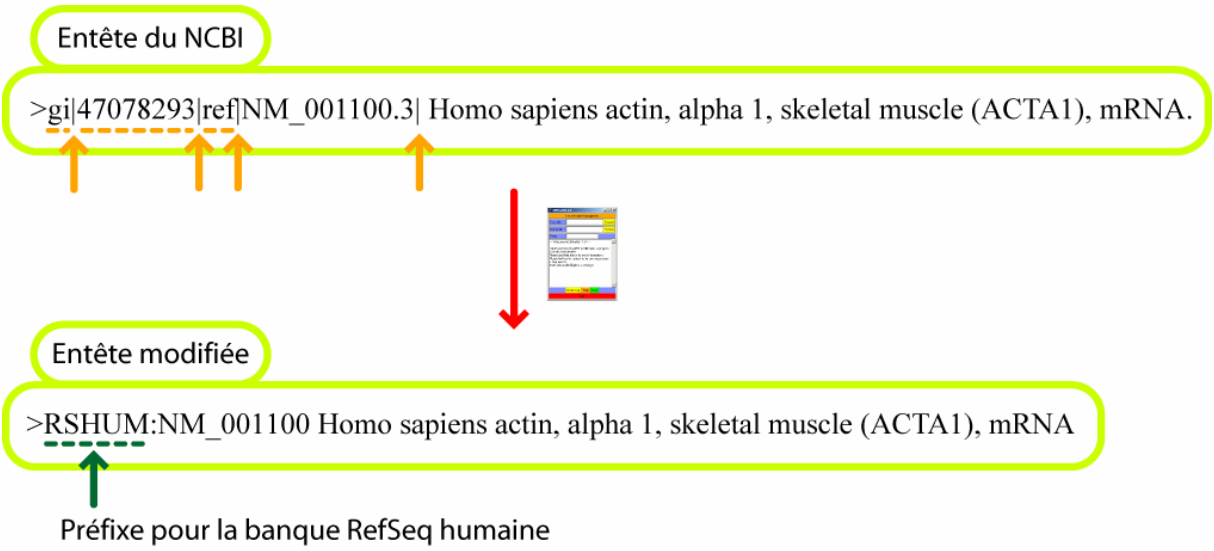


Figure 74 **Exemple de conversion de l'entête donnée par le NCBI par DbFastER.**
Les éléments indiqués en orange sont identifiés et éliminés.

10.2 GalActicA

Le programme GalActicA pour « Galfile Automatic Annotation » a été developpé dans le cadre des outils mis en place pour la plateforme de puce Actichip. Cet outil permet la gestion des fichiers GAL (*GenePix Array List*) et la gestion des fichiers TSV (*Tab Separated Value* ou fichier délimité par des tabulations). La plateforme de puce à ADN utilise les logiciels d'Axon (http://www.moleculardevices.com) pour créer les plaques, scanner les puces et analyser les images brutes (logiciel GenePix). Axon utilise et génère un format de fichier appelé ATF (*Axon Text Format*) décliné en 2 types de fichiers ; le fichier GAL (*GenePix Array List*) et le fichier GPR (GenePix Result). Ce format est caractérisé par une entête décrivant la structure des données qui vont suivre et par la liste des données organisées en colonnes séparées par des tabulations. Ces données peuvent être des chiffres ou du texte, mais dans ce cas, elles ne doivent pas contenir de caractères « spéciaux » (« , », « ; »…). Cette limitation peut être levée par l'utilisation de « guillemets » pour « protéger » le texte.

Le fichier GAL est un fichier texte contenant des informations spécifiques sur la localisation, la taille et le nom de chaque spot d'une puce. Il contient l'ensemble des informations utilisées pour la création de la puce. Typiquement les données d'un fichier GAL, en plus des coordonnées sur la puce (numéro de bloc, de ligne et de colonne), sont le nom (nom de gène par exemple) et une définition pour le spot. Le fichier GPR est le fichier de résultats d'analyse de GenePix. Ces 2 types de fichiers peuvent être édités par des tableurs comme, par exemple, Excel, mais sont souvent sources d'erreurs et de corruption du format (par exemple la perte des guillemets).

Le fichier TSV est un type de fichier utilisé couramment par les plateformes de bioinformatique ou de transcriptomique car elles permettent de stocker plusieurs informations ordonnées pour un élément (par un exemple un gène) par ligne. Ainsi, les fichiers d'annotation des éléments des puces à ADN sont souvent distribués sous cette forme.

La gestion informatique, sur la plateforme de puce à ADN, de ces types de fichiers par un programme dédié est essentielle pour en garantir la bonne utilisation. GalActicA permet ainsi d'ajouter et de supprimer les guillemets à une colonne de données, d'ajouter ou de remplacer une colonne par une autre. Le remplacement d'une colonne permet notammement de mettre à jour une colonne par de nouvelles données (par exemple une mise à jour des définitions des gènes). La mise à jour ou l'ajout d'une colonne combine les fichiers GAL (ou GPR) avec un fichier TSV. Outre cette possibilité d'enrichir les informations d'un fichier

GAL, GalActicA permet également de fusionner 2 fichiers TSV ayant des références communes.

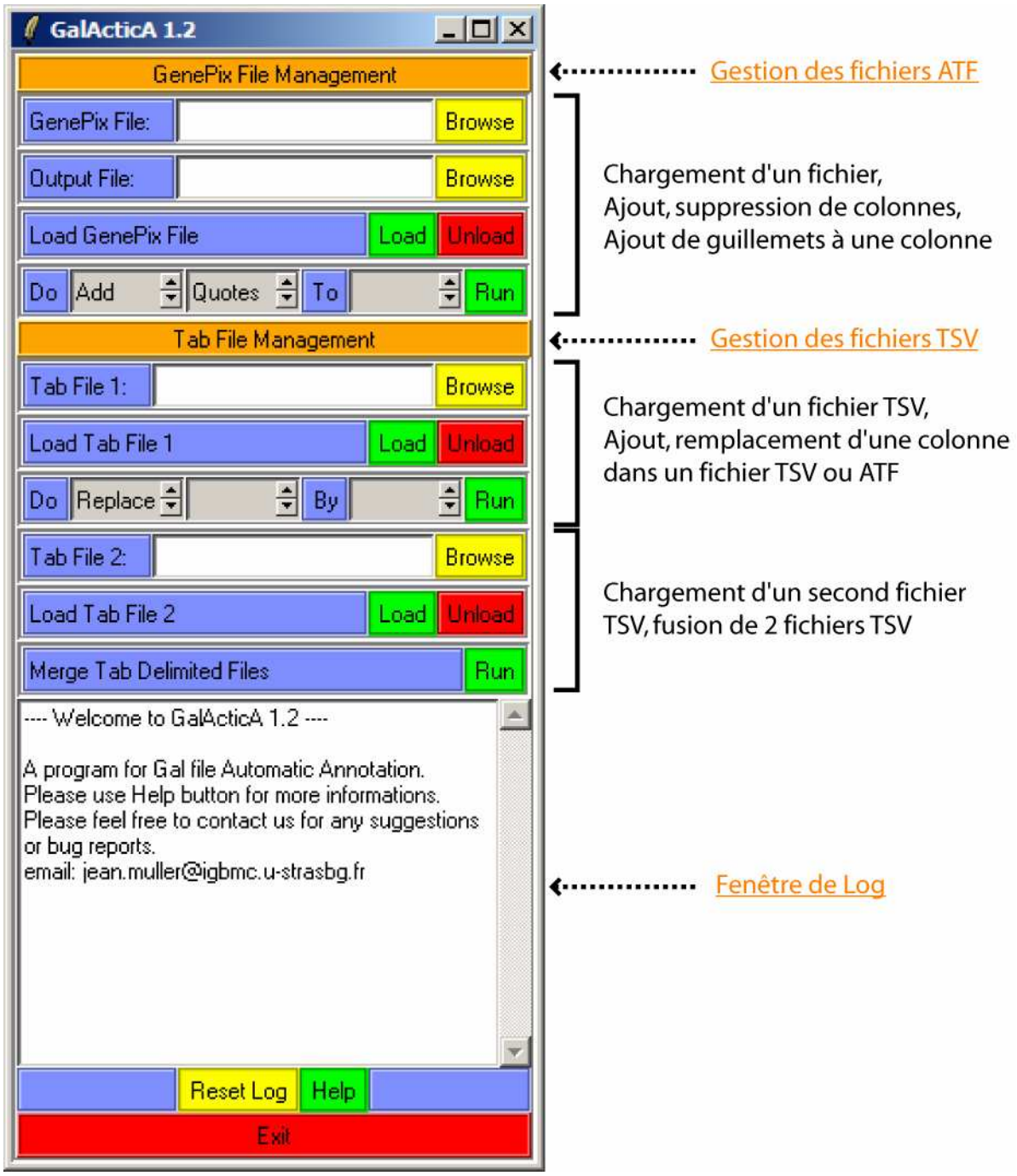


Figure 75 **Interface du programme GalActicA.**

D'un point de vue informatique, le chargement en mémoire de n'importe quel type de fichier supporté par le programme met à jour automatiquement son interface pour proposer la liste des colonnes disponibles. Ceci est rendu possible par l'implémentation d'un analyseur syntaxique indépendant des noms de colonnes et des données, et l'utilisation d'un seul point de référence géré par l'utilisateur via une boite de dialogue.

10.3 Comparaison de puces

Dans le cadre du développement de notre puce dédiée au cytosquelette, Actichip, nous avons comparé les résultats d'Actichip avec les résultats d'autres plateformes de puces à ADN. Pour pouvoir réaliser cette comparaison, il est nécessaire de connaître la liste des gènes partagés par les plateformes. En conséquence, nous avons développé un certain nombre de procédures informatiques dont le but est de répertorier les gènes en commun.

L'approche choisie combine à la fois une recherche textuelle en comparant des listes d'identifiants de séquences et une recherche de similarité en comparant les séquences disponibles pour les différentes plateformes (Figure 76).

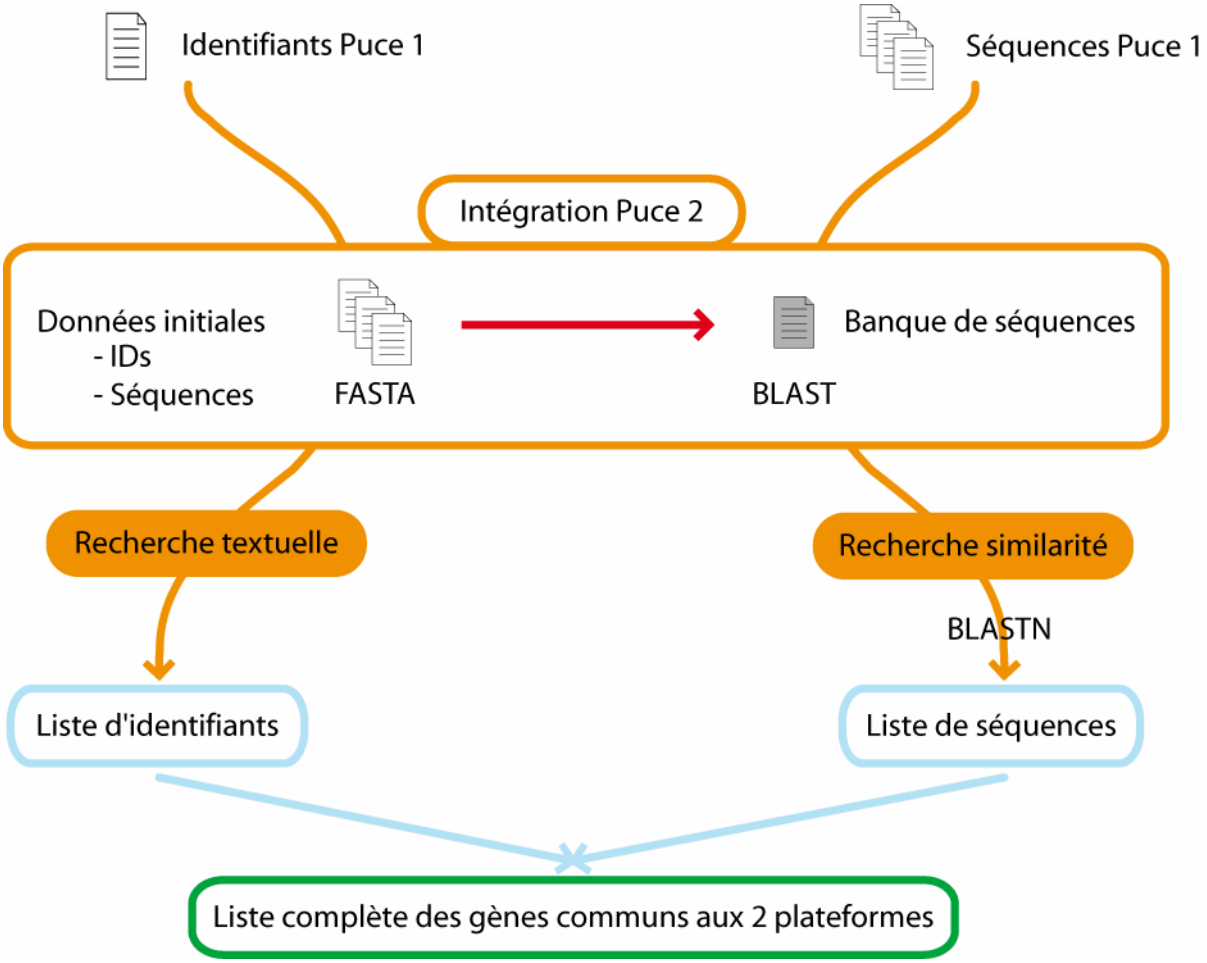


Figure 76 **Processus d'intégration et de comparaison des données d'une puce.**
Le cadre Orange délimite la zone d'intégration qui comprend un analyseur syntaxique qui extrait les identifiants et les séquences associées ainsi qu'une banque BLAST associée aux séquences extraites. La comparaison se fait ensuite à 2 niveaux, d'une part, les listes des identifiants des 2 puces sont comparées et d'autre part, les séquences des 2 puces sont alignées et analysées. Au final, on obtient une liste de gènes communs aux 2 puces.

Les différentes plateformes de puces commerciales fournissent rarement la liste des séquences sondes utilisées, mais plutôt une liste d'identifiants dans les banques nucléiques (GenBank ou RefSeq). A partir de ces identifiants, nous recherchons les séquences associées par des requêtes SRS.

Les programmes développés permettent d'une part, l'extraction des données importantes (identifiant de séquence et/ou séquences) par la création de parsers spécifiques à chaque plateforme, et d'autre part, de comparer leurs identifiants et/ou leurs séquences. La comparaison de séquences utilise le programme BLAST ainsi que les différents paramètres décrits plus haut comme le pourcentage d'identité et les pourcentages de recouvrement. Selon le type de plateforme et les types de séquences comparées, les limites seront ajustées.

Résultats

La génomique comparative

Chapitre 11 - Une approche globale : Actinome

« On ne doit pas escamoter l'incompréhensible,
mais non plus s'en servir comme d'une explication. »
Jean Rostand

Le cytosquelette est un système complexe composé de multiples familles de protéines dont certaines interactions avec d'autres membres du cytosquelette sont connues. Cependant, les implications de ces familles de protéines dans l'établissement du cytosquelette demeurent souvent inconnues. La génomique comparative cherche à mieux comprendre l'évolution des gènes et de leurs fonctions. En particulier, les profils phylogénétiques permettent, à partir d'un bilan de présence/absence dans plusieurs organismes, d'établir des liens entre les protéines qui ne partagent pas nécessairement de similarité de séquences.

Une étude préliminaire des profils phylogénétiques de l'ensemble des gènes du cytosquelette participe de la réflexion autour de la génomique comparative au sein du laboratoire. Elle permettra de développer les outils automatiques et les protocoles standards nécessaires à sa mise en place.

Elle s'articule autour de la banque Actinome qui permettra dans le cadre de cette approche globale et automatique de fournir l'ensemble des séquences protéiques attachées aux cytosquelettes. Dans une seconde approche, centrée sur les ARPs (Chapitre 12 - Les Actin-Related Proteins), elle fournira un sous ensemble de données dédiées.

A terme, cette étude vise à identifier les gènes susceptibles de participer, au sein du cytosquelette, aux mêmes complexes protéiques ou à des fonctions proches. Elle devrait également permettre d'identifier les groupes de gènes fonctionnels apparus au cours de l'évolution représentée par les 41 génomes eucaryotes utilisés (de l'algue à l'Homme).

Pour construire les profils phylogénétiques de l'ensemble des gènes du cytosquelette nous avons ainsi utilisé le programme ComIcs développé pour l'occasion. Ce programme décrit dans la partie Matériel et Méthodes (Chapitre 7 - ComIcs) permet de construire des profils phylogénétiques à partir de la recherche de similarité dans les banques de séquences protéiques.

11.1 Stratégie retenue

11.1.1 Séquences appâts et ComIcs

Les séquences appâts considérées dans cette étude correspondent à l'ensemble des 1742 séquences protéiques contenues dans la version 3.0 d'Actinome (6.1.1 Historique).

ComIcs a été paramétré pour rechercher un maximum de séquences dans les banques de protéines. Ainsi, BLASTP a été utilisée avec une limite d'expect à 10 et un maximum de séquences alignées et retenues à 5000. La présence d'une séquence dans un organisme est uniquement décrite si elle vérifie une valeur d'expect inférieure à 10^{-3}.

11.1.2 Banques protéiques utilisées

Nous avons utilisé 2 banques de protéines : UniProt qui contient des séquences annotées de haute qualité (pour la partie correspondant à Swiss-Prot) et RefSeq qui contient des séquences validées. L'utilisation de 2 banques permet de compenser d'éventuelles lacunes dans les couvertures de chacune de ces banques pour les différents organismes.

11.1.3 Clustering

Le regroupement ou « clustering » des profils les plus proches a été réalisé au moyen de 2 méthodes :

- Par une Analyse en Composantes Principale. Cette analyse a été développée par Nicolas Wicker au sein du laboratoire. L'analyse en composantes principales, communément appelée ACP, est une méthode statistique multidimensionnelle qui permet de synthétiser un ensemble de données en identifiant la redondance dans celles-ci. Dans notre étude, les données correspondent à un tableau de 41 lignes (organismes) x 1742 colonnes (séquences).
- Par le logiciel DPC (Wicker *et al.* 2002) pour *Density of Point Clustering.* Il est notamment utilisé au laboratoire pour différentes taches comme le clustering d'alignement multiples ou de données de puces à ADN.

11.2 Résultats et discussion

11.2.1 ComIcs

Le premier résultat de cette analyse est la création et la validation du programme ComIcs qui permet la définition de manière automatique de nombreux profils phylogénétiques. L'application de ces stratégies au travers de son utilisation a permis d'obtenir les 1742 profils phylogénétiques.

Cette analyse automatique nécessite pour chaque mise à jour 8710 fichiers répartis pour chaque banque (UniProt et RefSeq) en 1742 fichiers BLAST et 1742 fichiers « .report ». A cela s'ajoute 1742 fichiers « .report » pour la fusion des résultats des 2 banques. Ces fichiers représentent environ 2 Go de données stockées sur les serveurs. Au niveau du temps de calcul, l'ensemble des BLASTP peut être réalisé en 48H et l'analyse d'un seul de ces BLASTP peut prendre entre 10 secondes et 10 minutes (pour les BLASTPs les plus importants) sur les machines les plus puissantes disponibles (notamment *star*, voir le chapitre 5.1.1 IGBMC). Les résultats présentés ici concernent la dernière mise à jour effectuée en juin dernier.

L'interface graphique est capable de supporter les 1742 profils simultanément et permet la réalisation des différentes tâches (tris, mises à jours) dans des temps rapides. Ceci valide les différents choix de programmation décrits dans la partie Matériel et Méthodes (Chapitre 7 - ComIcs).

11.2.2 Bilan général de cette analyse

Dans un premier temps, dans le but de mieux comprendre les objets que nous manipulons, nous avons analysé les résultats de manière globale.

En lien avec la stratégie établie par défaut dans ComIcs, nous avons délibérément paramétré le programme BLASTP dans le but de détecter un maximum de séquences dans les banques de protéines. Les valeurs limites d'expect (option –e du BLAST) et du nombre de séquences alignées (options –v et –b du BLAST) ont été définies de manière à ce qu'elles soient le moins restrictives possible. Les paramètres sont donc pour l'expect et le nombre de séquences retenues, de 10.0 et 5000 respectivement. La présence d'une protéine dans un organisme est définie dans une première approche par l'utilisation de la valeur d'expect. Ce seuil a été placé à 10^{-3} par défaut. L'utilisation de différents seuils pour les valeurs d'expect nous permet d'un côté d'être plus spécifique (E= 10^{-3}) et de l'autre de conserver les résultats d'une

recherche exhaustive par le programme BLASTP (ces alignements de séquences pourront être questionnés par la suite).

L'utilisation de 2 banques de séquences protéiques (UniProt et RefSeq) est issue des observations effectuées lors d'analyses préliminaires et qui montraient l'absence de certaines protéines dans l'une ou l'autre banque. Ainsi, leur association nous permet de couvrir un plus grand nombre de protéines des organismes à génome complet retenus pour cette étude. La comparaison du nombre total de séquences détectées par les BLAST effectués sur l'ensemble des 1742 protéines en fonction des organismes et dans chacune des 2 banques, nous permet de mieux comprendre les différences de résultats observées (Figure 77). Il faut d'abord noter que ces chiffres sont à prendre avec beaucoup de précaution puisqu'ils représentent les résultats de la recherche d'un groupe de séquences très particulier, les protéines du cytosquelette, qui recouvrent des fonctions à la fois très anciennes et très récentes et dont les profils de présence/absence peuvent être très variés. Ainsi, une détection de peu de séquences dans un organisme peut reflèter la faible représentativité de cet organisme dans la banque ou le fait qu'il existe peu d'homologues des protéines du cytosquelette dans cet organisme. Néanmoins, la comparaison d'une banque par rapport à l'autre nous permet de lever en partie ce type d'ambiguïté. Par exemple, la distribution générale du nombre de séquences semble être comparable, puisque les organismes les mieux décrits dans la littérature sont les mieux représentés dans les 2 banques (l'homme, la souris, la mouche, le poisson zèbre, arabidospis et le riz). La différence est focalisée sur certains organismes en particulier ; l'homme est mieux représenté dans UniProt alors que la souris l'est beaucoup plus dans RefSeq. Ceci est du au nombre plus important de séquences prédites de la souris dans RefSeq.

Certains organismes ne sont pas représentés dans RefSeq ; c'est le cas pour *Tetraodon nigroviridis, Caenorhabditis briggsae, Debaryomyces hansenii, Cryptococcus neoformans, Thalassiosira pseudonana, Tetrahymena thermophila* et *Cyanidioschyzon merolae*. Cette observation est compréhensible, dans certains cas, comme celui de *Tetrahymena thermophila,* dont le génome est moins bien annoté, mais pose un problème pour les autres. A l'inverse, RefSeq semble plus complet sur certains organismes comme l'abeille (*Apis mellifera*), le rat, le poulet et l'oursin (*Strongylocentrotus purpuratus*).

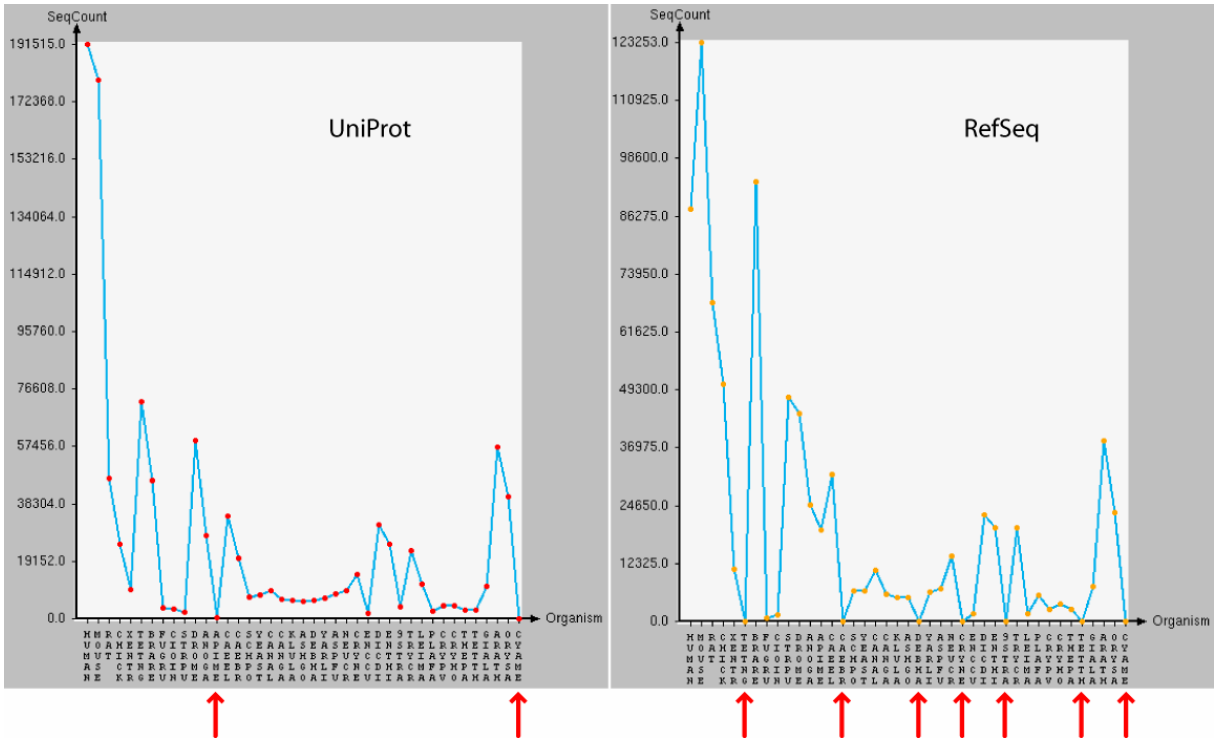


Figure 77 Nombre de séquences détectées au total pour chaque organisme.
Nombre de séquences détectées au total dans les 1742 BLASTPs par organisme en considérant une valeur d'expect de 10^{-3} pour chacune des 2 banques utilisées. Les flèches rouges indiquent pour la banque UniProt les organismes les moins représentés et pour RefSeq les organismes absents. Liste des organismes et des abrégés correspondants présentés de gauche à droite sur l'axe des organisme : ***Homo sapiens*** (HUMAN)***, Mus musculus*** (MOUSE)***, Rattus norvegicus*** (RAT)***, Gallus gallus*** (CHICK)***, Xenopus tropicalis*** (XENTR)***, Tetraodon nigroviridis*** (TETNG)***, Brachydanio rerio*** (BRARE)***, Takifugu rubripes*** (FUGRU)***, Ciona intestinalis*** (CIOIN)***, Strongylocentrotus purpuratus*** (STRPU)***, Drosophila melanogaster*** (DROME)***, Anopheles gambiae*** (ANOGA)***, Apis mellifera*** (APIME)***, Caenorhabditis elegans*** (CAEEL)***, Caenorhabditis briggsae*** (CAEBR)***, Schizosaccharomyces pombe*** (SCHPO)***, Saccharomyces cerevisiae*** (YEAST)***, Candida albicans*** (CANAL)***, Candida glabrata*** (CANGA)***, Kluyveromyces lactis*** (KLULA)***, Ashbya gossypii*** (ASHBA)***, Debaryomyces hansenii*** (DEBHA)***, Yarrowia lipolytica*** (YARLI)***, Aspergillus fumigatus*** (ASPFU)***, Neurospora crassa*** (NEUCR)***, Cryptococcus neoformans*** (CRYNE)***, Encephalitozoon cuniculi*** (ENCCU)***, Dictyostelium discoideum*** (DICDI)***, Entamoeba histolytica*** (ENTHI)***, Thalassiosira pseudonana*** (9STRA)***, Trypanosoma cruzi*** (TRYCR)***, Leishmania major*** (LEIMA)***, Plasmodium falciparum*** (PLAFA)***, Cryptosporidium parvum*** (CRYPV)***, Cryptosporidium hominis*** (CRYHO)***, Theileria parva*** (THEPA)***, Tetrahymena thermophila*** (TETTH)***, Giardia lamblia*** (GIALA)***, Arabidopsis thaliana*** (ARATH)***, Oryza sativa*** (ORYZA) ***et Cyanidioschyzon merolae*** (CYAME)***.***

Une première conclusion de cette analyse réside dans l'importance de l'utilisation des 2 banques de protéines disponibles. Leur complémentarité permet de couvrir au maximum le protéome de chaque organisme afin d'assurer la détection d'au moins un homologue (s'il existe) dans un organisme.

Une valeur d'expect égale à 10^{-3} est en général une limite acceptable pour assurer un lien entre 2 séquences alignées par le programme BLASTP. Cependant, l'organisation en domaines de certaines familles de protéines peut fausser ces résultats. Afin de mieux

observer l'effet de la valeur d'expect retenue comme significative sur la décision de la présence ou de l'absence d'une protéine dans un organisme, nous avons fait varier la limite de décision entre 10^{-3} et 10^{-100}. Les 1742 profils ont été affichés et classés en fonction de la distance euclidienne appliquée sur le profil de la séquence ACT0004 correspondant à une isoforme d'actine présente dans tous les organismes. Les profils sont ainsi classés du plus complet (présent chez tous les organismes) au moins complet (présent uniquement chez l'homme) (Figure 78).

Les zones les plus rouges correspondent aux parties des profils qui restent identiques quelque soit la valeur d'expect utilisée (Figure 78). Elles dénotent de la détection constante dans les résultats du BLASTP, d'homologues ayant une valeur d'expect très faible (10^{-100}). C'est le cas, par exemple, des zones 1 et 2 (voir Figure 78 B). Les profils de la zone 1 correspondent aux protéines les plus fortement conservées dans l'ensemble des organismes comme les actines, les tubulines ou certaines myosines. Les profils de la zone 2 sont ceux conservés chez les métazoaires et plus particulièrement les organismes vertébrés.

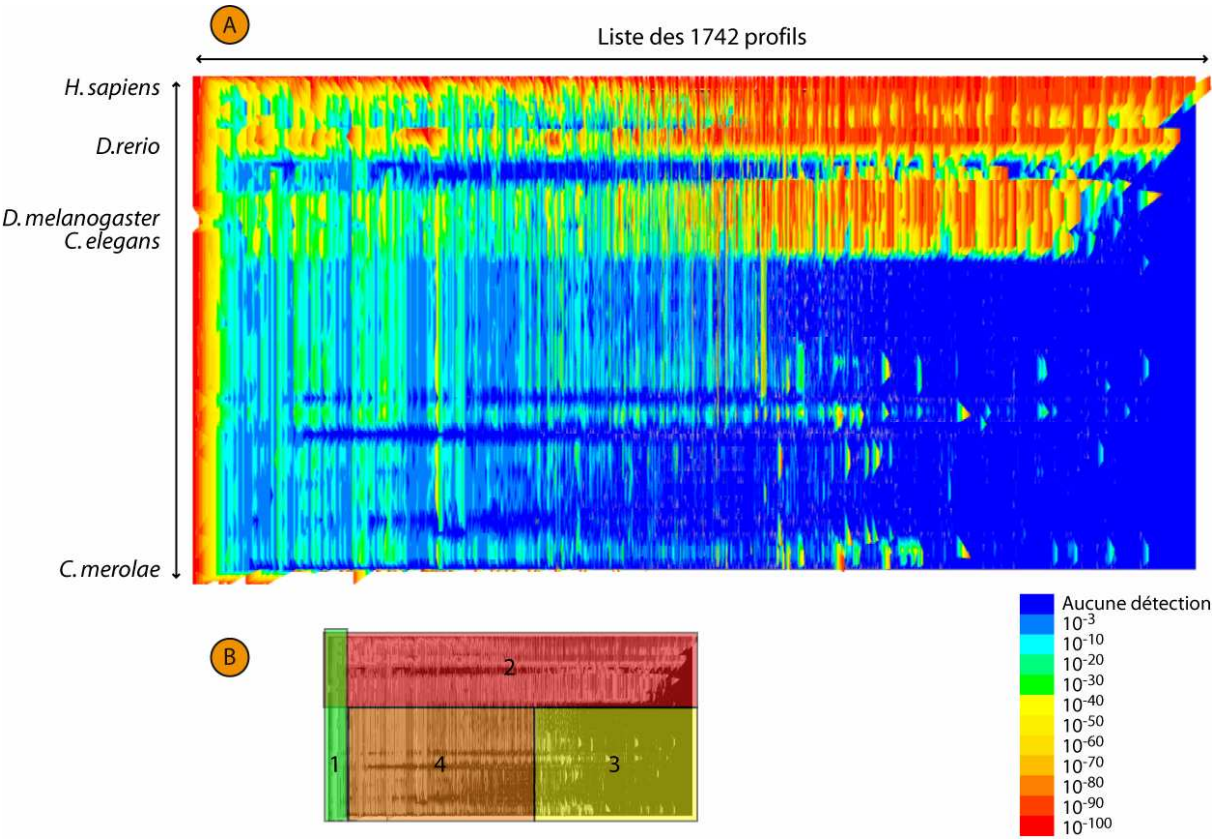


Figure 78 **Impact de la valeur d'expect du BLASTP dans la détection des séquences.**
(A) Schéma général représentant les valeurs d'expect à partir desquelles la détection de la présence d'un homologue dans un autre organisme est effective. La légende est indiquée à droite. (B) Vue réduite et annotée des différentes zones de la partie (A).

Les zones bleues foncées correspondent aux parties des profils qui ne détectent aucune séquence pour les organismes considérés. C'est le cas de la zone 3 de la Figure 78 B.

Enfin la zone 4 correspond à une zone d'incertitude ou l'affirmation de la présence d'un homologue de la protéine considérée dans un organisme n'est pas certaine.

On peut remarquer, sans que cela soit réellement une preuve, que l'utilisation de l'expect 10^{-3} est possible dans une première approche pour déterminer les profils phylogénétiques des organismes eucaryotes puisque la majorité des séquences ont des expects supérieurs à 10^{-10}. Cependant pour certaines zones ou familles de protéines ceci ne sera pas suffisant.

11.2.3 Analyse intégrée des clusters

Dans le cadre d'une analyse plus avancée des 1742 profils phylogénétiques, nous avons introduit une étape supplémentaire. Cette étape consiste à définir des groupes ou des clusters contenant les profils phylogénétiques les plus proches. L'utilisation de 2 méthodes de clustering développées par Nicolas Wicker au laboratoire, l'une basée sur une ACP et l'autre sur le programme DPC (*Density of Point Clustering*), a permis d'obtenir respectivement 13 et 64 clusters. La comparaison des 64 clusters obtenus par DPC montre une tendance à la surestimation du nombre de clusters sur ce jeu de données. En effet, beaucoup de clusters sont redondants et contiennent des profils très proches. Au contraire, l'analyse par ACP semble caractériser un nombre de clusters plus proche de la réalité. Seul un cluster semble abérrant en regroupant 2 profils très distincts et 2 clusters pourraient être fusionnés sur la vue de la proximité de leurs profils respectifs.

L'analyse du contenu des clusters et leur interprétation n'en est encore qu'à un stade préliminaire mais cependant, un certain nombre de points semblent se dégager. Le nombre de profils ou de protéines est relativement différent (Tableau 11).

Numéro de cluster	Nombre de protéines
Cluster 0	97
Cluster 1	107
Cluster 2	103
Cluster 3	385
Cluster 4	93
Cluster 5	265
Cluster 6	40
Cluster 7	119
Cluster 8	72
Cluster 9	168
Cluster 10	231
Cluster 11	60
Cluster 12	2

Tableau 11 Nombre de protéines par cluster.
Les 13 clusters obtenus par le clustering par ACP sont indiqués avec leur nombre de profils ou de protéines.

La composition des différents clusters est relativement homogène (Figure 79) du point de vue des différents profils regroupés et des protéines ou fonctions qu'ils représentent. Ainsi, certains types de profils sont clairement définis. Par exemple, le cluster 3 contient les protéines universelles, le cluster 5 est composé de profils restreints aux métazoaires. Le cluster 8 est clairement spécifique des organismes vertébrés. Le cluster 9 propose des profils dont les protéines sont absentes des parasites et des algues.

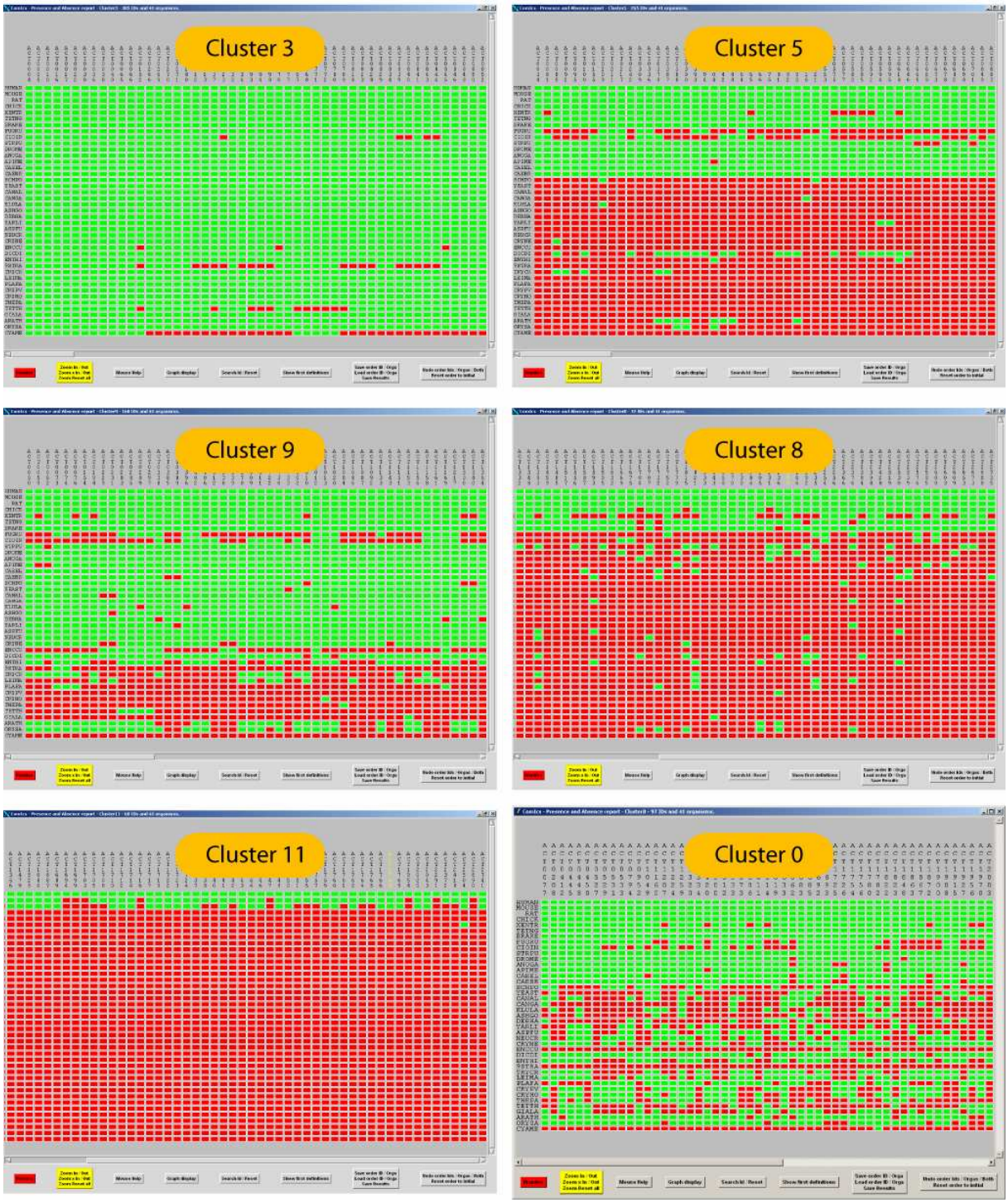


Figure 79 **Exemple de clusters de profils phylogénétiques.**
La composition des clusters est relativement homogène du point de vue des profils contenus (cluster 3, 5 11). La composition de certains clusters est néanmoins un peu plus délicate à analyser (Cluster 0).

Les clusters contiennent des familles de protéines complètes ou partielles comme, par exemple, les actines et les ARPs dans le cluster 3 (>10 protéines différentes), les protéines associées aux kératines dans le cluster 11 (>30 protéines différentes) ou encore dans le cluster 5 avec des protéines impliquées dans l'adhésion (les 3 desmocolines, les 3 desmogléines, la

plakophiline, les caténines alpha) et dans des fonctions liées au muscle (les troponines, les tropomyosines), des fonctions exclusivement observées dans les organismes multicellulaires.

11.3 Les limites d'une telle approche

Bien que ces analyses soient prometteuses, certains problèmes ou limitations sont néanmoins apparus. Ces problèmes doivent être résolus ou minimisés pour obtenir automatiquement des profils phylogénétiques de grande qualité. Cette étape est essentielle afin de pouvoir analyser cette grande quantité de données. En particulier, les familles de protéines très conservées (actine) ou qui contiennent de longs domaines (protéines moteur) ou motifs (coiled-coil) conservés posent des problèmes de détection de faux positifs. La superfamille des actines est un exemple clair dont les profils phylogénétiques montrent une présence ubiquitaire (Figure 80).

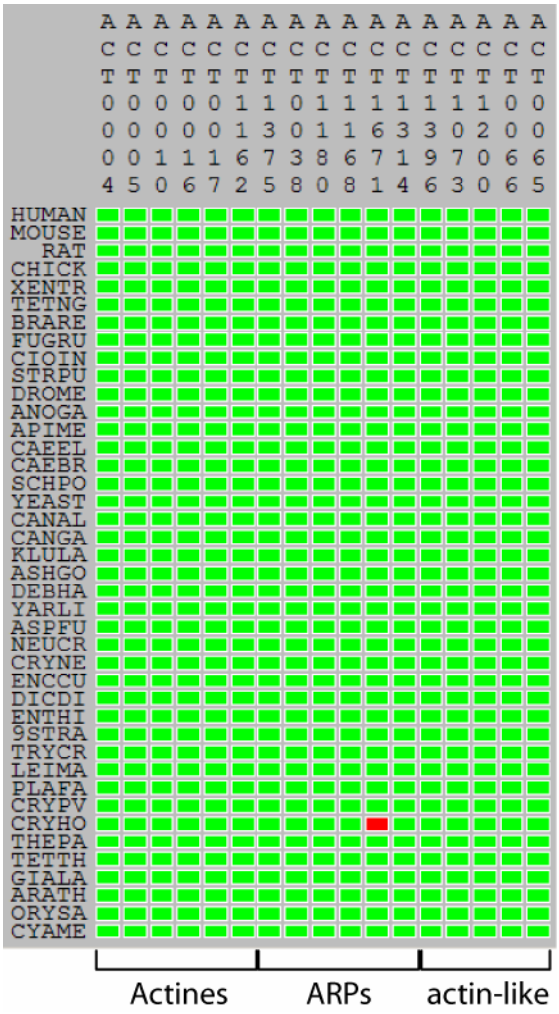


Figure 80 **Profils phylogénétiques des membres de la superfamille des Actine.**
Extrait du cluster 3 contenant les membres de la superfamille des actines

L'actine est connue pour son extrême conservation de séquence et sa présence dans l'ensemble des organismes eucaryotes. Cette distribution universelle n'est pas nécessairement partagée par les autres membres de cette famille comme les Actin-Related Protein (Goodson *et al.* 2002). L'analyse détaillée des résultats de la recherche de similarité montre effectivement la complexité de ce type de famille puisque ce sont près de 800 séquences qui sont simplement détectées avec une valeur d'expect inférieure à 10^{-3}. Au-delà du nombre important de séquences détectées, le problème vient de l'impossibilité de différencier certaines actines éloignées ou orphelines d'une ARP1, sur la seule base de la valeur d'expect. Dans le but de mieux répondre à ce type de problématique et de définir les véritables profils phylogénétiques de cette famille de protéines, une étude plus approfondie a été réalisée et est décrite dans le chapitre suivant (Chapitre 12 - Les Actin-Related Proteins).

Les faux négatifs sont également problématiques dans la définition des profils. Ils incitent à croire à tort en l'absence d'une protéine dans un organisme. Ce problème n'est pas directement lié aux caractéristiques des familles de protéines. En effet, une des raisons possibles est la mauvaise couverture des protéomes de certains organismes dans les banques protéiques. Toutefois, ce problème n'est pas irréversible car plusieurs solutions existent. La première consiste à attendre que les banques de protéines soient complétées au fur et à mesure des nouvelles versions. Cette solution n'est cependant pas très satisfaisante. La seconde solution consiste à revenir aux sources de l'information, c'est-à-dire le génome. En effet, la validité d'une absence ne peut etre assurée que par une recherche exhaustive directement dans le génome des organismes considérés. Cependant, ce type d'approche implique que l'on soit à même de différencier automatiquement les différents homologues d'un gène ou l'éventuelle présence de pseudogène.

11.4 Conclusion et perspectives

Ce travail préliminaire a permis de développer une plateforme d'analyse de génomique comparative sous la forme d'un programme (ComIcs) capable de créer des profils phylogénétiques et de les analyser. Nous avons ainsi appliqué cette plateforme aux protéines du cytosquelette décrites dans la version 3.0 de la banque Actinome. Sur la base des valeurs d'expect et l'utilisation des banques UniProt et RefSeq, 1742 profils phylogénétiques ont pu être établis de manière automatique dans 41 génomes eucaryotes. L'analyse de la dernière mise à jour de ces profils phylogénétiques en est encore à un stade préliminaire, mais un certain nombre de points ont pu être traités et des problèmes identifiés.

Dans notre cas, nous avons utilisé une valeur d'expect de 10^{-3} ce qui est peu restrictif mais a permis d'obtenir des résulats relativement cohérents. L'analyse des clusters de profils

requiert encore un travail important de validation et d'interprétation. L'interprétation sera effectuée en couplant des informations issues de l'expertise des biologistes au laboratoire et de l'enrichissement en annotations diverses comme les termes GO, les voies métaboliques ou de signalisation (KEGG). Cette analyse devra permettre de révéler des grands mouvements dans l'évolution des cytosquelettes, comme par exemple, l'apparition de telle ou telle fonction dans les vertébrés. Elle devra également révéler des associations entre certaines protéines co-présentes dans les mêmes organismes.

Dans toute expérience biologique ou bioinformatique, seule la présence d'un signal doit être considérée. En effet, l'absence de signal ne reflête pas forcément un non-signal mais parfois une limitation de la technique ou des données utilisées. Dans le contexte de la génomique, la présence d'une protéine (ou la fonction qu'elle représente) dans un organisme est liée à la detection d'une similarité.

Bien que l'utilisation des valeurs d'expect soit préconisée pour les procaryotes et pour *S. cerevisiae* leur utilisation peut poser un certain nombre de problèmes (Snitkin *et al.* 2006). Les profils phylogénétiques restent difficilement applicables ou tout du moins limités dans le cadre d'une analyse automatique (Snitkin *et al.* 2006). Il est ainsi important de connaître les limites de chaque technique et de définir des critères bornés en fonction de ce que l'on veut faire. ComIcs permet également d'utiliser le pourcentage d'identité et/ou le recouvrement pour discriminer les différents résultats. Une approche combinée de l'ensemble de ces paramètres pourront sans doute affiner les résultats.

Dans le cadre de familles de protéines très conservées, comme le cytosquelette en contient, la présence d'une protéine peut être faussement révélée. Une connaissance plus approfondie de ces familles de protéines doit alors être utilisée pour discriminer les résultats. Cependant ce type d'information est rarement connu ou disponible. L'alignement multiple peut alors être un filtre intéressant pour enrichir les informations disponibles. Ce travail constitue le point de départ de l'analyse approfondie de la superfamille des actines. Cette nouvelle analyse, décrite dans le prochain chapitre, s'inscrit dans la stratégie générale du laboratoire à savoir une compréhension des limitations des analyses automatiques en bioinformatique couplée à des solutions manuelles fines qui devront conduire à l'amélioration des traitements automatiques initiaux.

Un autre problème majeur représenté par les faux négatifs et nécessitent l'intégration d'un niveau de recherche supplémentaire, la recherche dans les génomes. Nous avons déjà développé dans ComIcs les étapes initiales de l'intégration de ce niveau, avec la possiblité d'effectuer des recherches dans les 41 genomes complets (par le programme TBLASTN). La mise à disposition des séquences des génomes complets et leur mise à jour a été le fruit d'une

collaboration avec la plateforme de bioinformatique de Strasbourg. Un protocole de détection automatique des séquences dans les génomes devra encore y être intégré pour parfaire la définition des profils phylogénétiques.

Un nombre important de développements peut encore être attaché à la plateforme ComIcs. L'intégration de la recherche génomique en est un. Une approche utilisant plusieurs séquences intiales issues de plusieurs organismes éloignés permettrait de s'abstraire des problèmes de détection liés à l'utilisation d'une seule séquence d'un organisme en particulier. L'intégration des outils récents, comme MACSIMS (Thompson *et al.* 2006) combinant l'alignement multiple et la détection automatique des domaines peut également aider à valider la présence des protéines dans certains organismes.

Chapitre 12 - Les Actin-Related Proteins

« Rien en biologie n'a de sens, si ce n'est à la lumière de l'évolution »

Theodosius Dobzhansky

Les Actin-Related Proteins ou ARPs sont des membres d'une large famille de protéines homologues nommée superfamille des actines. Cette superfamille constitue le cœur du cytosquelette d'actine. Elle est caractérisée par une conservation importante des séquences qui la compose à la fois au sein d'un même organisme, mais également entre les séquences d'organismes distincts. Cette conservation de séquences constitue un véritable bruit de fond dans les recherches de similarité. Ce bruit de fond devient un problème pour clairement distinguer les ARPs entre elles ou même une ARP d'une actine. Comme nous l'avons dans le chapitre précédent, une autre conséquence est la création de profils phylogénétiques erronés par la détection de faux positifs. Cette famille de protéines pose donc de nombreux problèmes bioinformatiques nécessitant un traitement particulier.

Ce travail a donné lieu à une publication en 2005 (Muller *et al.* 2005) présentée à la fin du chapitre. Après la présentation de la superfamille des actines, je présenterai les données obtenues et préciserai certaines hypothèses émises au moment de la publication. Je discuterai également de certains aspects de l'utilisation du serveur web ARPAnno.

12.1 La superfamille des Actines

La superfamille des actines est une famille essentielle du cytosquelette d'actine. Cette superfamille est constituée des actines à proprement parlé, des Actin-Related Proteins (ARP) et de plusieurs Actin-Like Proteins (ALP). Les ALPs bien que de plus en plus nombreuse sont souvent spécifiques d'un seul organisme et ne seront pas traitées dans ce manuscrit (Chadwick *et al.* 1999; Eichinger *et al.* 2005; Gordon *et al.* 2005; Kuribara *et al.* 2006). Elles sont souvent assimilées à des membres orphelins de la superfamille.

La molécule d'actine constitue le composant principal des microfilaments. L'actine est une protéine ubiquitaire de 375 acides aminés trouvée dans tous les organismes vivants. Selon les organismes, il peut exister une ou plusieurs isoformes ; l'Homme par exemple dispose de 6 isoformes qui sont exprimées de façon tissu spécifique (Khaitlina 2001) (Tableau 12).

Définition	Nom de gène HUGO	Numéros d'accession (RefSeq)
actin alpha 1 skeletal muscle	ACTA1	NM_001100
actin alpha 2 aortic smooth muscle	ACTA2	NM_001613
actin beta	ACTB	NM_001101
actin alpha cardiac muscle	ACTC	NM_005159
actin gamma 1	ACTG1	NM_001614
actin gamma 2 enteric smooth muscle	ACTG2	NM_001615

Tableau 12 Liste des isoformes d'actine chez l'homme.

Une des caractéristiques principales des actines réside dans l'extrême conservation des séquences à la fois intra-espèces, puisque les isoformes partagent de 93 à 99% d'identité, et à la fois inter-espèces puisque par exemple l'actine alpha de muscle squelettique humaine est strictement identique à celles de la souris, du rat, du bœuf et du lapin.

Au niveau nucléique, la conservation est également très importante puisque les ARNm des différentes isoformes d'actine chez l'homme partagent jusqu'à 99% d'identité sur leurs parties codantes (CDS) (Tableau 13). Cette caractéristique constitue notamment une difficulté lors du choix de sondes effectué lors de la réalisation de la puce à ADN Actichip (Chapitre 14 - Actichip une puce dédiée au cytosquelette).

Gène	Taille (bases) mRNA	% identité mRNA	% recouvrement CDS/mRNA	% identité CDS
ACTA1	1509	100,00%	100,00%	100,00%
ACTB	1793	86,00%	61,29%	97,43%
ACTG1	1919	86,00%	57,22%	97,34%
ACTC	1512	85,00%	68,32%	91,09%
ACTG2	1345	84,00%	83,57%	99,38%
ACTA2	1330	85,00%	80,68%	94,70%
Moyenne		87,67%	75,18%	96,66%

Tableau 13 Pourcentage d'identité (au niveau nucléique) des différentes isoformes d'actine chez l'homme.

Les pourcentages d'identité ont été calculés à la fois sur la partie codante (la partie la plus conservée) et sur la séquence complète des ARNm. Les pourcentages d'identité ont été caclulés en utilisant une seule et même référence, l'actine de muscle squelettique humaine (ACTA1). La fraction de la partie alignée entre la séquence de l'isoforme d'actine et la référence est exprimée par le recouvrement.

Cette grande identité de séquence est d'autant plus intéressante qu'elle se traduit par la conservation au cours de l'évolution d'une même architecture structurale appelée « actine fold » (Figure 81) (Bork *et al.* 1992; Holmes *et al.* 1993; Kabsch *et al.* 1995). L'actine fold consiste en 2 domaines organisés autour d'une poche centrale de fixation de l'ATP.

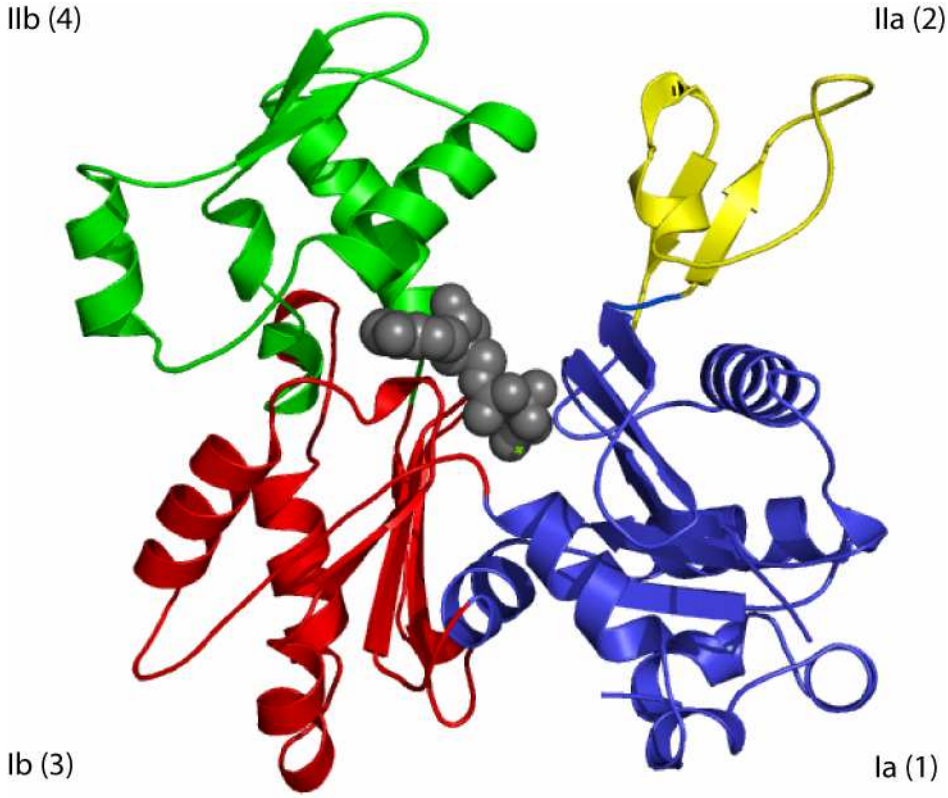


Figure 81 **Structure de l'actine.**
La structure tridimensionnelle de l'actine représentée est celle de la levure ***S. cerevisiae*** (code PDB 1yag) dans son orientation standard. Les 4 sous domaines sont représentés ; Ia ou 1 en bleu, IIa ou 2 en jaune, Ib ou 3 en rouge et IIb ou 4 en vert. La molécule d'ATP est représentée en gris foncé en position centrale dans la poche de fixation du nucléotide.

Les 2 domaines nommés I et II sont divisés en 4 sous-domaines ; Ia ou 1, IIa ou 2, Ib ou 3 et IIb ou 4 (Figure 81). Les domaines 1 et 3 définissent la face de la molécule présentée à l'extrémité barbée d'un filament d'actine alors que les domaines 2 et 4 définissent l'extrémité pointue du filament (voir 1.2.1 Les filaments d'Actine).

L'actine est une protéine du cytosquelette qui possède un nombre important de partenaires. Le répertoire des ABP est composé de plusieurs architectures structurales déclinées dans plusieurs familles de protéines différentes avec des effets différents sur l'actine (Van Troys *et al.* 1999). Une des faces de contact majeures de l'actine avec ses partenaires est la cavité hydrophobe (revue dans (Dominguez 2004)). Cette partie de la molécule est notamment responsable de la fixation de la gelsoline, de la Vitamin D-binding protein (DBP), des actin-depolymerizing factor/cofilin (ADF/cofiline), de la protéine ciboulot et des interactions actine-actine dans les microfilaments.

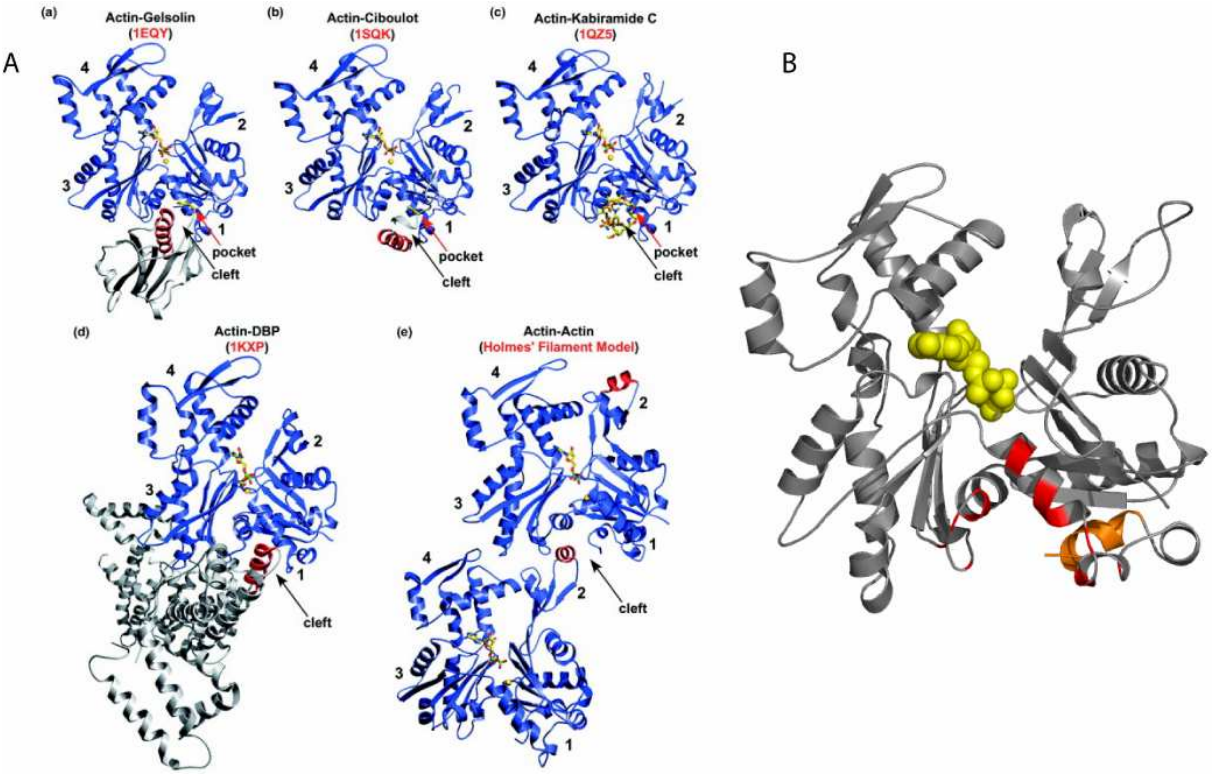


Figure 82 **La cavité hydrophobe, une interface de fixation avec les ABPs.**
(A) Mise en évidence des interactions existantes entre l'actine et certaines ABPs (gelsoline, Vitamin D-binding protein, ADF/cofiline, ciboulot et des interactions actine-actine) au travers de la cavité hydrophobe (Dominguez 2004). (B) Mise en évidence des régions participant à la définition de la clé hydrophobe (rouge). La partie C-terminale (orange) pourrait également participer à la définition de cette région (Dominguez 2004).

12.2 Les ARPs

Les ARPs sont des protéines similaires à l'actine retrouvées dans plusieurs organismes. La classification unifiée des ARPs a été proposée en 1994 par Schroer *et al.* (Schroer *et al.* 1994) et complétée en 1997 par Poch et Winsor (Poch *et al.* 1997). Cette seconde étude est basée sur la séquence complète du génome de la levure *S. cerevisiae* et a permis de définir 10 sous-familles d'ARP. Cette classification s'appuie sur les pourcentages d'identité et de similarité des différentes ARPs par rapport à une séquence d'actine conventionnelle. Ainsi, ARP1 est la séquence la plus proche de l'actine alors qu'ARP10 en est la plus éloignée. En 1997, seules les ARP1 à 3 étaient définies par plusieurs séquences issues de différents organismes. Les familles d'ARP4 à 10 étaient basées sur uniquement 1 ou 2 séquences. Ces familles ont été depuis validées à plusieurs reprises dans des organismes modèles différents (Eckley *et al.* 1999; Harata *et al.* 2001; Goodson *et al.* 2002). A ce jour, cette classification semble se maintenir. Une seule nouvelle famille d'ARP, ARP11, a été identifiée par Eckley *et al.* chez les métazoaires (Eckley *et al.* 1999). ARP11 est, à l'image d'ARP10, très éloignée de l'actine. Il

faut également noter qu'ARP7, ARP9 et ARP10 sont uniquement présents dans les champignons (comme *S. cerevisiae*).

A la vue de la proximité de séquences des ARPs avec l'actine, il était raisonnable de penser que l'architecture structurale (actine fold) soit également conservée. Ceci a été vérifié récemment pour ARP2 et ARP3 dont les structures ont été résolues dans plusieurs états (Robinson *et al.* 2001; Volkmann *et al.* 2001; Nolen *et al.* 2004; Rodal *et al.* 2005) (Figure 83). Cette conservation est d'autant plus intéressante que certaines ARPs possèdent des insertions de séquences plus ou moins conséquentes (de 10 à quelque centaines de résidus). Ces insertions, initialement observées chez *S. cerevisiae* (Poch *et al.* 1997), ne sont pas identifiées clairement pour chaque sous-famille d'ARP et leurs fonctions sont encore inconnues. Néanmoins, la présence d'insertions chez les ARPs peut d'une part, constituer un élément distinctif par rapport aux actines, et d'autre part, conférer une certaine plasticité à l'actine fold.

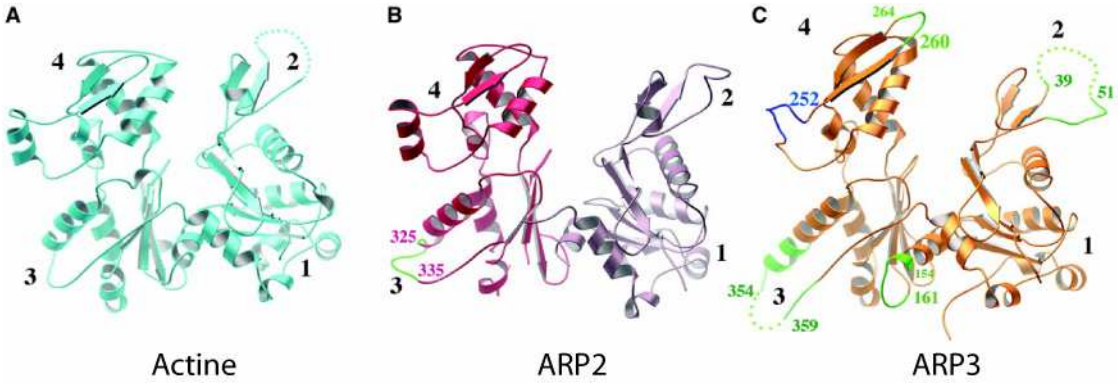


Figure 83 **Comparaison des structures de l'actine, ARP2 et ARP3.**
(A) Structure de l'actine de muscle squelettique de lapin (code PDB 1ATN). (B) ARP2 (code PDB 1K8K). Les sous domaines 1 and 2 ne sont pas présents dans la carte de densité électronique et ont ainsi été modélisés à partir de l'actine et positionnés en fonction de la structure du complexe ARP2/3. (C) ARP3 (code PDB 1K8K). Les insertions sont matérialisées en vert. Les zones absentes des cartes de densité sont en pointillées (adapté de (Robinson *et al.* 2001)).

D'un point de vue fonctionnel, les ARPs sont impliquées à la fois dans des mécanismes classiques du cytosquelette (nucléation et transport) et dans des mécanismes totalement différents, comme la régulation de la transcription. Ces nouvelles fonctions permettent d'étendre la palette des possibilités du cytosquelette d'actine.

Ainsi, on retrouve ces protéines à la fois dans le cytoplasme et dans le noyau. ARP1-3 et, plus récemment ARP10 et ARP11 sont essentiellement localisées dans le cytoplasme (Schafer *et al.* 1999; Machesky *et al.* 2001; McKinney *et al.* 2002). On les qualifie alors d'ARPs cytoplasmiques. A l'opposé, les ARP4-9 sont qualifiées d'ARPs nucléaires car elles sont localisées majoritairement dans le noyau de la cellule. Elles participent notamment à des

fonctions comme le remodelage de la chromatine et la régulation de la transcription (Weber *et al.* 1995; Grava *et al.* 2000; Harata *et al.* 2000; Olave *et al.* 2002; Blessing *et al.* 2004).

Les ARPs sont rarement monomériques et jouent souvent un rôle prépondérant au sein de complexes multi-protéiques. Elles sont en général associées par paires. Ainsi, ARP1 et ARP11 sont des constituants du complexe dynactine (composé de 11 sous-unités) impliqué dans le transport de cargos et d'organelles le long des microtubules (Eckley *et al.* 1999; Eckley *et al.* 2003). ARP2 et ARP3 sont les principaux constituants du complexe de nucléation des microfilaments. Le complexe ARP2/3 contient 7 sous unités. Pour les ARPs nucléaires, les combinaisons sont un peu plus complexes puisque autour d'ARP4, on associera différents partenaires en fonction de l'organisme et du complexe (Tableau 14). ARP4 est donc co-présent avec ARP5 et ARP8, avec ARP6 et avec l'actine. Bien que le rôle des ARPs nucléaires soit encore peu connu, elles sont cependant essentielles à l'activité enzymatique des complexes auxquelles elles appartiennent (Galarneau *et al.* 2000; Gorzer *et al.* 2003; Shen *et al.* 2003). ARP7 et ARP9 ont également été décrits comme des partenaires obligatoires chez *S. cerevisiae* mais leur activité n'est pas essentielle au complexe RSC (*Remodel the Structure of Chromatin*) (Szerlong *et al.* 2003). A l'inverse des ARPs cytoplasmiques, les mécanismes fonctionnels des ARPs nucléaires et leur emplacement dans les complexes n'ont pas encore été élucidés.

L'analyse de la famille des ARPs est un exemple des difficultés existantes en bioinformatiques pour caractériser un ensemble de protéines. C'est une famille extrêmement intéressante qui autour d'une architecture spatiale commune a su développer au sein de la cellule des fonctions diverses et majeures.

Tableau 14 Les différentes associations d'ARPs dans les complexes nucléaires.

L'ensemble des données sont extraites des publication suivantes (Cairns *et al.* 1998; Papoulas *et al.* 1998; Peterson *et al.* 1998; Zhao *et al.* 1998; Galarneau *et al.* 2000; Ikura *et al.* 2000; Nie *et al.* 2000; Shen *et al.* 2000; Fuchs *et al.* 2001; Kuroda *et al.* 2002; Olave *et al.* 2002; Kitagawa *et al.* 2003; Mizuguchi *et al.* 2004; Mohrmann *et al.* 2004; Cai *et al.* 2005).

Organisme	Type	Complexe	actine	ARP4	ARP5	ARP6	ARP7	ARP8	ARP9
S.cerevisiae	chromatin remodeling complex	SWI/SNF					1		1
		RSC					1		1
		Ino80	1	1	1			1	
		SWR1	1	1		1			
	complexe HAT	NuA4	1	1					
D.melanogaster	chromatin remodeling complex	BAP	1	1					
		PBAP	1	1					
H.sapiens	chromatin remodeling complex	BAF, SWI/SNF-A	1	1					
		SWI/SNF (bBAF)	1	1					
		PBAF (SWI/SNF-B)	1	1					
		WINAC		1					
		p400	1	1					
		SRCAP		1		1			
	complexe HAT	TIP60	1	1					

12.3 L'alignement multiple de séquences complètes

Comme nous l'avions envisagé suite à l'étude globale des profils phylogénétiques (Chapitre 11 - Une approche globale : Actinome), afin de mieux comprendre la famille des ARPs et pour caractériser chacune des sous-familles, nous avons décidé de construire un alignement multiple des séquences complètes de toutes les ARPs disponibles et de nombreuses actines et actine-like. Cet outil nous permet d'intégrer les caractéristiques majeures d'une famille de protéines comme l'organisation en domaines des protéines, la distribution des insertions et des délétions, la conservation de résidus à des positions particulières, et la diversité des organismes disponibles (Lecompte *et al.* 2001).

Dans ce but, nous avons recherché de manière exhaustive l'ensemble des séquences similaires dans les banques de protéines à partir de plusieurs séquences d'ARP et d'actine distinctes et d'organismes aussi variés que la drosophile, la levure et l'homme. Dans ce type de recherche, le problème du bruit de fond lié à la forte similarité devient un avantage et nous permet d'être le plus complet possible.

Les séquences ainsi extraites sont alignées par PipeAlign (Plewniak *et al.* 2003) afin de constituer un seul alignement multiple de séquences complètes. Cet alignement multiple, appelé ARP-MACS, a ensuite été corrigé manuellement en tenant compte des structures secondaires pour obtenir un alignement de haute qualité. L'alignement des ARPs est disponible sur le site http://bips.u-strasbg.fr/ARPAnno/ARPMACS.html.

ARP-MACS contient 692 séquences dont 148 ARPs. Ces 148 ARPs sont à comparer aux 29 séquences disponibles en 1997 (Poch *et al.* 1997). L'analyse de ces 148 ARPs a révélé la consistence de la classification unifiée.

12.3.1 Caractérisation des sous familles

Le bilan des insertions et délétions par rapport à la séquence d'actine humaine prise comme référence est déterminé pour chaque sous-famille d'ARPs (cf Figure 2 de l'article (Muller *et al.* 2005)). Ce bilan des indels mais également des résidus caractéristiques par famille a de nombreuses retombées : la compréhension de l'evolution séquence/structure de la famille, une base solide pour faire un profil de présence/absence, une base pour le développement d'ARPAnno… Il est emblématique des possibilités offertes par ce type d'études.

Ce bilan a mis en évidence le nombre important d'insertions et le nombre réduit de délétions. D'un point de vue fonctionnel, ces insertions peuvent être capitales et définir de nouveaux

domaines. Cependant, à ce jour, aucune donnée bioinformatique ou biologique n'a encore permis d'élucider précisément les fonctions de ces insertions/délétions.

La distribution de ces insertions sur la séquence de référence a montré leur compatibilité avec la conservation de l'actine fold car elles sont souvent localisées dans des boucles situées sur les faces extérieures de la molécule et un regroupement en 4 points chauds. Ces points chauds constituent a priori des zones de plasticité plus importante de l'actine fold.

Ce bilan a permis également de déterminer les insertions propres à chaque sous-famille. Complété par la détermination de résidus discriminants pour certaines familles, ces 2 bilans (insertions et résidus discriminants) déterminent un ensemble de connaissances essentielles pour caractériser et filtrer chacune des sous-familles d'ARP.

Ce filtre de connaissances est la base à partir de laquelle nous avons pu définir ARPAnno (voir Chapitre 8 - ARPAnno). ARPAnno est un serveur d'annotation des séquences proches de l'actine. Il permet de déterminer, au moyen d'un score et la consultation d'un alignement multiple, à quelle sous-famille appartient une séquence soumise. Cet outil est devenu un moyen efficace pour intégrer l'ensemble des connaissances recueillies sur les sous-familles d'ARPs. Ainsi, nous avons mis à profit la richesse des séquences recherchées dans les banques de séquences, la qualité de l'alignement multiple et le filtre de connaissances.

12.3.2 Les profils phylogénétiques

Les profils phylogénétiques des ARPs et des actines posent un certain nombre de problèmes par l'introduction de faux positifs, comme nous avons pu le dire précédemment. La détermination des profils phylogénétiques réels des ARPs a permis de dégager plusieurs enseignements généraux.

- Premièrememt, l'utilisation de l'alignement est un filtre nécessaire et fonctionnel.
- Deuxièmement, la validation des présences et absences dans les gènomes est possible mais fastidieuse.
- Troisièmement, les ARPs constituent un exemple probant de la corrélation possible entre les profils phylogénétiques et leur biologie.

Ainsi, la caractérisation des séquences d'ARPs au moyen de l'alignement multiple, nous a permis de déterminer dans les banques de protéines la distribution des différents types d'ARP par organisme. Néanmoins, pour déterminer de manière fine leurs profils phylogénétiques, il est nécessaire de valider leur présence/absence dans les différents

génomes complets disponibles. ARPAnno a été mis à profit à chaque étape de cette validation en discriminant les séquences prédites dans les différents génomes.

La comparaison entre les profils initiaux déterminer en automatique et ceux obtenus dans le cadre de cette étude, illustre la différence flagrante (Figure 84 et Figure 85) qui existe entre les 2 types de profils.

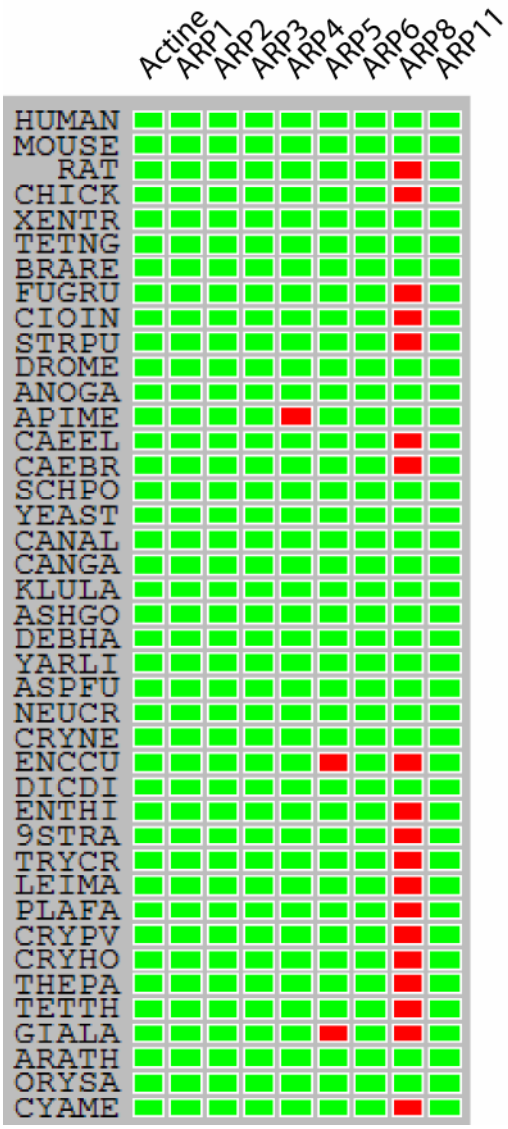


Figure 84 **Profils phylogénétiques d'une actine et des ARPs humaines.**
Les profils sont ceux obtenus de façon automatique avec ComIcs en utilisant une limite de détection inférieure à une valeur expect de 10^{-10}

L'étude des profils phylogénétiques des sous-familles d'ARP est une illustration parfaite de la corrélation entre la co-présence ou la co-absence de partenaires d'un même complexe protéique. Ainsi, comme nous l'avons décrit précédemment les ARPs sont souvent présentes par paires dans les complexes protéiques. Ces paires fonctionnelles connues peuvent être

déduites de l'analyse des profils et sont corrélées de manière quasi parfaite pour ARP2 et ARP3, ARP4 et ARP6, ARP5 et ARP8, ARP7 et ARP9 (Figure 85).

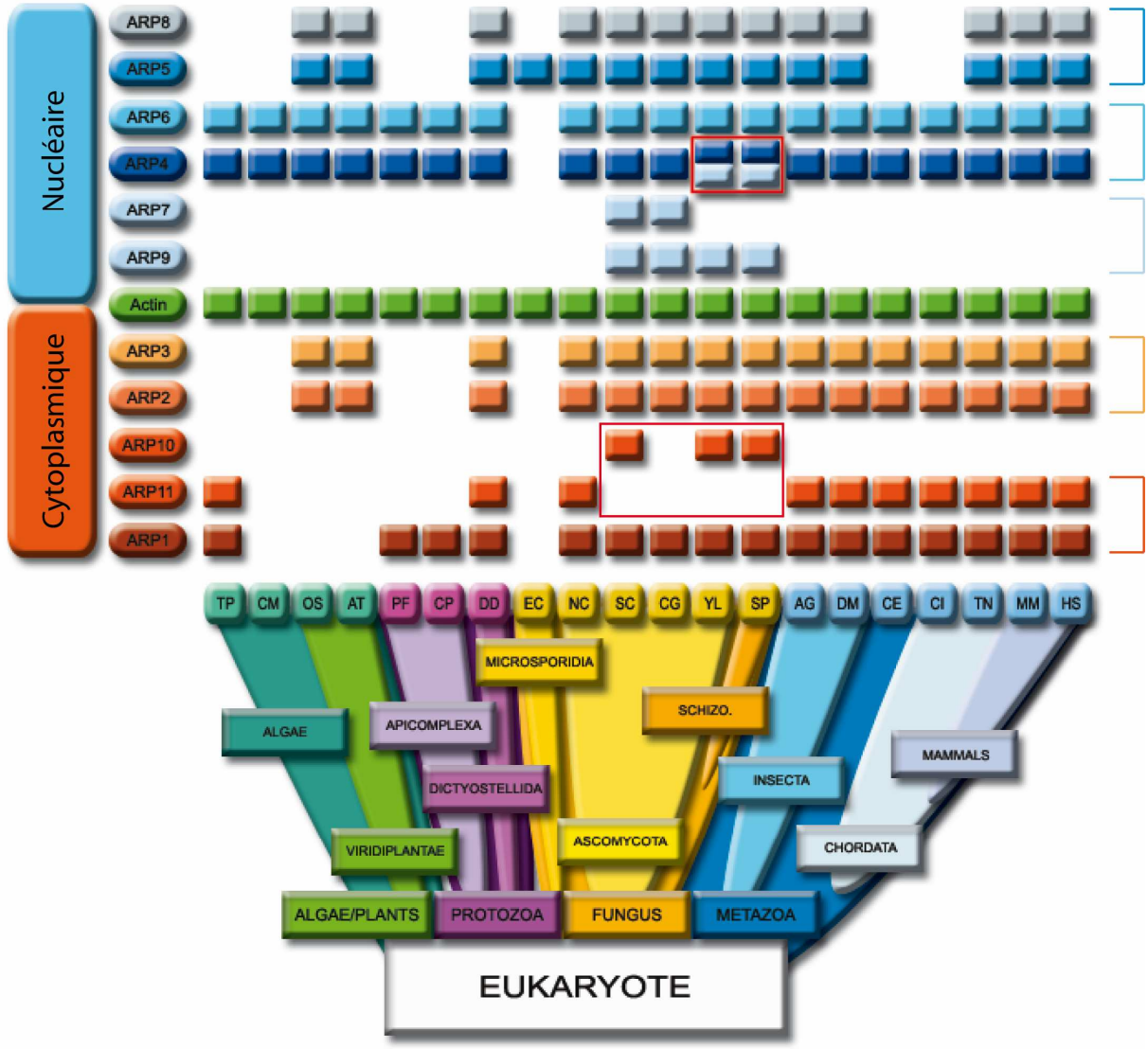


Figure 85 **Profils phylogénétiques des ARPs**
Les ARPs ont été regroupées par paires d'ARPs fonctionnelles (cf les accolades à droite). L'actine est représentée en position centrale (vert) et délimite la séparation entre les ARPs dites cytoplasmiques (ARP1, ARP2, ARP3 ARP10 et ARP11) et les ARPs dites nucléaires (ARP4, ARP5, ARP6, ARP7, ARP8 et ARP9). Les profils des ARP10 et ARP11 peuvent être regroupés aux vues de données biologiques récentes.

Dans une moindre mesure, une ARP11 est toujours présente en même temps qu'une ARP1 (bien qu'ARP1 soit observée dans beaucoup plus d'organismes). Nous avions déjà suggéré dans l'article, l'ambiguïté entre les 2 sous-familles ARP10 et ARP11, sans avoir pu décider (faute d'argument vraiment décisif) entre le fait de les regrouper en une seule sous-famille ou de conserver cette séparation. En 2005, seule ARP11 avait été clairement identifiée avec ARP1 dans le complexe dynactine chez les métazoaires. De façon troublante, bien qu'aucune indication de séquence ne regroupait les ARP10 et ARP11 (mise à part le fait qu'elles soient

toutes les deux très divergentes de l'actine et qu'elles présentent de grandes délétions), leurs profils phylogénétiques se complémentaient quasi parfaitement. La réponse définitive est venue en 2006 avec l'identification d'ARP10 dans le complexe dynactine de *S. cerevisiae* (Clark *et al.* 2006). ARP10 et ARP11 ne forment ainsi qu'une seule et même sous-famille.

Depuis plusieurs années, le nombre d'études décrivant les ARPs nucléaires s'est accrue ; démontrant leur appartenance à des complexes protéiques ou encore leurs implications dans des fonctions nucléaires majeures. Les ARPs cytoplasmiques, dont les fonctions sont directement reliées au cytosquelette, ont depuis longtemps retenues l'attention des biologistes. En particulier, ARP2 et ARP3, responsables de la nucléation et le branchement des filaments d'actine, apparaissaient indispensables à la cellule eucaryote. Cependant, ces 2 ARPs sont notamment absentes des apicomplexes (Figure 85 et (Gordon *et al.* 2005)) et de certaines algues. Or, d'autres nucléateurs (les formines et spires voir 1.2.1 Les filaments d'Actine) ont été identifiés.

Notre étude amplifie encore l'importance des ARPs nucléaires. En effet, contre toute attente, nous avons pu identifier ARP4 et ARP6 comme les seules ARPs présentes dans tous les organismes eucaryotes testés, à l'exception du parasite *E. cuniculi*. Ainsi, les ARPs universelles sont des ARPs nucléaires. Parmi ses 2 ARPs, ARP4 apparaît centrale si on considère sa position particulière dans la composition des complexes nucléaires contenant des ARPs. Ainsi, on la retrouve dans tous les complexes avec à la fois ARP6, l'actine, ARP5 et ARP8.

Dans la superfamille des actines, l'existence de plusieurs membres universels, l'actine ARP6 et ARP4, soulève la question de leur origine et de l'origine de l'actine moderne dans les organismes eucaryotes. De part son implication plus large, ARP4 apparaît d'autant plus importante et soulève certaines interrogations. ARP4 et ARP6 aurait-elle un homologue structurale et fonctionnel chez les procaryotes ? Les ARP4 et les actines ont-elles divergé à partir d'un même ancêtre commun ? Ceci est très probable. Quand a eu lieu cet évènement ? Avant ou après la séparation procaryotes/eucaryotes ?

12.4 De l'utilisation du serveur ARPAnno

12.4.1 Statistiques du serveur

ARPAnno a été mis en service en juin 2005 et sa publication officielle date de décembre 2005. Depuis cette date, 363 requêtes différentes ont été soumises à ARPAnno concernant 62 utilisateurs différents (Figure 86). Le temps moyen d'une requête est inférieur à 1 minute et représente pour l'ensemble des 363 requêtes un peu plus de 6 heures de temps de calcul. L'analyse de la répartition des requêtes et du nombre d'utilisateurs montre une corrélation entre les dates de référencement du serveur au sein de PubMed (octobre 2005) et de sa publication dans le journal (25 décembre 2005) et les plus fortes demandes.

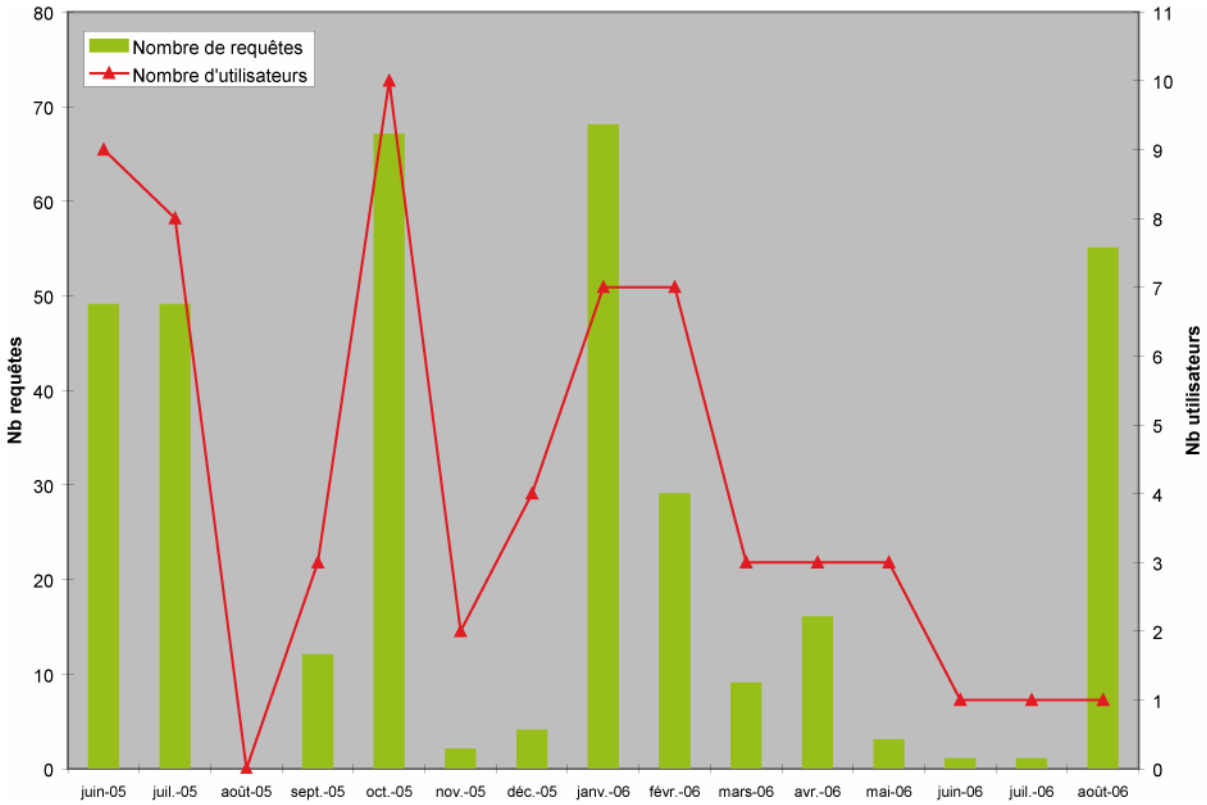


Figure 86 **Distribution du nombre de requêtes et du nombre d'utilisateurs du serveur web ARPAnno pour la période d'utilisation de juin 2005 à août 2006.**

12.4.2 Quelles séquences ? Quels organismes ?

Afin de mieux observer l'utilisation du serveur ARPAnno, on peut analyser le type de requêtes soumises (Figure 87 A). On notera plusieurs points importants :

- Le serveur a été très disponible et a traité la plupart des requêtes soumises avec seulement 5 requêtes (1,4%) non abouties (catégorie « Problem »).

- 237 des requêtes (2/3) correspondent à des prédictions de séquence appartenant à des groupes d'ARP (catégorie « Prediction »).
- Le nombre de séquences soumises et identifiées directement (100% identiques) comme appartenant à la catégorie Actin ou Orphan représente environ 18% (69) des requêtes.
- 42 requêtes (~14%) concerne des séquences n'appartenant pas à la superfamille des actines.

Il est également intéressant de constater la diversité du type et d'origine des séquences soumises au serveur ARPAnno (Figure 87 B). En effet, du point de vue des prédictions déterminées par le serveur (catégorie « Prediction » dans la Figure 87 A), les utilisateurs ont soumis un bon nombre de familles d'ARP différentes (ARP1-6, ARP8 et ARP11). Seules 3 familles d'ARPs, ARP7, ARP9 et ARP10, n'ont pas encore été soumises par des utilisateurs mais il est important de noter que ces 3 types d'ARPs sont également les moins représentées dans les organismes connus. Encore une fois, il est surprenant de voir que bon nombre des soumissions concernaient des actines conventionnelles. Cependant, ces actines conventionnelles sont des formes plus éloignées ou nouvellement apparues dans les banques de séquence protéiques.

Parmi les catégories d'organismes soumis à ARPAnno, on remarquera à la fois des génomes séquencés et publiés comme *Homo sapiens*, *Gallus gallus*, *Xenopus tropicalis*, *Cyanidioschyzon merolae*, *Cryptosporidium hominis*, *Arabidopsis thaliana*, *Oryza sativa*, *Dictyostelium discoideum*, *Saccharomyces cerevisiae*, *Plasmodium falciparum*, *Trypanosoma cruzi*, *Trypanosoma brucei*, ou encore *Leishmania major*, mais également un nombre important de prédiction de gènes à partir d'organismes en cours d'annotation ou de séquençage comme *Paramecium tetraurelia*, *Tetrahymena thermophila*, *Volvox carteri*, *Dunaliella salina* et *Chlamydomonas reinhardtii*.

On notera également un intérêt marqué pour les organismes isolés dont aucun programme de séquençage n'est en cours, mais qui ont soulevé l'intérêt comme des algues ou des plantes ; *Ernodesmis verticillata*, *Ostreobium quekettii*, *Boergesenia forbesii*, *Siphonocladus tropicus*, *Caledenia catenata*, *Coelothrix*, *Brassica oleracea*, *Chlamydomonas moewusii*, *Thalassiosira weissflogii* et *Selaginella apoda*.

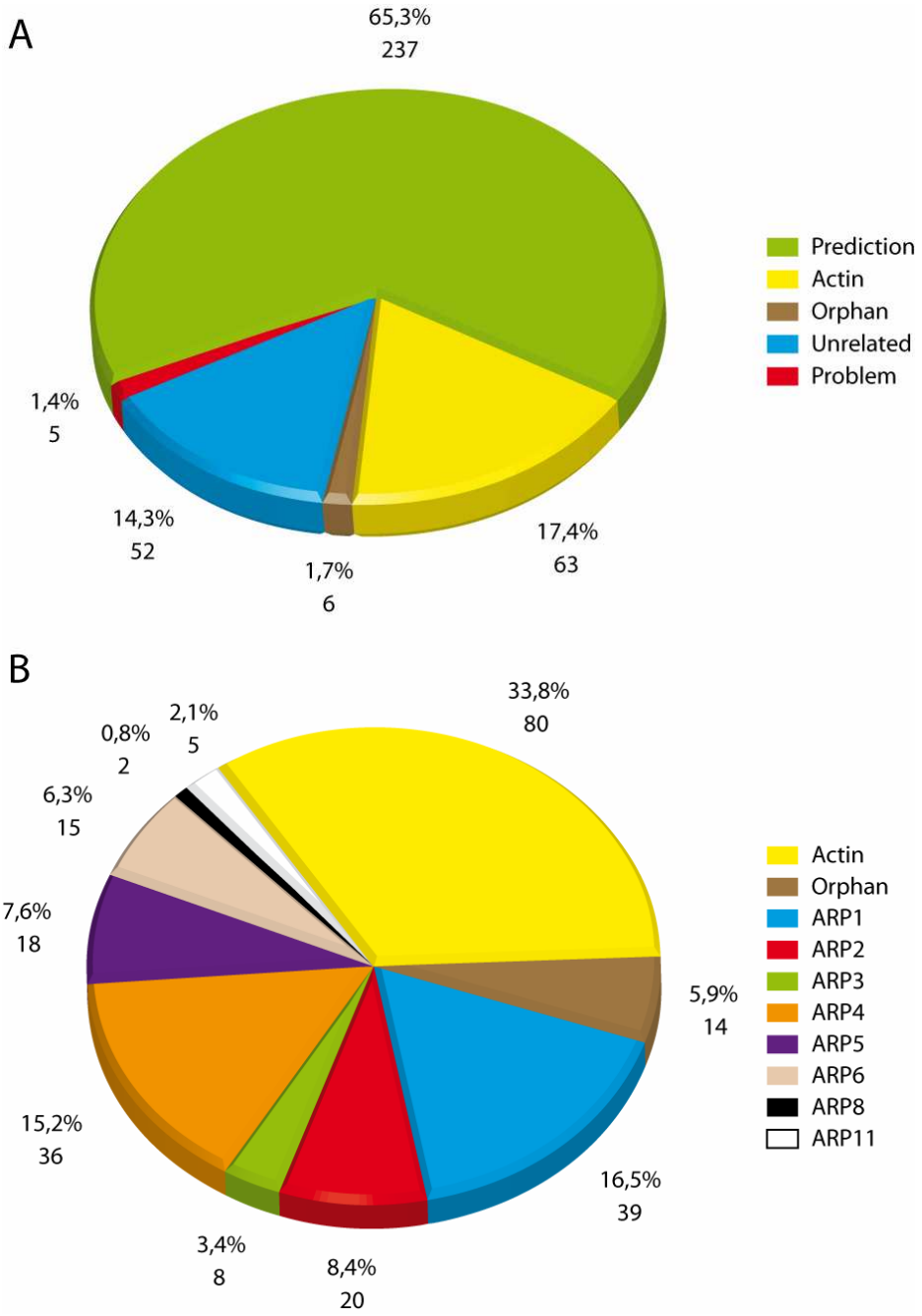


Figure 87 **Répartition du type de requêtes soumises à ARPAnno.**
(A) les requêtes sont catégorisées en 5 types ; « Prediction » (donnant lieu à une analyse approfondie), « Actin » et « Orphan » (pour une soumission d'une séquence 100% identiques à l'une de ses 2 catégories), « Unrelated » (soumission d'une protéine n'appartenant pas à la superfamille des actines) et enfin « Problem » (requête dont le déroulement s'est interrompue brutalement, souvent lié à des problèmes informatiques). (B) Répartition des résultats des requêtes de la catégorie « Prediction » (A). La contribution de chaque catégorie est indiquée ainsi que sa valeur réelle.

12.5 Conclusion et perspectives

La superfamille des actines et en particulier les ARPs, constitue une famille de protéines problématique en bioinformatique. Bon nombre d'analyses ou de recherches effectuées par des programmes bioinformatiques automatiques sont faussées par la grande similarité de séquences partagée par les membres de cette superfamille. Face à cette problématique, nous avons entrepris une analyse complète de la famille des ARPs. Cette étude nous a permis de mieux caractériser chacune des sous-familles et de dégager un ensemble de caractéristiques discriminantes. L'application directe de ces résultats est la mise à disponibilité de 2 outils essentiels en bioinformatique : ARP-MACS, l'alignement multiple de l'ensemble des séquences d'ARP et d'actines disponibles et ARPAnno un serveur d'annotation capable d'identifier les sous-familles d'ARPs. Les efforts consentis pour caractériser de manière fine cette famille de protéines montrent l'importance des problèmes posés et identifiés par la génomique comparative. L'utilisation de l'alignement multiple s'est révélée primordiale pour résoudre ces problèmes. Il constitue une piste sérieuse pour valider l'ensemble des profils phylogénétiques.

On notera que cette famille de protéines constitue un véritable cas-test ou « *benchmark* » pour de nombreuses applications en bioinformatique au laboratoire. Les analyses que nous avons menées sont des exemples emblématiques de ce que la bioinformatique peut apporter dans la compréhension d'une famille de protéines.

Depuis 2005, un certain nombre de nouvelles séquences d'ARPs se sont accumulées dans les banques de séquences protéiques et de nouveaux génomes sont rendus disponibles. Une mise à jour de l'alignement multiple et la validation du filtre de connaissance doit être entreprise. L'ensemble de ces informations sera par la suite intégré à ARPAnno et permettra une identification encore plus sûre des différentes sous-familles.

ARPAnno est un serveur actif depuis maintenant plus d'un an. L'utilisation qu'en font les biologistes montre la diversité et le nombre des applications couvertes par ARPAnno. Ils considérent à la fois des séquences issues des projets de séquençages de nouveaux organismes et des séquences provenant d'organismes modèles dont le génome complet est disponible. Ceci montre l'importance de ce type d'outils dédié à des familles de protéines complexes et dont une connaissance approfondie est nécessaire.

Une procédure automatique de recherche de nouvelles séquences a déjà été mise en place par l'utilisation de DbW (Prigent *et al.* 2005). L'ensemble des séquences ainsi détectées doit

être soumis à ARPAnno pour identifier la sous-famille correspondante, puis être intégré à ARP-MACS.

Enfin, bon nombre de prédictions émises au cours de ces travaux constituent autant d'hypothèses sur lesquelles les biologistes pourront baser leurs travaux.

Publication N°1

Chapitre 13 - Application à une maladie génétique : BBS

« Le surnaturel baisse comme un lac qu'un canal épuise;
la science à tout moment recule les limites du merveilleux. »
Guy de Maupassant

13.1 Présentation du syndrome Bardet-Biedl

Observé par Laurence et Moon dès 1866, le Syndrome de Bardet-Biedl (BBS) (OMIM 209900) a été décrit en 1920 par Georges Bardet (Bardet 1920) puis en 1922 par Arthur Biedl (Biedl 1922). Ce syndrome est une maladie autosomique récessive caractérisée par divers signes cliniques dont les plus importants sont : la rétinopathie pigmentaire précoce, l'obésité, la polydactylie, l'hypogonadisme, le déficit cognitif variable et les anomalies rénales (Figure 88).

Figure 88 **Phénotypes associés au BBS.**
(A) Photo d'une patiente atteinte du syndrome de Bardet-Biedl caractérisée par l'obésité. (B) Fond d'œil caractéristique d'une rétinite pigmentaire. (C) Exemple de polydactylie.

Le BBS est une affection rare. Sa prévalence dans des familles européennes est évaluée à 1 enfant sur 150000. Cette prévalence est plus importante dans les familles consanguines et dans des régions isolées du monde comme le Koweït (1/13500) (Farag *et al.* 1988; Farag *et al.* 1989) ou le Newfoundland au Canada (1/18000) (Moore *et al.* 2005).

Les signes cliniques du Bardet-Biedl sont nombreux et variés. Ainsi, en 1999, Beales *et al.* ont proposé une conduite à tenir pour standardiser la définition des signes cliniques du BBS et améliorer ainsi la détection des patients (Beales *et al.* 1999). Ces critères sont décrits dans le Tableau 15. Il faut noter que le diagnostique est établi à partir de 4 critères majeurs ou de 3 critères majeurs associés à des critères mineurs.

Critères majeurs	Critères mineurs
Rétinite pigmentaire	Myopie sévère, strabisme, cataracte, astigmatisme
Polydactylie postaxiale	Syndactylie/brachydactylie
Obésité	Ataxie/troubles de la coordination
	Spasticité modérée des membres inférieurs
Lenteur d'idéation	Retard psychomoteur
Reins hyperéchogènes ou kystiques	Troubles du langage
Hypogénitalisme masculine	Polyurie-polydipsie (diabète insipide néphrogénique)
Hypertension d'origine inconnue	Anomale de l'émail dentaire/de l'éruption dentaire
	Intolérance glucidique
	Hypertrophie du ventricule gauche/cardiopathie
	Fibrose hépathique

Tableau 15 Liste des critères utilisés pour identifier un patient atteint par le syndrome Bardet-Biedl.

13.1.1 Une maladie hétérogène

BBS est une maladie caractérisée par une hétérogénéité génique impliquant 9 gènes (BBS1-9). Fin 2005, ces 9 gènes ne rendaient compte que de 52% des patients connus (Figure 89). Le gène BBS1 était le gène BBS muté le plus fréquemment observé, puisqu'il représentait plus de 20% des patients testés. Parmi l'ensemble des mutations de ce gène, la mutation M390R (Mykytyn *et al.* 2003) est retrouvée, selon les cohortes de patients, dans 75% des patients BBS mutés dans le gène BBS1 (Beales *et al.* 2003). Cette mutation est qualifiée de mutation récurrente. BBS1 constitue ainsi le gène majeur dans le diagnostique de la maladie. Les autres gènes étaient mutés dans des proportions bien moins importantes, jamais supérieures à 6% des patients.

Enfin, on notera que la plus grande proportion, 48% des patients, désigne un groupe de patients pour lequel aucun gène connu n'avait encore pu être identifié. Classiquement des études à partir de ces patients permettent d'isoler de nouveaux gènes. Les question qui se

posent sont mulitples : Comment caractériser les gènes mutés ? Combien de gènes restent-ils ? Quelles seront les proportions ? Etc…

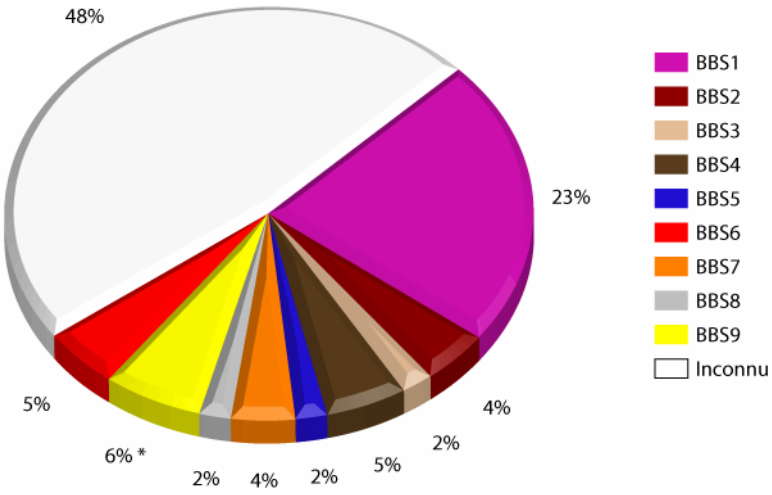


Figure 89 **Présentation de la proportion de chacun des gènes dans la maladie.**
Ces données sont issues d'une cohorte de patients français séquencés au sein de l'équipe d'Hélène Dollfus. Les données pour BBS9 sont basées sur une cohorte américaine (Nishimura *et al.* 2005).

De façon surprenante, bien que les 9 gènes BBS fussent identifiés comme des gènes responsables de la maladie, peu d'informations fonctionnelles étaient disponibles. En effet, hormis leurs localisations chromosomiques, les mutations et quelques assignations de domaines protéiques, il n'existait que peu d'indications sur leurs interactions avec d'autres protéines ou leur rôle précis dans la cellule (Tableau 16).

Gène	HUGO	Localisation génomique	Nombre d'éxons	Définition	Référence
BBS1	BBS1	11q13	17	Bardet-Biedl syndrome 1 protein	(Mykytyn *et al.* 2002)
BBS2	BBS2	16q21	17	Bardet-Biedl syndrome 2 protein	(Nishimura *et al.* 2001)
BBS3	ARL6	3p13	9	ADP-ribosylation factor-like 6	(Chiang *et al.* 2004)
BBS4	BBS4	15q22	16	Bardet-Biedl syndrome 4 protein	(Mykytyn *et al.* 2001)
BBS5	BBS5	2q31	12	Bardet-Biedl syndrome 5 protein	(Li *et al.* 2004)
BBS6	MKKS	20p12	6	McKusick-Kaufman/Bardet-Biedl syndromes putative chaperonin	(Katsanis *et al.* 2000) (Slavotinek *et al.* 2000)
BBS7	BBS7	4q27	9	Bardet-Biedl syndrome 7 protein	(Badano *et al.* 2003)
BBS8	TTC8	14q32	15	Tetratricopeptide repeat protein 8	(Ansley *et al.* 2003)
BBS9	B1	7p14	25	Parathyroid hormone-responsive B1	(Nishimura *et al.* 2005)

Tableau 16 **Tableau récapitulatif des gènes BBS connus fin 2005.**
Les noms de gènes officiels sont référencés par HUGO (Human Genome Organisation).

13.1.2 De l'ombre à la lumière

De manière remarquable en 2003, après plus de 100 ans de recherche, un lien fonctionnel majeur a pu être mis en évidence reliant le syndrome de Bardet-Biedl à une structure cellulaire ancestrale, le cil (Ansley *et al.* 2003; Beales 2005). Le cil est une structure du cytosquelette constituée de microtubules qui joue un rôle essentiel dans la motilité et la perception de stimuli à la fois chimique ou mécanique pour la cellule (voir 1.3.2 Les structures basées sur les microtubules). Le cil joue un rôle essentiel dans la photoréception, dans l'olfaction et dans la perception du mouvement de fluides à la surface des cellules épithéliales du rein (revue dans (Pazour *et al.* 2002)). La perturbation du fonctionnement ou de l'établissement du cil conduit ainsi à la perte d'un élément essentiel dans ces processus sensitifs.

Plusieurs approches ont été nécessaires pour établir cette liaison. Des études fonctionnelles ont permis de mettre en évidence l'importance de certains des gènes BBS dans le maintien ou la genèse de la structure (Blacque *et al.* 2004; Mykytyn *et al.* 2004; Snell *et al.* 2004; Ross *et al.* 2005). D'autres études ont démontré l'implication des gènes BBS dans le transport associé au cil (Kim *et al.* 2004) ou leur localisation cellulaire dans le cil (Kim *et al.* 2005). Parallèlement, une étude de génomique comparative (détaillée dans le chapitre 3.2.1 Soustraction de génomes) a permis de révéler, par soustraction à des protéomes d'organismes ciliés (Homme et Chlamydomonas) du protéome d'un organisme non cilié (*A.thaliana*), la présence du gène BBS5 parmi des groupes de gènes spécifiques aux cellules ciliées (Li *et al.* 2004). Cette étude a ouvert de nouveaux horizons pour l'étude bioinformatique de cette maladie. Ainsi depuis 2003 plusieurs études combinant la génomique comparative et des puces à ADN ont permis d'identifier les gènes BBS3 (Chiang *et al.* 2004) et BBS9 (Nishimura *et al.* 2005).

La superposition quasi parfaite des phénotypes considérés chez les patients atteints de BBS et la localisation des cellules ciliées dans le corps humain pour le rein, les yeux et les oreilles, (Figure 90) ne fait que renforcer ce lien désormais évident. La polydactylie reste encore énigmatique bien que le cil puisse être impliqué dans le développement et donc dans l'établissement des doigts. Concernant une éventuelle affection de l'appareil reproductif, aucune relation claire n'a pu être encore établie.

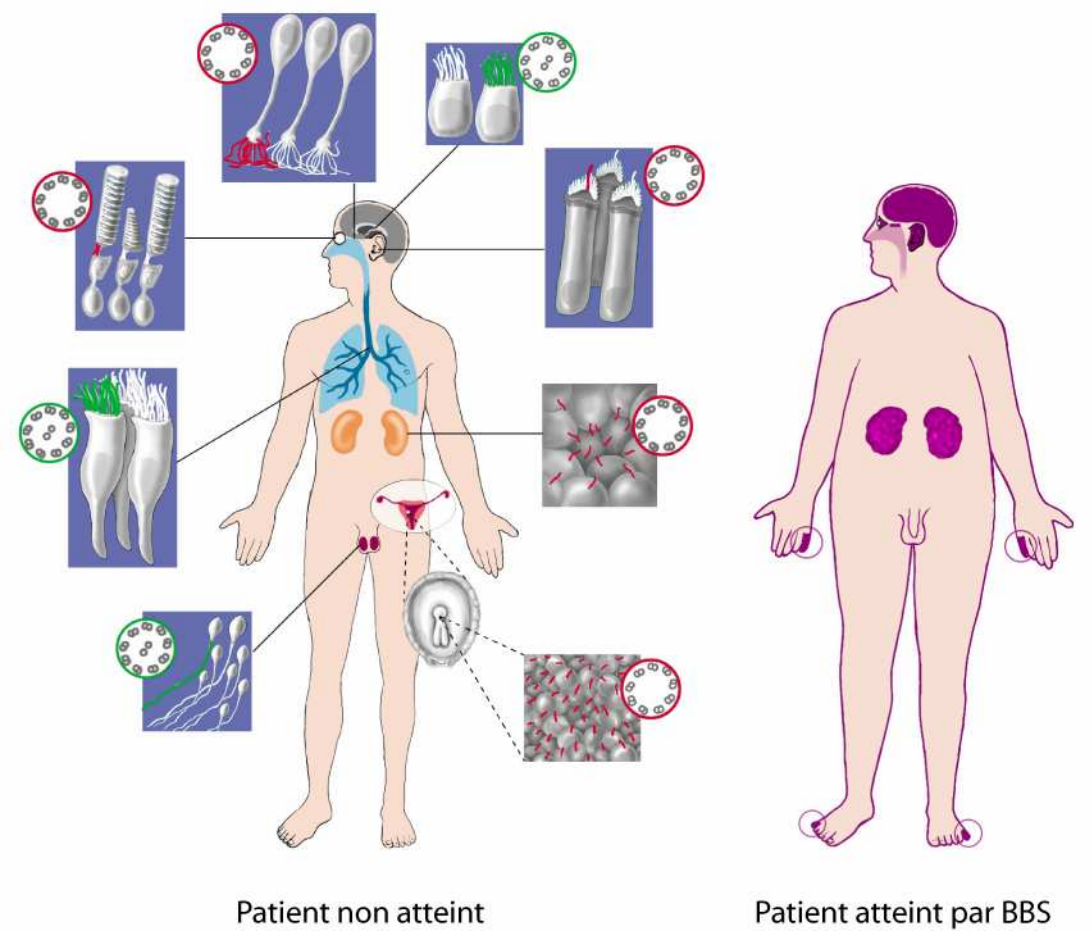


Figure 90 Comparaison de la distribution des cellules ciliées et des phénotypes associés aux patients atteint par BBS.
Les phénotypes sont indiqués en violet foncés.

13.2 Identification de BBS10 et BBS12

Dans le contexte d'une maladie rare où les médecins disposent de peu de patients et avec l'apport de la génomique comparative dans la caractérisation des derniers gènes de BBS, nous avons décidé de rechercher de nouveaux gènes en combinant une approche par cartographie homozygote sur des familles consanguines isolées et une approche bioinformatique de caractérisation des gènes candidats.

Au cours de cette thèse, nous avons ainsi appliqué avec succès cette stratégie pour identifier 2 nouveaux gènes responsables de BBS, BBS10 et BBS12.

13.2.1 Détection de zones chromosomiques candidates

La recherche de nouveaux gènes candidats pour une maladie génétique dont on dispose de peu de familles est basée sur une stratégie éprouvée, l'« Homozygosity Mapping » ou cartographie des régions homozygotes.

Proposée en 1987, cette méthodologie a été décrite comme une stratégie efficiente pour détecter des gènes impliqués dans des maladies rares pour lesquelles le nombre des familles atteintes sont souvent limitées en nombre et en taille (Lander *et al.* 1987). Elle est basée sur le principe que 2 individus atteints issus de la même parenté peuvent hériter d'un même allèle responsable de la maladie. Les zones homozygotes entre les individus atteints ont ainsi une plus forte probabilité de contenir le ou les gènes responsables de la maladie. Naturellement, du fait de la consanguinité, d'autres zones, non liées à la maladie, seront homozygotes chez les patients atteints. Ces zones homozygotes non liées à la maladie sont éliminées par comparaison avec des individus sains de la même famille. Au final, la détection de zones candidates, revient à détecter des zones homozygotes communes aux individus atteints et hétérozygotes chez les individus sains.

Cette technique a déjà permis de mettre en évidence plusieurs gènes responsables de maladie comme l'ataxie de Friedreich avec déficit de vitamine E sur le chromosome 8 (Ben Hamida *et al.* 1993), l'alkaptonuria, une maladie du métabolisme sur le chromosome 3 (Pollak *et al.* 1993). Dans notre cas, ce sont 2 gènes BBS10 (Laurier *et al.* 2006; Stoetzel *et al.* 2006) et BBS12 (manuscrit accepté dans l'*American Journal of Human Genetics*) qui on été identifiés et dont les articles sont disponibles à la fin de ce chapitre.

Pour le gène BBS10, l'analyse d'une famille libanaise hautement consanguine, a permis de mettre en évidence 2 zones chromosomiques. Cette famille originaire d'un petit village au Nord du Liban est composée de 5 fratries atteintes de BBS (Figure 91). Une première approche classique de cartographie par microsatellites (400 marqueurs sur tout le génome) n'avait pas permis d'observer de zones communes. C'est l'utilisation de puces Affymétrix SNP 10K qui a permis d'obtenir la résolution nécessaire à l'obtention des zones homozygotes communes aux différentes familles analysées. Ces puces représentent un gain d'information d'un facteur 8 (10000 SNPs par rapport aux 400 marqueurs microsatellites). Ce sont donc deux zones d'intérêt qui ont été identifiées ; une première zone sur le chromosome 16 correspondant à la région du gène BBS2 (pour une fratrie) et une seconde zone sur le chromosome 12 (pour 3 fratries) ne correspondant à aucun gène BBS connu (Laurier *et al.* 2006).

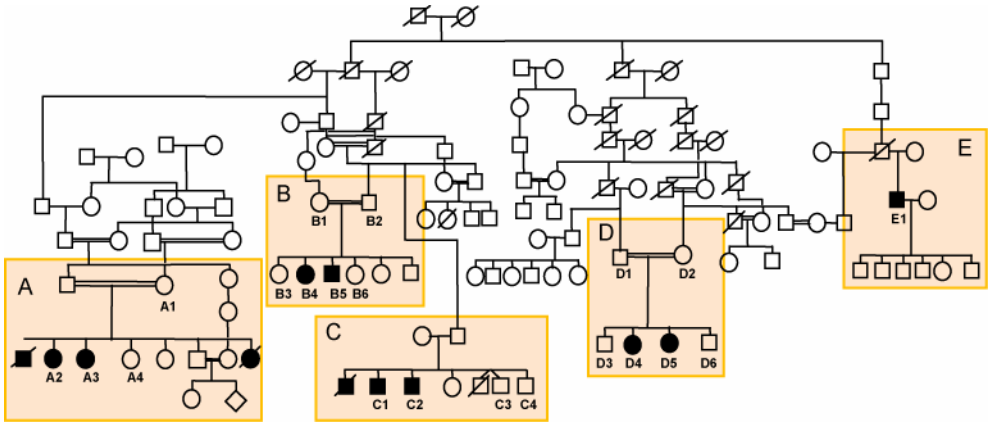


Figure 91 **Famille libanaise étudiée pour la mise en évidence de BBS10.**
Les 5 fratries sont indiquées par les rectangles oranges (A-E). Les individus atteints par le syndrome sont représentés par des rectangles ou cercles noirs. Les individus décédés sont indiqués par des barres noires.

Pour le gène BBS12, plusieurs familles consanguines de gens du voyage (roumaines et libanaises) ont été analysées au moyen des mêmes puces Affymétrix. Parmi toutes ces familles, 3 d'entre elles ont révélé un lien génétique pour une zone homozygote commune sur le chromosome 4. De manière intéressante, cette zone contient le gène BBS7 mais aucun patient ne portait de mutations dans ce gène. On peut noter pour la petite histoire qu'au final, seules 2 familles se sont révélées être mutée dans le gène BBS12, ce qui suggère qu'il existe sans doute d'autres gènes inconnus responsables de la maladie chez la troisième famille.

13.2.2 Analyse bioinformatique

Dans le contexte de BBS10, l'analyse de la cartographie homozygote avait révélé 2 zones. La première isolée pour le groupe de patient D (Figure 91) est localisée sur le chromosome 2 (54,88 à 55,11 Mb). Elle est de petite taille (<1 Mb) et contient le gène BBS2. La seconde isolée pour les patients des groupes A, B et C est localisée sur le chromosome 12 (65,64 à 78,92 Mb). Elle est de grande taille (13,3 Mb) et contient au moins 57 gènes. Des expériences complémentaires ont permis par la suite de réduire cette zone à 8,7 Mb (de 70,22 à 78,92 Mb). Cette nouvelle zone ne contient alors plus que 23 gènes.

Pour BBS12, une seule région a été identifiée sur le chromosome 4 de 118,87 à 124,7 Mb. Cette zone de taille comparable à celle de BBS10 contient 46 gènes dont le gène BBS7. Cette région était incluse dans une zone plus grande de 58 Mb déjà publiée en 2005 par Nishimura *et al.* (Nishimura *et al.* 2005). Les auteurs avaient alors suggérer l'existence d'un locus BBS dans cette grande zone.

Dans le but d'établir une liste des priorités dans l'étude des cibles potentielles, nous avons en premier lieu extrait les séquences des 69 (23 et 46) gènes présents dans les zones incréminées, puis appliqué les différents protocoles bioinformatiques disponibles comme RetScope et ComIcs à ces 2 ensembles de gènes. Ceci nous a permis de caractériser les homologues de ces gènes dans les banques de protéines, de construire les alignements multiples, de calculer les profils phylogénétiques, de définir l'organisation en domaines des protéines, d'étudier leurs fonctions potentielles, de déterminer les termes Gene Ontology associés, etc.

Tenant compte des récentes découvertes caractérisant BBS comme une ciliopathie, nous avons choisi de focaliser nos recherches sur 2 approches principales. La première approche est ciblée sur les gènes associés aux cils. Pour cela, nous avons comparé les gènes de nos intervalles avec les listes de gènes obtenues par la génomique soustractive et enrichies en protéines du cil (Chiang *et al.* 2004; Li *et al.* 2004). Dans la seconde approche, nous avons recherché les gènes annotés comme faisant parti du cytosquelette et plus particulièrement des microtubules. Dans ce but, nous avons analysé les termes Gene Ontology et les annotations des gènes considérés. Dans ce cas, les profils phylogénétiques nous ont permis d'ordonner les protéines et d'analyser en priorité les gènes avec des profils proches de gènes communs aux listes de protéines du cil (Figure 92 et Figure 93).

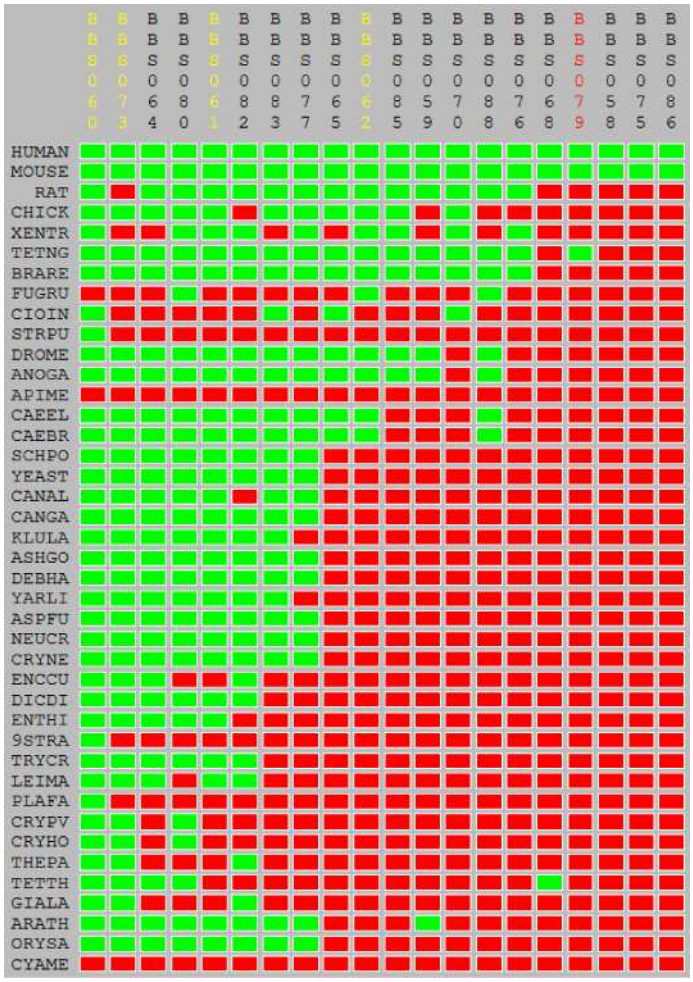


Figure 92 **Profils phylogénétiques des gènes de la zone du chromosome 12.**
Les noms indiqués en jaune correspondent aux séquences communes aux listes enrichies en protéines du cil (Chiang *et al.* 2004; Li *et al.* 2004). BBS10 est indiqué en rouge.

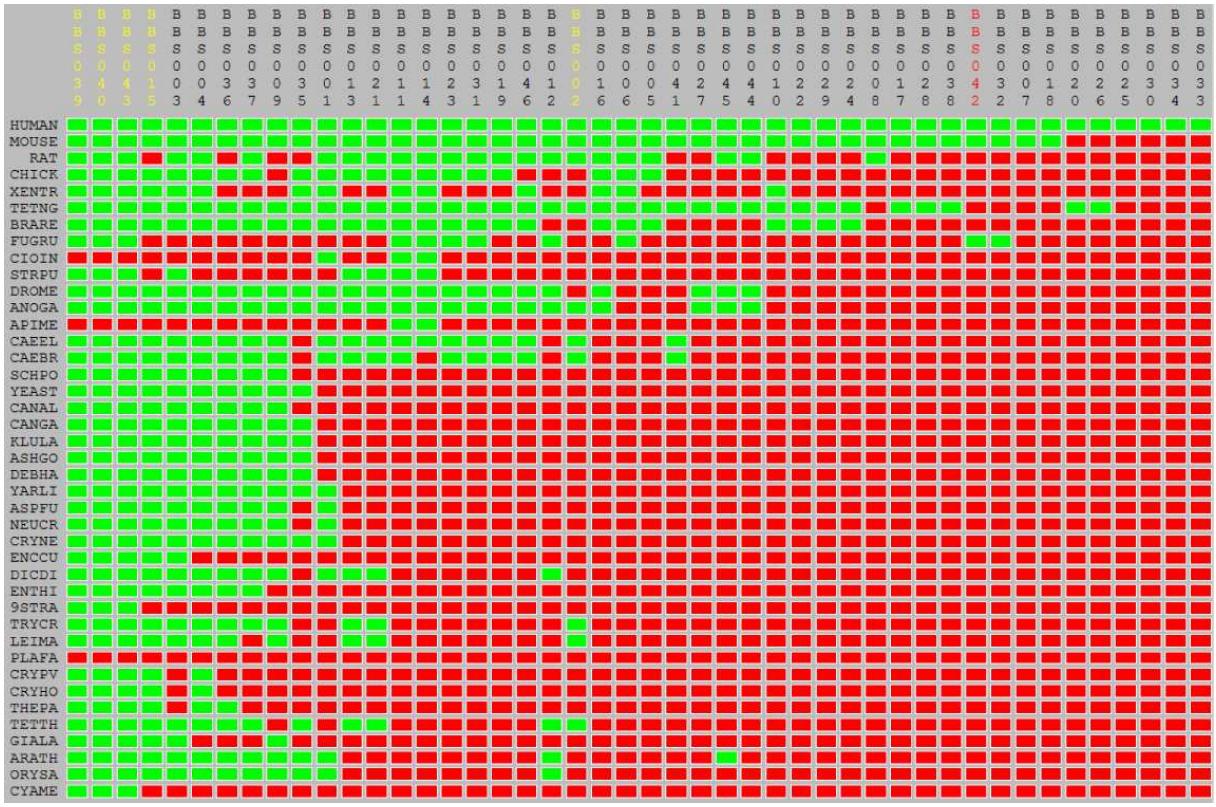


Figure 93 **Profils phylogénétiques des gènes de la zone du chromosome 4.**
Les noms indiqués en jaune correspondent aux séquences communes aux listes enrichies en protéines du cil (Chiang *et al.* 2004; Li *et al.* 2004). BBS12 est indiqué en rouge.

La première approche nous a permis d'isoler 4 et 5 gènes respectivement pour les zones sur le chromosome 12 (locus BBS10) et sur le chromosome 4 (locus BBS12) (Tableau 17). Cette approche n'a pu révéler aucun des 2 gènes, puisque le séquençage des gènes retenus chez les patients atteints, n'a identifié aucune mutation. La seconde approche a permis d'isoler des gènes supplémentaires. Le séquençage de ces gènes a permis d'identifier les gènes BBS10 (FLJ23560) et BBS12 (FLJ35630) comme étant mutés de manière significative (divers types de mutations, comme des mutations non sens et des pertes du cadre de lecture) et responsables de phénotypes observés.

Chromosome 12 (BBS10)		Chromosome 4 (BBS12)	
Gène	Position (Mb)	Gène	Position (Mb)
RAB21 (Ras-related protein Rab-21)	70,4	BBS7	122,9
TPH2 (tryptophan 5-monooxygenase 2)	70,3	PRSS12 (Neurotrypsin)	119,4
TBC1D15 (TBC1 domain family member 15)	70,5	SEC24D (Protein transport protein)	119,9
HRB2 (HIV-1 Rev binding protein 2)	74,2	SPATA5	124,06
		SPAF	124,07
Loc387869 (similar to microtubule)	73,3	NP_078850 (fibronectin III like domain)	122,2
NP_078961 (Hypothetical protein FLJ23560)	75,2	NP_689612 (Hypothetical protein FLJ30834)	122,8
ZDHHC17 (Huntingtin-interacting protein 14)	75,7	NP_689831 (FLJ35630 hypothetical protein LOC166379)	123,9
SYT1 (synaptotagmin-1)	77,8		

Tableau 17 Liste des séquences identifiées pour la mise en évidence des gènes BBS10 et BBS12.
La partie haute du tableau contient la liste des gènes identifiés par la première approche. La partie basse contient les gènes identifiés par la seconde approche. Les gènes BBS10 et BBS12 sont indiqués en rouge respectivement à gauche et à droite.

13.3 Analyse manuelle de l'alignement multiple

Les 2 gènes BBS10 et BBS12 sont des gènes non annotés du génome humain. Ces cadres ouverts de lecture ont été prédits par les programmes de prédiction de gènes lors de l'annotation structurale du génome humain. Néanmoins, aucune information supplémentaire quant à une éventuelle fonction ou une interaction avec d'autres protéines n'est disponible. Hormis les séquences nucléiques et protéiques, il n'y avait initialement aucune donnée disponible. Dans ce contexte, nous avons ainsi entrepris une analyse de séquence approfondie par l'utilisation des recherches de similarités dans les banques de séquences et la caractérisation des familles de protéines au moyen du MACS (2.4 L'alignement multiple), dans le but d'extraire et d'apporter un maximum d'information sur ces 2 gènes. Ces études ont révélé que ces gènes encodent pour des membres distants d'une même famille de protéines, les chaperonines de type II. Il faut remarquer que le gène BBS6 avait également été identifié comme un membre de cette famille et nous l'avons ainsi inclus dans nos analyses (Stone *et al.* 2000; Kim *et al.* 2005).

13.3.1 Les chaperonines

Les chaperonines sont des complexes protéiques capables d'aider d'autres protéines à obtenir leur conformation spatiale définitive. Il existe plusieurs types de chaperonines. Parmi les chaperonines cylindriques qui possèdent une cavité centrale, on distingue 2 types : les chaperonines de type I ou GroEL/ES et les chaperonines de type II ou CCT (TRiC). Le type I est retrouvé chez les bactéries et chez les eucaryotes uniquement dans les organelles d'origine endosymbiotique (mitochondrie et chloroplaste) alors que le type II est présent des archées à l'homme. La structure globale du complexe est identique, cependant le type I est composé de 3 parties (2 anneaux fermés par une coiffe) alors que le type II est composé de seulement 2 parties (2 anneaux fermés) (Figure 94). L'organisation des sous unités en 3 domaines est la même (apical, intermédiaire et équatorial), la différence se faisant au niveau d'un protrusion spécifique au type II qui permet la fermeture de la cavité (Figure 94). Les substrats des chaperonines de type II sont essentiellement des protéines du cytosquelette (revue dans (Spiess *et al.* 2004)), en particulier l'actine et la tubuline (Gao *et al.* 1992; Sternlicht *et al.* 1993).

Chez les eucaryotes, un anneau contient 8 sous unités différentes, nommés de alpha à zéta, d'une taille d'environ 550 acides aminés chacune (Kubota *et al.* 1994). BBS6, BBS10 et BBS12 ont respectivement des tailles de 570, 723 et 710 acides aminés.

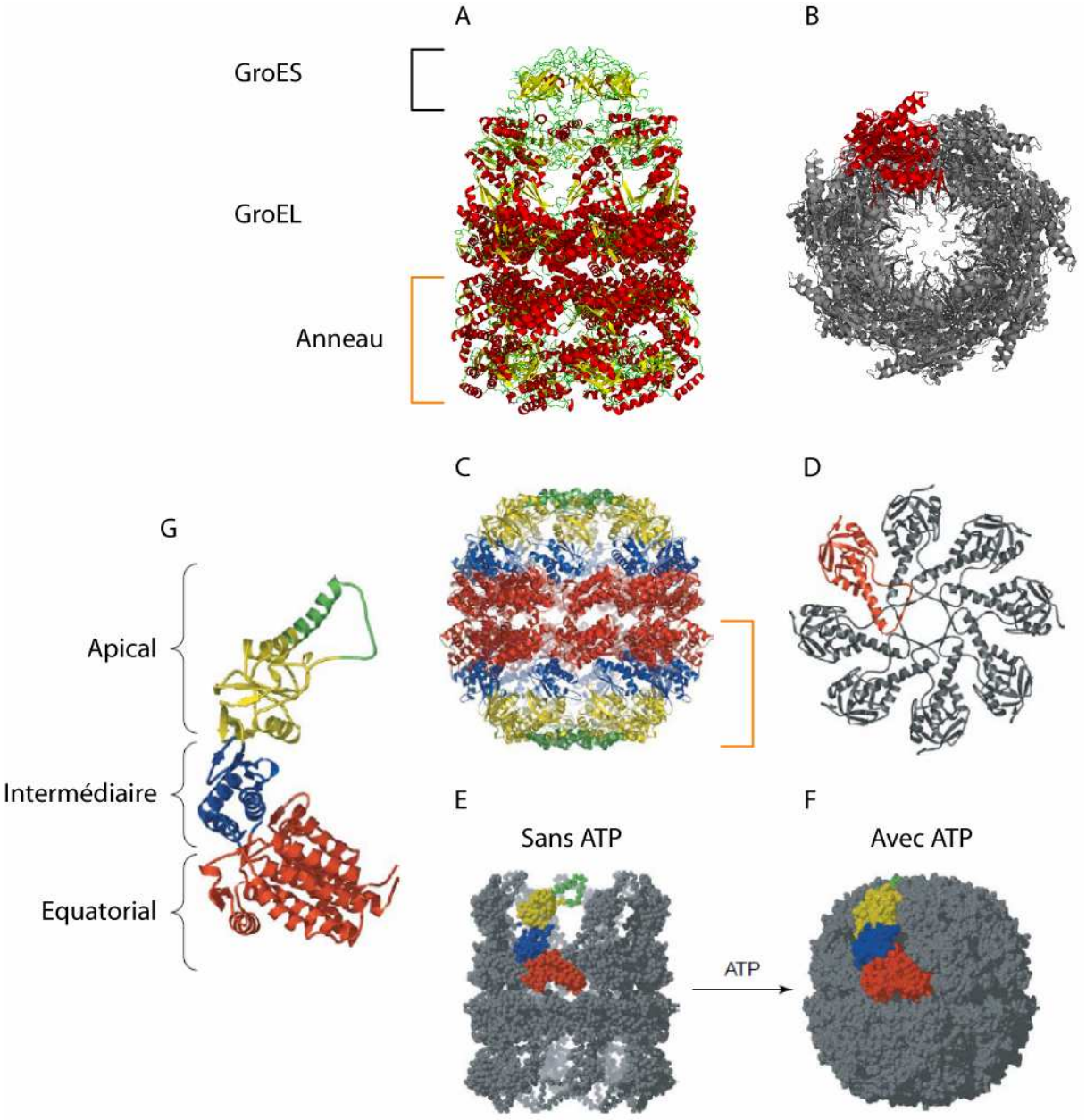


Figure 94 **Structures des chaperonines de type I et de type II.**
(A) et (B) montrent le complexe de type I (GroEL/ES) (PDB 1AON) en vue latérale et apicale respectivement. Il est composé de 3 domaines, 2 anneaux (GroEL) et une coiffe (GroES). De la même manière le complexe de type II (CCT ou TRiC) (adapté de (Spiess *et al.* 2004)) est illustré par (C) et (D). Une sous unité du complexe est illustrée en rouge dans chacune des vues apicales. Pour le type II, les 2 états ouverts et clos correspondent à l'absence (E) ou non (F) de la molécule d'ATP. Le détail de l'organisation d'une sous unité en 3 domaines (apical, intermédiaire et équatorial) est illustré en (G). La différence majeure entre les 2 types complexes est représentée par la petite extension (partie verte de la sous unité) (G) qui permet aux complexes de type II de se fermer.

13.3.2 Organisation en domaines

L'alignement multiple complet (http://bips.u-strasbg.fr/BBS) de tous les types de sous-unités de chaperonines de type II incluant les 3 gènes BBS (BBS6, BBS10 et BBS12), nous a permis

d'identifier les 3 domaines connus de ces protéines à savoir le domaine apical, intermédiaire et équatorial (Figure 94). Il faut noter que la similarité entre ces protéines et les chaperonines de type II canonique est faible. Il semble ainsi que seule l'organisation en domaines et l'organisation structurale (« fold ») de ces protéines soient conservées. Leur architecture est qualifée de discontinue, puisque les domaines structuraux sont composés de parties non linéaires dans la séquence primaire.

L'alignement multiple a permis de corriger les prédictions faites sur la protéine BBS10 humaine dont la partie N-terminale était absente dans les banques de séquences protéiques. En effet, environ 70 acides aminés codés par le premier exon avaient été mal prédits par le choix d'un codon initiateur interne.

Cette analyse a également mis en évidence des séquences additionnelles dans les 3 protéines BBS par rapport à la séquence canonique d'une chaperonine de type II. Ces insertions, plus ou moins grandes (de 8 à 170 acides aminés) sont disposées tout le long de la séquence. Néanmoins, comme nous avons pu le voir sur le modèle réalisé (cf Figure 1b de (Stoetzel *et al.* 2006) et Figure 2d de l'article BBS12), leur positionnement spatial est quasi exclusivement réservé à la même face de la molécule. Ceci suggère fortement que ces insertions puissent participer à la définition de nouveaux domaines structuraux et/ou de nouvelles fonctions. Une tentative de modélisation par homologie par le programme SWISS-MODEL (Arnold *et al.* 2006) en utilisant la séquence de BBS10 et une séquence canonique d'une chaperonine de type II permet de mieux visualiser ces interactions (Figure 95). La grande insertion de BBS10 est apparemment constituée majoritairement d'hélices alpha.

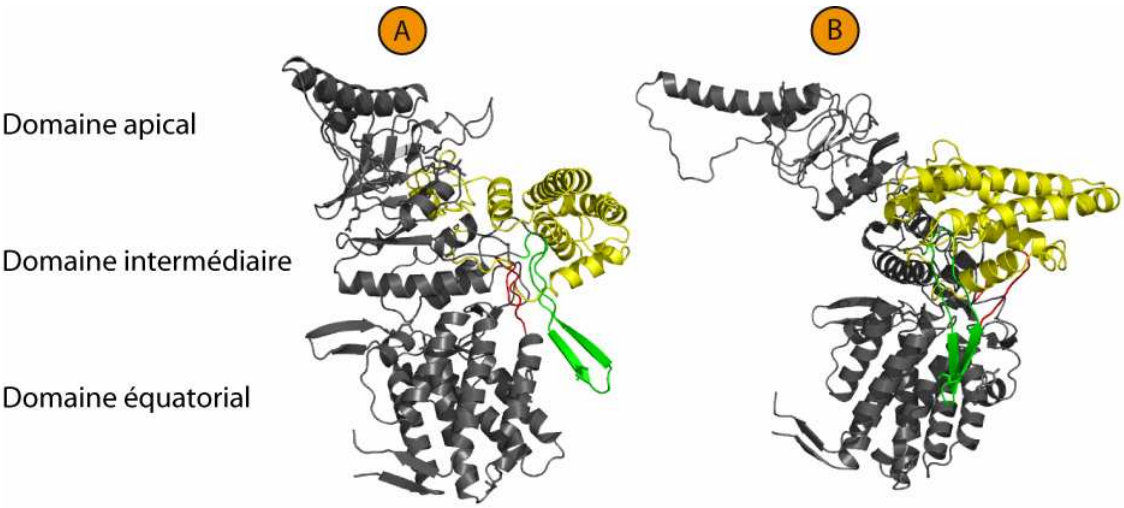


Figure 95 **Modèle de la structure de BBS10.**
Le modèle de la structure de BBS10 est montré en vue frontale (A) et en vue latérale (B). Les 3 domaines typiques sont indiqués. La partie commune aux sous-unités de type II est représentée en gris foncé. Les insertions 1, 2, et 3 sont représentées en rouge, vert et en jaune, respectivement. Le modèle a été obtenu en utilisant le serveur de prédiction SWISS-MODEL.

Il est intéressant de noter que, par rapport à la structure canonique, certains points d'insertion sont communs entre BBS6 et BBS12 et entre BBS10 et BBS12. Outre la plasticité de ces régions au sein du fold, ceci pourrait suggérer que ces régions sont les zones privilégiées d'ajout de séquences au sein des chaperonines liées au BBS.

Un des modes de fonctionnement possible de ces gènes est leur association au sein du complexe canonique des chaperonines de type II. Elle pourrait en effet remplacer une ou plusieurs sous-unités du complexe et participer à la modulation de son activité. Cependant, la présence d'insertions de taille relativement conséquente sur une seule façade de la molécule pourrait constituer une gêne stérique et empêcher ce scénario. Ceci a été suggéré pour BBS6, qui possède les insertions les moins conséquentes et qui ne s'associe pas avec le complexe canonique (Kim *et al.* 2005). Toutefois à ce stade des connaissances, plusieurs autres modes de fonctionnement sont envisageables, comme des homodimères ou hétérodimères ou trimères… Une première réponse à cette question sera donnée par la caractérisation de la structure 3D complète de ces protéines qui a été entamée à l'IGBMC et qui permettra de juger de façon plus exacte de la structure et du rôle de ces insertions. Indépendamment de ces conclusions, au niveau des séquences, ces insertions constituent des éléments caractéristiques, distinctifs des chaperonines de type II canonique.

13.3.3 Conservation du site ATP

L'analyse de la conservation moyenne des résidus impliqués dans la fixation et l'utilisation de l'ATP (Ditzel *et al.* 1998) a révélé une possible utilisation de l'ATP pour les fonctions de BBS6, BBS10 et BBS12 (Figure 96). Cependant, il faut noter que BBS10 est la seule protéine à conserver quasi parfaitement un motif majeur de fixation de l'ATP, GDGTT[T/S]. Ces résultats sont des prédictions et devront être validés par des tests biologiques.

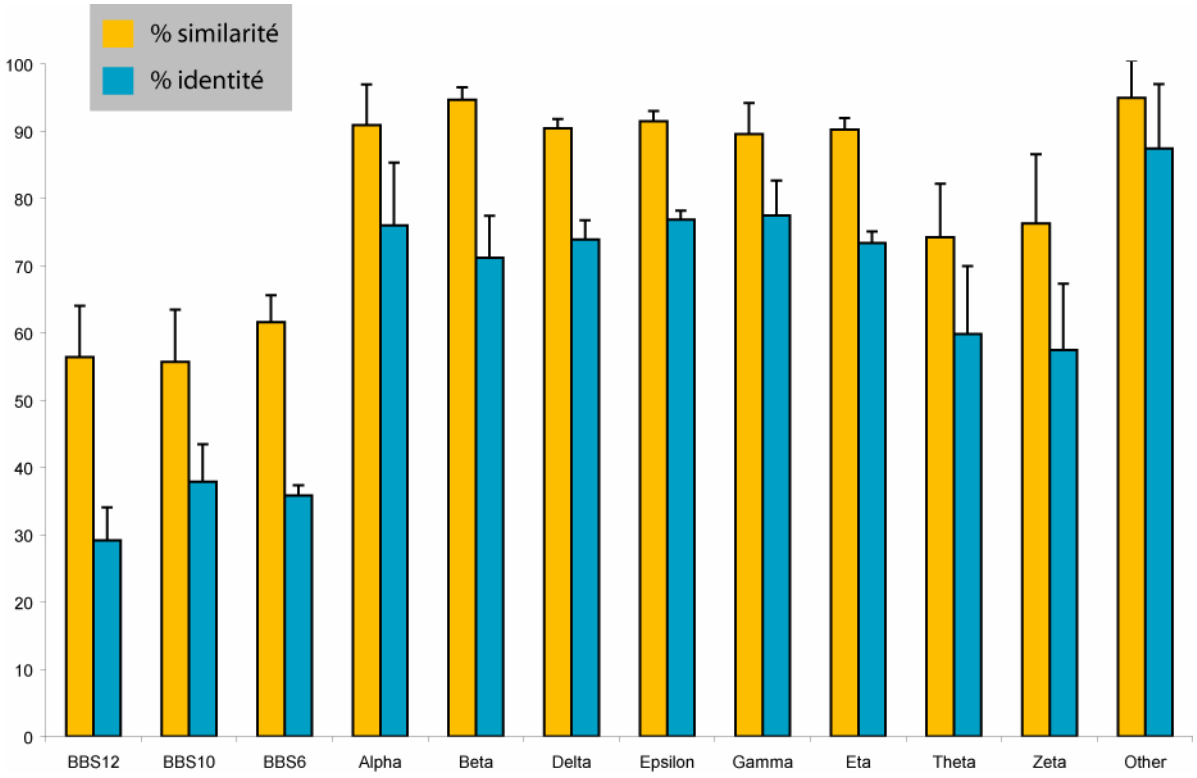


Figure 96 Histogramme de la conservation des résidus impliqués dans la fixation et l'utilisation de l'ATP.
Les moyennes des conservations (identité et similarité) des résidus sont calculées pour l'ensemble des protéines de chaque sous famille de l'alignement multiple. Les résidus sont identifiés à partir des travaux de Ditzel ***et al.*** sur les sous unités de chaperonines de type II (Ditzel *et al.* 1998).

13.3.4 Profil phylogénétique

L'analyse du profil phylogénétique de BBS10 et BBS12 a permis d'illustrer plusieurs limites ou dangers de la génomique comparative. Il faut tout d'abord noter que du fait de la pauvre caractérisation de ces protéines, peu d'homologues sont repertoriés dans les banques de protéines. La conséquence directe de ceci est la définition de profils phylogénétiques erronés (BBS10 ne serait présent que chez l'homme, la souris et le poisson *T. nigroviridis* et BBS12 que chez l'homme, la souris et le poison *T. rubripes*) (Figure 92 et Figure 93). Une analyse plus fine, basée sur l'utilisation des séquences génomiques des organismes dont le génome est

complet, nous a permis de caractériser davantage d'homologues aux gènes BBS10 et BBS12 humain et de déterminer ainsi leur véritable distribution phylogénétiques. Leur présence a ainsi été validée pour les seuls organismes vertébrés (des poissons à l'homme en incluant le poulet, le rat, la grenouille, le taureau et le chien).

Cet exemple souligne encore une fois l'importance de la validation des profils phylogénétiques par l'exploration des génomes. Une autre conséquence de cette distribution tout à fait particulière et inattendue est la remise en question des approches uniquement basée sur la génomique soustractive. En effet, il apparaît désormais évident que les études menées par Li *et al.* et Chiang *et al.* (Chiang *et al.* 2004; Li *et al.* 2004) basées sur l'hypothèse d'une présence au sein de tous les organismes ciliés (incluant Chamydomonas et l'homme) ne pouvait pas isoler BBS10 et BBS12. La distribution restreinte aux vertébrés pour BBS10 et BBS12 et aux chordés pour BBS6 (un homologue très éloigné pourrait exister chez *Ciona intestinalis*) illustre une apparition récente de cette famille de gènes. On peut proposer que ces protéines constituent des marqueurs de l'évolution ou plutôt de la spécialisation du cil chez les vertébrés.

13.3.5 Evolution

L'analyse de l'arbre phylogénétique et notamment la comparaison des longeurs de branches au sein des organismes vertébrés montre que par rapport aux autres chaperonines, les protéines BBS6, BBS10 et BBS12 ont une distance évolutive beaucoup plus importante. Considérant que cette distance correspond à un même temps d'évolution (la divergence des vertébrés), ceci se traduit par une divergence de séquence plus grande. Il est intéressant également de noter que les 3 gènes BBS ont le même comportement et suggèrent donc une co-évolution. L'ensemble de ces évidences suggère que BBS6, BBS10 et BBS12 possèdent une vitesse d'évolution plus grande ou une pression de sélection moins importante qui pourrait correspondre à leur récente implication dans l'établissement de fonctions ciliaires récentes spécifiques des vertébrés.

Un autre point intéressant de comparaison avec les chaperonines classiques est le nombre et la conservation des exons. Il est intéressant de noter que le nombre d'exons codant pour les chaperonines BBS est faible ; BBS6 possède 4 exons, BBS10 possède 2 exons et BBS12 possède 1 exon. Ensuite, en comparaison avec les autres gènes des sous-unités de chaperonines de type II qui comptent toutes entre 12 et 14 exons, ceci suggère encore une fois une histoire évolutive distincte. Ce découpage exonique est de surcroit conservé dans tous les génomes étudiés.

Enfin pour mieux comprendre la relation entre BBS et les chaperonines, il faut rappeler que de nombreux substrats des chaperonines de type II sont des protéines du cytosquelette. La caractérisation de nouvelles protéines de cette famille n'est pas tellement surprenante, sachant que le phénomène de duplication a guidé l'évolution des chaperonines (Archibald *et al.* 2000). Plusieurs hypothèses suggérent une évolution spécifique et parallèle de ces chaperonines avec les cytosquelettes (Archibald *et al.* 2000; Valpuesta *et al.* 2002). Ces 2 protéines constituent peut-être une nouvelle étape de cette évolution chez les vertébrés mélant à la fois les chaperonines et le cytosquelette, puisque BBS10 et BBS12 sont très certainement impliqués dans le fonctionnement du cil (une structure des microtubules). La caractérisation de leurs fonctions et de leurs partenaires seront autant de clés pour mieux comprendre cette implication dans la fonction du cil et leur origine en tant que chaperonine.

13.4 Conclusion et perspectives

La découverte de ces 2 nouveaux gènes est une avancée majeure pour l'étude du syndrome de Bardet-Biedl, qui a des conséquences multiples.

Premièrement, le diagnostique de la maladie est amélioré et simplifié. En effet, le gène BBS10 couvre près de 20% des patients et ne possède que 2 exons à tester. BBS12 est identifié pour 5% de patients et ne contient qu'un seul exon codant (Figure 97).

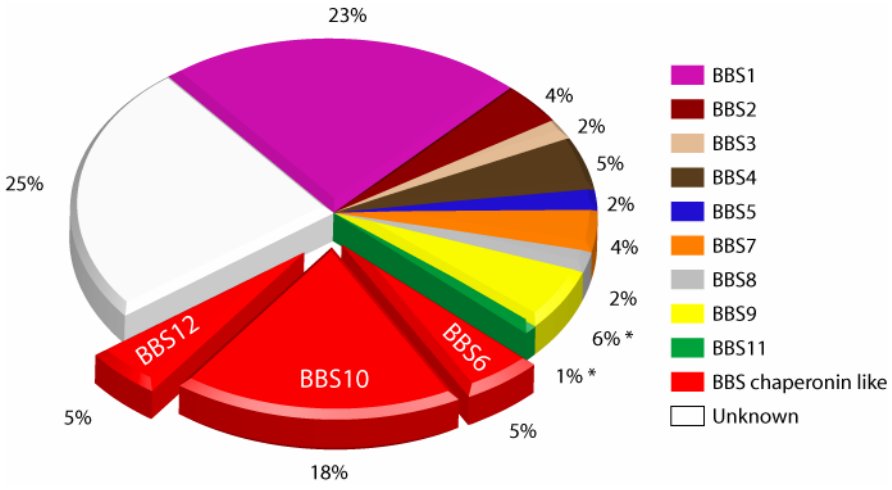


Figure 97 **Proportion de chacun des gènes dans la maladie en 2006.**
Ces données sont issues d'une cohorte de patients français séquencés au sein de l'équipe d'Hélène Dollfus. Les données pour BBS9 sont basées sur une cohorte américaine (Nishimura *et al.* 2005) et les données pour BBS11 sont basées sur les données publiées par Chiang ***et al.*** (Chiang *et al.* 2006).

Avec BBS6, les gènes BBS10 et BBS12 définissent une nouvelle branche de la famille des chaperonines, que l'on nommera « chaperonin-like ». Elles représentent ainsi 1/4 des gènes responsables du BBS et sont détectées dans près de 30% des patients. En 2006, un nouveau

gène BBS, BBS11 (TRIM32) a été identifié (Chiang *et al.* 2006). Cependant, avec une seule mutation faux sens caractérisée pour ce gène et l'absence de mutation dans plusieurs cohortes de patients américains et français (Dollfus et Katsanis communications personnelles), son statut de gène responsable de BBS semble compromis.

Ces 2 gènes sont les gènes BBS les mieux caractérisés à ce jour. La famille de protéines définie avec BBS6 ouvre de nouvelles perspectives fonctionnelles au sein de la maladie. Les prédictions que nous avons effectuées ouvrent le champ à bon nombre d'expériences complémentaires qui permettront de mieux comprendre les implications de ces protéines dans la maladie.

Ainsi, bien que la fonction de ces chaperonines-like n'ait pas encore été clairement élucidée, le lien fonctionnel avec le cil est bel et bien maintenu. Ainsi, tout comme BBS6 (Kim *et al.* 2005), BBS10 est localisé dans le cil (Katsanis, communication personnelle) et plus particulièrement dans la zone péricentriolaire (proche du centrosome). Il a été récemment montré que le cil des vertébrés possède certaines spécialisations l'impliquant dans le développement embryonnaire aux moyens des voies de communication cellulaires (revue dans (Davis *et al.* 2006)). L'existence de BBS10 et BBS12 spécifiques aux vertébrés est peut-être le reflet de cette spécialisation.

Deuxièmement, leur analyse met en avant quelques précisions quant à l'utilisation de la génomique comparative dans le syndrome de Bardet-Biedl. D'une part, tous les gènes ne sont pas forcément spécifiques de tous les organismes ciliés. D'autre part, la complémentarité des approches doit être privilégiée afin de caractériser de nouveaux gènes. Enfin, l'analyse des profils phylogénétiques de BBS10 et BBS12 soulève l'importance de la validation au niveau génomique de ces profils.

Récemment, plusieurs banques de données enrichies en protéines du cil ont été construites (Gherman *et al.* 2006; Inglis *et al.* 2006). Ces nouvelles ressources permettront d'identifier de nouveaux candidats pour les maladies du cil comme BBS et de mieux comprendre les éventuelles interactions entres les protéines du cil.

Ce travail a permis de révéler leur existence et de les caractériser au mieux, cependant il reste encore de nombreuses questions sans réponses ; Quel est leur fonction ? Définissent-elles un complexe (dimère, trimère ou plus) ? Redondance de fonctions ? Utilisation de l'ATP ? Conformation spatiale ?

Publication N°2

La transcriptomique

Chapitre 14 - Actichip une puce dédiée au cytosquelette

« Quand on prend la peine de découvrir les ficelles, on se sent moins marionnette... »

Robert Blondin

La création de la puce Actichip s'inscrit dans la thématique générale du laboratoire à Luxembourg qui concentre ses efforts sur l'étude du cytosquelette et en particulier, du cytosquelette d'actine. Le cytosquelette est un système dynamique dont les éléments sont contrôlés dans le temps et dans l'espace. La création d'une puce thématique dédiée au cytosquelette permettra de disposer d'un outil capable de détecter les changements de l'expression de l'ensemble des gènes de ce système complexe lors de certaines situations physiologiques ou pathologiques.

Ma contribution dans ce projet a consisté à développer les outils bioinformatiques attachés à la création d'Actichip, notamment au choix de sondes spécifiques des gènes du cytosquelette. Ce projet a combiné des étapes de bio-analyse, afin de mieux comprendre les problèmes posés et a donné lieu au développement d'un logiciel de design de sondes (CADO4MI), mais également à d'autres outils ou protocoles comme DbFaster ou GalActicA. L'ensemble des logiciels est décrit dans la partie Matériel et Méthodes (voir le Chapitre 9 - CADO4MI et le Chapitre 10 - Autres outils développés). Deux manuscrits sont en cours de correction en vue d'une soumission avant la fin de l'année 2006.

Dans cette partie, après un bref rappel sur l'historique de la puce, je décrirai les différents aspects du design de sondes qui ont mené à la création d'Actichip et les problèmes rencontrés. Enfin, j'évoquerai les différentes validations et les premiers résultats liés à ce travail.

14.1 Les différentes versions d'Actichip

Le développement de la puce Actichip se déroule en plusieurs phases. On distingue ainsi 3 phases à l'issue desquelles l'intégralité des séquences référencées dans notre banque Actinome sera représentée sur la puce (Figure 98).

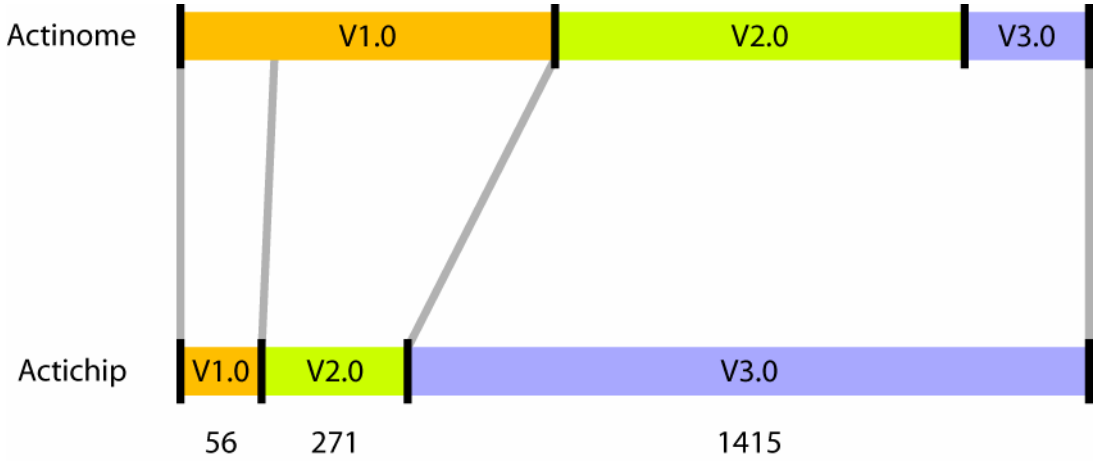


Figure 98 **Les différentes versions de la puce Actichip.**
Les versions de la puce Actichip couvrent les différentes versions de la banque Actinome.

Dans un premier temps, nous avons établi un prototype de la puce Actichip (v1.0). Cette version contient 56 gènes issus de la première version d'Actinome. Ce prototype a permis de vérifier la faisabilité du projet et de maitriser les différents outils de fabrication de la puce. Ce travail a été mené par Laurent Vallar et ne sera pas discuté au cours de cette thèse.

La seconde version de la puce à ADN, Actichip v2.0 comporte essentiellement des gènes du cytosquelette d'actine et des marqueurs de tissus épithéliaux et mésenchymateux. Actichip 2.0 correspond à la version en cours d'utilisation et fait l'objet d'un manuscrit en cours de soumission.

Enfin la dernière version, Actichip v3.0, contiendra l'ensemble des gènes représentés dans Actinome. Cette dernière version qui couvrira l'ensemble des protéines du cytosquelette est en cours de développement.

Dans ce chapitre, nous aborderons le choix des sondes de manière plus général, puis le design de la deuxième version d'Actichip à laquelle nous inclurons les sondes de la première version. La version 3 sera mentionée sur quelques aspects en particuliers.

14.2 Le design de sondes spécifiques

Le design de sondes pour puces à ADN est un processus multicritère qui consiste à choisir pour chaque ARNm la, ou les, parties de sa séquence qui servira de canevas à la séquence complémentaire. Un oligonucléotide d'une taille fixe et comportant cette séquence sera synthétisé et déposé sur la puce. Cet oligonucléotide s'hybridera spécifiquement à la séquence coplémentaire d'un ADN cible marqué par un fluorochrome. Les divers critères nécessaires au choix d'une sonde sont décrits dans la partie Matériel et Méthodes (Chapitre 9 - CADO4MI). Parmi ces critères, le critère de spécificité est très certainement le plus

important (Kane *et al.* 2000; Hughes *et al.* 2001). En particulier, 2 paramètres, le pourcentage d'identité et le nombre de bases consécutives identiques permettent de déterminer si une séquence similaire à la sonde sera capable de s'hybrider à elle lors de l'étape d'hybridation sur la puce à ADN.

Un certain nombre de programmes de design existaient au début de ce projet et d'autres sont apparus au cours de sa réalisation (Tableau 18). Ils diposent tous d'un certain nombre de caractéristiques communes comme l'évaluation de la spécificité ou le calcul de la température de fusion (Tm). Le pourcentage de GC (%GC), les structures secondaires et la position de la sonde sur la séquence sont des critères qui ne sont pas considérés par tous les programmes.

Le design de sondes dans le cadre des gènes du cytosquelette a été un véritable challenge. En effet, le cytosquelette contient un nombre important de familles de gènes partageant une forte identité de séquences et comportant de nombreuses isoformes ou de variants d'épissage. La superfamille des actines en est un exemple éloquent. Les isoformes d'actine partagent ainsi jusqu'à 99% d'identité (Tableau 13) sur la partie codante. Ceci réduit considérablement le nombre de zones potentiellement disponibles pour choisir une sonde spécifique. Les cytokératines, les myosines en sont d'autres exemples.

Il apparaît ainsi nécessaire de disposer d'un outil capable de réaliser en automatique cette tâche tout en assurant une validation importante des résultats. De plus, un accès à l'ensemble des paramètres et des informations du design est un des éléments essentiels à la maitrise du choix des sondes. L'interface graphique d'analyse des résultats permet de visualiser, valider et contrôler de manière précise le choix des sondes et ainsi de faciliter le « re-design » des cas problématiques. Ce second outil faisait défaut sur la majorité des logiciels au moment de la réalisation de notre projet.

Nous avons ainsi développé CADO4MI (*Computer Assisted Design of Oligonucleotide For Microarray*) pour réaliser le design des oligonucléotides de la puce Actichip.

Nom du logiciel	ProbeSelect	OligoArray 1.0	OligoArray 2.0	OligoWiz	Picky	ROSO	YODA	GoArrays	CADO4MI
Langage de programmation	C++	Java	Java	Java	C++	C	Java	Java	Tcl/Tk
Système	U	W/M/U	W/M/U	W/M/U	W/M/U	W/M/U	W/M/U	W/M/U	W/M/U
LC/IG/IGA	LC/IG	LC/IG	LC/IG	IG/IGA	IG	LC/IG	LC/IG	IG	LC/IGA
Similarité	myersgrep	BLAST	BLAST	BLAST	Suffix array	BLAST	SeqMatch	BLAST	BLAST
Méthode de calcul de Tm	NN	NN	NN	NN	NN	NN	NN	NN	NN/MW/MOL
Intervalle de Tm	Oui	Oui	Calculé	Calculé	Calculé	Oui	Oui	Oui	Oui
%GC	Oui	Non	Oui	Non	Oui	Oui	Oui	Non	Oui
Structure secondaire	Non	Oui	Oui	Non	Oui	Oui	Oui	Oui	Non
Séquences prohibées	Non	Oui	Oui	Non	Non	Oui	Oui	Oui	Oui
Distance du 3’	Non	Oui	Oui	Calculé	Non	Oui	Non	Oui	Oui
Taille de la sonde	Oui	Variable	Variable	Calculé	Variable	Oui	Variable	Oui	Oui
Pas de la fenêtre d’analyse	Non	10 nts	1 nt	Non	Non	Non	Non	Non	Variable
Particularités	Calcul de l'énergie libre		Modèle thermodynamique	Fléxibilité	Puissance de calcul	Séparations des étapes	Calcul de la similarité	sonde discontinue	Estimation du Tm Utilisation de 2 banques
Références	(Li *et al.* 2001)	(Rouillard *et al.* 2002)	(Rouillard *et al.* 2003)	(Nielsen *et al.* 2003)	(Chou *et al.* 2004)	(Reymond *et al.* 2004)	(Nordberg 2005)	(Rimour *et al.* 2005)	Manuscrit en préparation

Tableau 18 Comparaison de différents logiciels de design de sondes pour puces à ADN.
W=Windows, M=Macintosh, U=Unix, LC=ligne de commande, IG=interface graphique, IGA=IG avancée pour l’analyse des résultats, NN=Nearest Neighbor, MW=Méthode de Wallace, MOL=Méthode des oligonucléotides longs

14.2.1 Quelques aspects de CADO4MI

Dans le développement de notre outil de design de sondes, nous avons voulu associer une flexibilité nécessaire et la validation des résultats. Ceci dans le but d'obtenir des données de qualité.

Suivant cette démarche, tous les paramètres du design sont modifiables, des plus anodins comme le choix du répertoire de sauvegarde des résultats jusqu'aux plus fondamentaux comme les critères de spécificité (Figure 66). Enfin, l'ensemble des résultats est consultable au moyen de fichiers intermédiaires.

Dans un souci de qualité générale du processus, nous avons intégré plusieurs niveaux de validation. Le premier niveau de validation consiste à tester la présence du gène considéré dans la banque de séquence choisie (voir 9.9.2 Validation de la séquence appât). Cette étape, qui paraît anodine voir inutile, trouve tout son sens lorsque l'on considère que les séquences soumises peuvent provenir de banques de séquences différentes de celle utilisées pour le design. Il peut en résulter des différences de tailles entre les séquences des ARNm du même gène, notamment au niveau des parties 3' et 5' UTR. Si la séquence présente dans la banque utilisée pour le design est plus longue dans sa partie 3' UTR, les conséquences peuvent être dramatiques. Le risque réside dans la possibilité de choisir une sonde positionnée à une distance trop importante de l'extrémité 3' du véritable ARNm. Cette sonde ne pourra alors pas détecter sa cible (souvent produite par RT-PCR à partir du poly-T).

Dans notre cas, les séquences d'Actinome utilisées pour le design sont validées par RetScope à partir de la banque généraliste GenBank. Cependant les banques de séquences utilisées lors du design sont RefSeq et UniGene (voir ci-dessous). Par exemple, dans le cas du design d'Actichip v2.0, 40 séquences (<10%) ne remplissaient pas les critères définis par ce module. L'étude de chaque cas a permis de valider la séquence initiale et d'assurer la validité du design de sondes (exemple Figure 99).

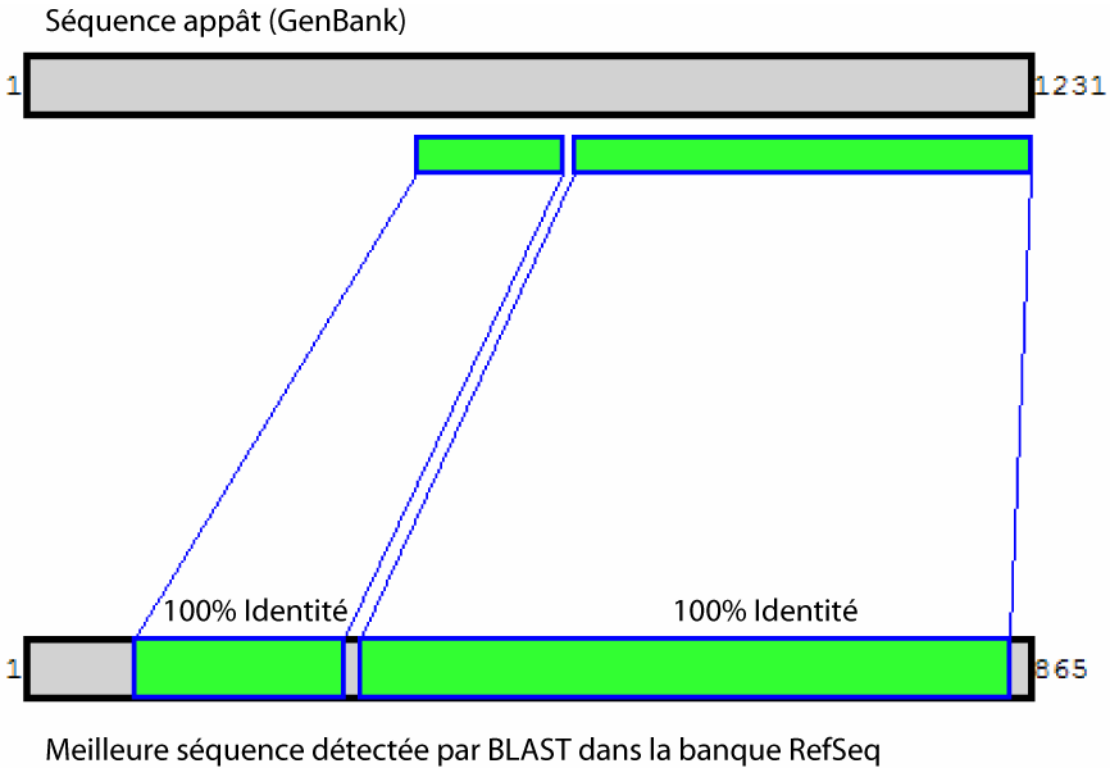


Figure 99 Exemple d'une différence entre une séquence appât et l'équivalent dans la banque RefSeq.

Illustration dans l'interface de BlastPanel (Chalmel F. non publié) des alignements entre la séquence appât soumise au design et son meilleur représentant dans la banque RefSeq. Le design pour le gène HSP27 illustre la différence qu'il peut y avoir entre une séquence issue de GenBank et son équivalent dans RefSeq. Le pourcentage d'identité garanti le fait que ces 2 séquences correspondent bien au même gène. Dans ce cas, la différence porte sur la partie 5' de la séquence et n'aura aucune incidence sur le choix d'une sonde.

Cependant, le design préliminaire de la version 3.0 a révélé près de 50% de séquences problématiques par rapport à celles de la banque RefSeq. Nous avons de fait analysé l'état de mise à jour de nos séquences GenBank utilisées. Nos séquences initiales sont en effet stockées dans des fichiers dans la banque Actinome. Néanmoins, bien que stockées nos séquences n'ont pas subi de mises à jours majeures pouvant expliquer cette différence. La raison probable est l'effort plus important de validation réalisé sur ce groupe en particulier dans RefSeq. Pour le set Actichip 3.0, ceci nous a incité à revoir notre stratégie en remplaçant nos séquences initiales par leurs équivalents RefSeq. Ce travail est en cours de réalisation par Arnaud Muller sur la plateforme de bioinformatique de Luxembourg et devrait mener au design final de la version 3.0 d'Actichip.

La spécificité d'une sonde pour un gène en particulier doit être évaluée en tenant compte de l'ensemble des autres gènes existant dans la cellule. Il existe différentes banques d'ARNm susceptibles de contenir l'ensemble des ARNm exprimés par une cellule. Les 2 banques majeures sont RefSeq et UniGene.

Dans le but d'utiliser ces différentes banques de séquences pour valider les sondes choisies, nous avons implémenté la possibilité de comparer et de choisir une sonde au moyen de 2 banques en même temps. Ce point constitue le deuxième niveau de validation de CADO4MI. De façon pratique, le design est réalisé sur chaque banque de séquence séparément puis la « sélection croisée » permet de choisir une sonde en tenant compte du résultat obtenu avec les 2 banques. Le résultat est consultable et modifiable au travers de l'interface graphique.

Ainsi, pour Actichip, nous utilisons les 2 banques de références que sont RefSeq et UniGene. RefSeq est une collection de séquences de références généralement validées par une expertise humaine (voir 2.2.2.2 RefSeq). L'avantage de cette banque réside justement dans la qualité des séquences qui y sont représentées, mais également par ses annotations et l'absence de redondance. Cependant, RefSeq est également moins complète en terme de nombre de transcrits. Ce dernier point semble cependant grandement s'améliorer depuis les dernières versions de la banque en 2006. UniGene est une banque contenant des groupes de séquences (gènes connus, ESTs…) représentant des gènes (voir 2.2.2.3 UniGene). Cette banque plus exhaustive représente l'ensemble des séquences exprimées (le transcriptome) de la cellule. Cependant, cette exhaustivité se traduit par un degré de redondance élevé et des annotations incomplètes voir fausses.

14.2.2 Les paramètres du design

Le choix des paramètres de design pour Actichip a été réalisé de façon à tenir compte, et ceci en dépit des différentes phases de sa construction, de l'ensemble des séquences qui y seront intégrées.

Les sondes utilisées sur Actichip ont une taille de 60 bases. Cette taille de sondes représente le meilleur rapport entre la spécificité et la sensibilité de détection du signal (Hughes *et al.* 2001). Nous avons choisi d'éliminer les séquences contenant une succession de 7 bases identiques (par exemple GGGGGGG). Ces séquences sont susceptibles de poser des problèmes notamment lors de la synthèse. Le design a été effectué sur la base d'une région englobant au maximum 3000 bases à partir de l'extrémité 3' de chaque ARNm. Le Tm doit être de 92 ± 5°C et le pourcentage en GC compris entre 35 et 70%.

Les choix du Tm et du GC ont été guidés par l'estimation de toutes les valeurs de Tm et de pourcentage en GC possibles pour toutes nos séquences (Figure 100). L'estimation des paramètres est une étape importante qui permet au final d'optimiser les résultats (Li *et al.* 2001). En effet, des valeurs de Tm et de pourcentage en GC homogènes assureront une hybridation uniforme. La moyenne de toutes les valeurs de Tm calculée avec le modèle du

plus proche voisin est de 91,8°C. La distribution des pourcentages en GC est centrée autour de 51,1%.

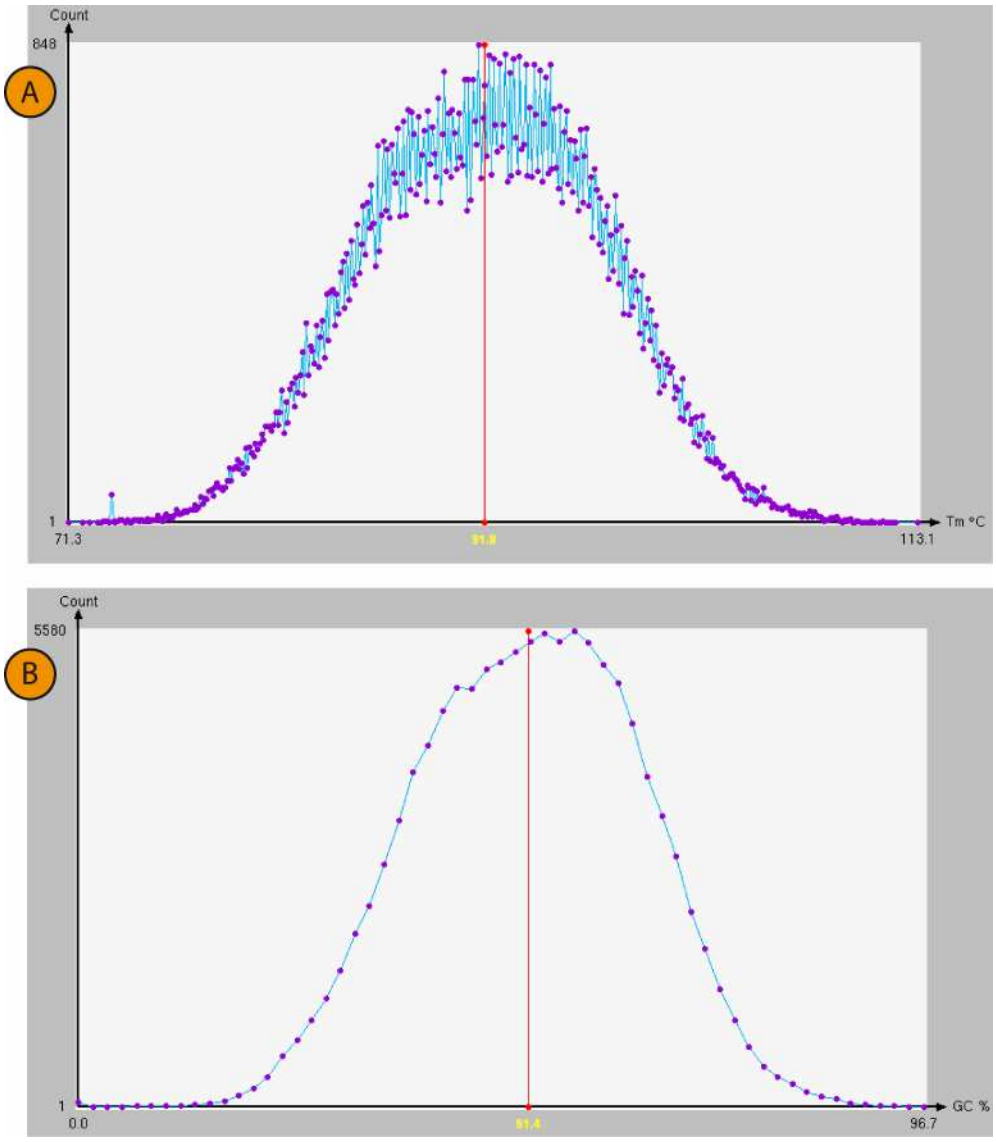


Figure 100 Estimation du Tm et du GC pour les séquences utilisées par Actichip.
(A) Représentation de la distribution de toutes les valeurs de Tm calculées par le modèle du plus proche voisin. (B) Représentation de toutes les valeurs de GC calculées. Les moyennes sont indiquées en rouge.

Il est intéressant de comparer ces résultats avec les valeurs calculées sur l'ensemble des séquences disponibles dans la banque RefSeq (Figure 101). Le groupe de séquences utilisé pour Actichip ne diffère pas sur ces 2 critères de la moyenne de toutes les séquences disponibles dans la banque RefSeq.

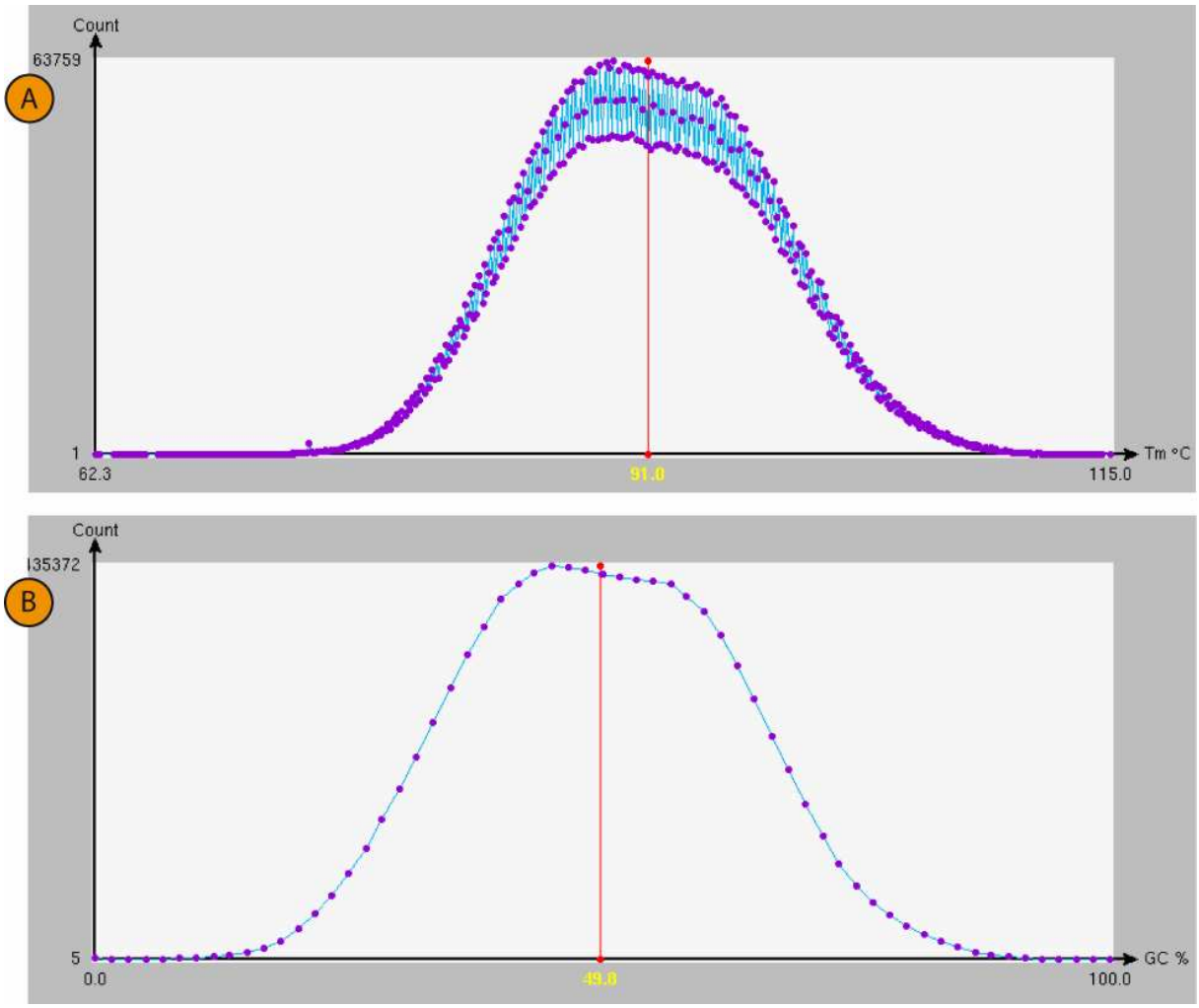


Figure 101 Estimation du Tm et du GC pour les séquences présentes dans RefSeq.
(A) Représentation de la distribution de toutes les valeurs de Tm calculées par le modèle du plus proche voisin. (B) Représentation de toutes les valeurs de GC calculées. Les moyennes sont indiquées en rouge. Les données sont calculées avec la « release 16 » de la banque RefSeq.

Le design a été réalisé une première fois en utilisant la banque RefSeq et une seconde fois avec la banque UniGene. La sélection finale des sondes a été conduite en combinant les 2 designs.

14.2.3 Résultat du design des sondes

Pour les 327 gènes, la procédure automatique de design a permis de choisir sans aucune intervention, 81,7% des sondes pour RefSeq seule, 86,4% dans le cas d'UniGene seule et 73,4% pour la combinaison des 2 banques. Ces résultats correspondent au pourcentage de sondes spécifiques, c'est-à-dire aux sondes qui ne détectent qu'une seule séquence *in silico*.

Ces données illustrent plusieurs points. D'une part, il existe bel et bien une différence entre les 2 banques de séquences utilisées (Figure 102). Leur utilisation simultanée ne permet pas d'augmenter le nombre de séquences automatiquement choisies, mais à l'inverse, de le

baisser. Ceci signifie que l'utilisation des 2 banques en simultanée permet d'éliminer des faux positifs de part et d'autres (voir 14.2.4 L'apport de l'utilisation de 2 banques).

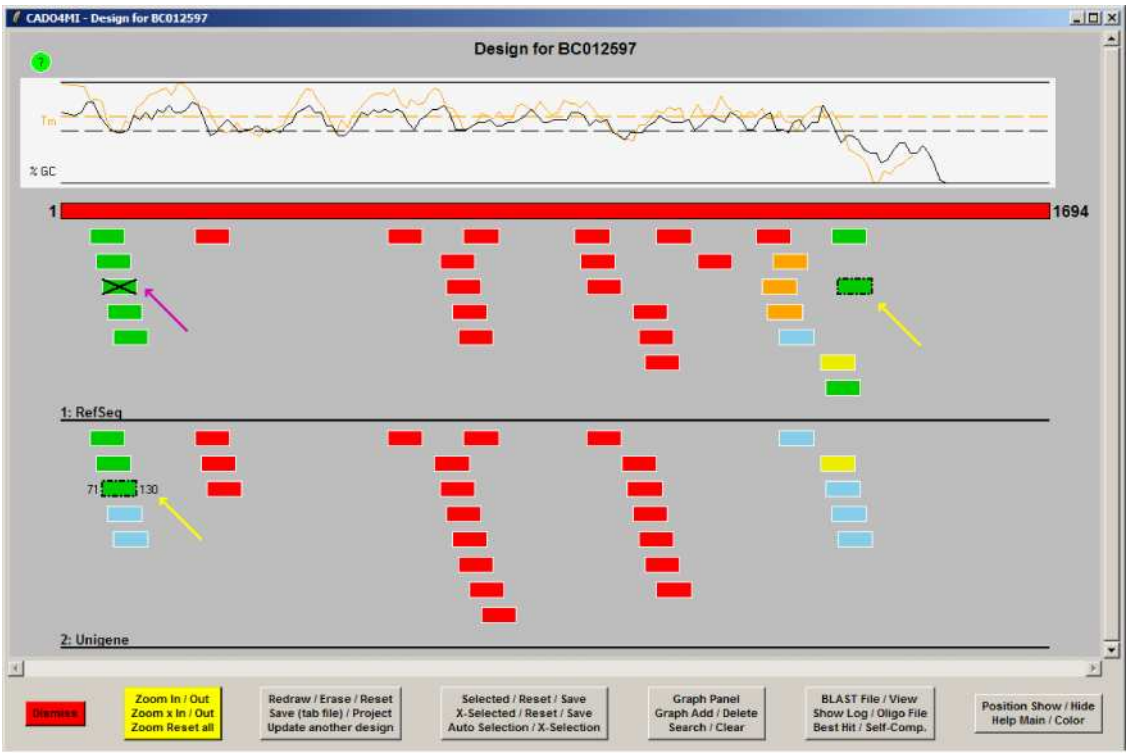


Figure 102 Exemple de l'utilisation de 2 banques de séquences différentes.
Cette fenêtre de CADO4MI illustre le résultat du design pour le gène de l'actine alpha de muscle squelettique dans les 2 banques de séquences, RefSeq (partie du haut) et UniGene (partie du bas). La sonde choisie pour chaque banque individuellement est indiquée par une flèche jaune (également avec un contour en pointillés noirs). Dans ce cas, on constate que la sonde choisie pour RefSeq correspond à une sonde non spécifique (couleur bleue) dans la banque UniGene. L'utilisation des 2 banques permet de choisir une autre sonde plus spécifique indiquée par une flèche mauve (également marquée par une croix).

D'autre part, 26,6% soit 87 sondes doivent être réexaminées afin de les valider ou de les écarter et, puis de choisir de nouvelles sondes. Ce sont des sondes potentiellement capables de détecter plusieurs séquences. Parmi les différentes complications observées, la plus fréquente concerne la détection de plusieurs variants d'épissage du même gène (42 cas). Dans ce cas précis, nous avons décidé de choisir la sonde capable de détecter le maximum de variants. Parmi les autres complications, on notera par exemple, la présence de doublons dans la banque UniGene (14 cas) et la détection de pseudogènes (12 cas). Pour choisir des sondes spécifiques, l'analyse de ces gènes a nécessité le redesign de 37 sondes en utilisant des critères moins restrictifs. Une sonde a pu être trouvée pour chacun des gènes avec néanmoins quelques écarts par rapport aux critères optimaux. Ainsi, la limite inférieure du Tm a du être abaissé dans 4 cas, de même pour le pourcentage de GC dans 1 cas et pour la

distance de l'extrémité 3' dans 5 cas. Enfin, 5 sondes ne permettront pas de distinguer uniquement le gène correspondant.

Au final, pour les 327 gènes inclus dans la deuxième version d'Actichip, nous avons choisi 355 sondes. Ce set d'oligonucléotides est caractérisé par un Tm moyen de 91,3°C, un %GC de 50,2% une distance moyenne à la fin 3' de 701 bases. Seules 4 sondes ne possèdent pas les critères de Tm requis, le Tm le plus bas est de 83,3°C. Une seule sonde possède un %GC de 31,7% par rapport au 35% préconisé.

14.2.4 L'apport de l'utilisation de 2 banques

Dans le point précédent, nous avons brièvement évoqué une différence entre les designs réalisés en automatique selon les banques de données utilisées. Cependant, qu'apporte réellement l'utilisation de 2 banques RefSeq et Unigene au choix des sondes ?

Nous avons ainsi comparé la sonde choisie par l'utilisation des 2 banques simultanément (la sélection croisée) à celle choisie si nous avions considéré chaque banque individuellement.

En considérant uniquement la banque RefSeq, 164 sondes sont identiques au choix final (sélection croisée). De même pour la banque UniGene, 206 sondes sont identiques au choix final. 108 sondes (environ 1/3) sont communes à ces résultats et indiquent le nombre de sondes correctement choisies en considérant ces banques individuellement (Figure 103). 219 sondes ne correspondent donc pas aux choix possibles avec les banques individuelles. En conséquence, l'utilisation combinée des banques permet de remettre en question le design d'environ 2/3 des sondes et ainsi d'améliorer le choix des sondes. Ce chiffre n'est pas négligeable et montre tout l'intérêt de notre stratégie.

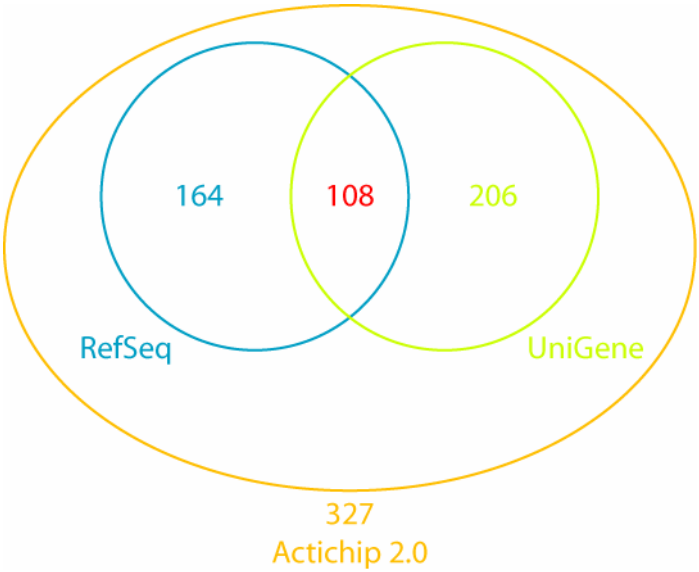


Figure 103 Intérêt de l'utilisation de 2 banques dans CADO4MI.
Comparaison du nombre de sondes retenues par la séléction croisée avec la sélection initiale pour chacune des 2 banques (UniGene et RefSeq). Pour les 327 gènes de la version 2.0 d'Actichip, seules 108 sondes (environ 1/3) sont identiques au choix initial d'UniGene et de RefSeq. Ainsi, 219 sondes (environ 2/3) ont bénéficiées de l'utilisation d'un design combinant ces 2 banques.

14.2.5 Les sondes de contrôle

Actichip intègre des contrôles négatifs et positifs. Ces contrôles internes permettent de valider le bon déroulement d'une expérience en vérifiant par exemple que les contrôles négatifs ne détectent aucun signal, ce qui serait révélateur d'une hybridation non-spécifique. Les contrôles négatifs sont en général des séquences provenant d'autres organismes que celui analysé. CADO4MI a ainsi permis de tester l'absence des séquences sondes provenant de 51 gènes bactériens et de 10 gènes de la plante *A. thaliana* dans les transcrits humain.

14.3 Validation et premiers résultats de la puce à ADN Actichip

La séquence des sondes étant définie, la prochaîne étape est la fabrication et la validation de la puce. Cette étape cruciale a été réalisée par la plateforme de puces à ADN du Luxembourg. L'ensemble de ces résultats est décrit dans une publication en cours de soumission et sera résumé dans les points qui suivent.

14.3.1 Fabrication d'Actichip

La synthèse des sondes a été réalisée par la société Eurogentec (Seraing, Belgium) à partir de la liste des sondes définies par CADO4MI. L'ensemble des sondes et des contrôles a été

déposé ou « spotté » en 3 exemplaires à une concentration de 25 μM, sur des lames ArrayIt pré-traitées à l'époxy (ArrayIt, Sunnyvale, CA, USA) par un robot Microgrid II microarrayer (Genomic solutions, Huntingdon, United Kingdom).

14.3.2 Comportement des sondes

Expérimentalement, le bon fonctionnement des sondes répond à 2 critères : la réactivité et la spécificité. La réactivité d'une sonde rend compte de sa capacité à s'hybrider à une séquence nucléotidique. Sur l'ensemble des expériences menées au laboratoire, seules 22 des 355 sondes n'ont jamais donné de signal. Plusieurs possibilités s'offrent alors. Soit la sonde a été mal choisie et ne peut donc pas détecter sa cible, soit la cible n'est pas présente ou en très faible quantité. Ce dernier cas est certainement le plus probable car l'expression de certains gènes est restreinte à des types cellulaires spécifiques. Par exemple, parmi les 22 sondes, 2 correspondent à des isoformes d'actine : l'actine du muscle lisse aortique (Alpha-actin 2) et l'actine alpha cardiaque. Les sondes représentant ces 2 isoformes spécifiques du muscle cardiaque et de l'appareil aortique n'ont jamais été testées dans des expériences transcriptomiques avec ce type de tissus. De plus, comme nous le verrons plus bas les sondes représentant ces 2 isoformes sont parfaitement fonctionnelles. La raison pour laquelle les 20 autres sondes n'ont réagi dans aucune expérience n'a pas pu être identifiée pour l'instant et les propriétés de ces sondes seront étudiées plus en détail ultérieurement.

La spécificité d'une sonde, c'est-à-dire sa capacité à s'hybrider non pas simplement à une séquence nucléotidique mais exclusivement à la séquence du gène pour lequel elle a été choisie, a été testée dans une expérience menée par André Mehlen sur la plateforme de puces à ADN. Les isoformes d'actine représentent par leur haut degré de similarité un exemple parfait pour tester la spécificité des sondes designées par CADO4MI. Dans ce contexte nous avons donc choisi de tester spécifiquement les sondes des 6 isoformes d'actine.

Cette expérience consiste à produire des fragments de PCR de chacune des isoformes d'actine et de les tester séparément dans différentes réactions d'hybridations sur la puce Actichip. Le but est de vérifier qu'il n'y a pas d'hybridations croisées entre les cibles et les sondes représentant les différents gènes d'actine

Comme on peut le voir sur la Figure 104 chaque sonde a été capable de détecter spécifiquement sa cible, démontrant par là même la qualité des sondes « designées ». De plus, aucun autre signal indiquant des réactions croisées n'a été observé pour les autres actines et pour l'ensemble de la puce.

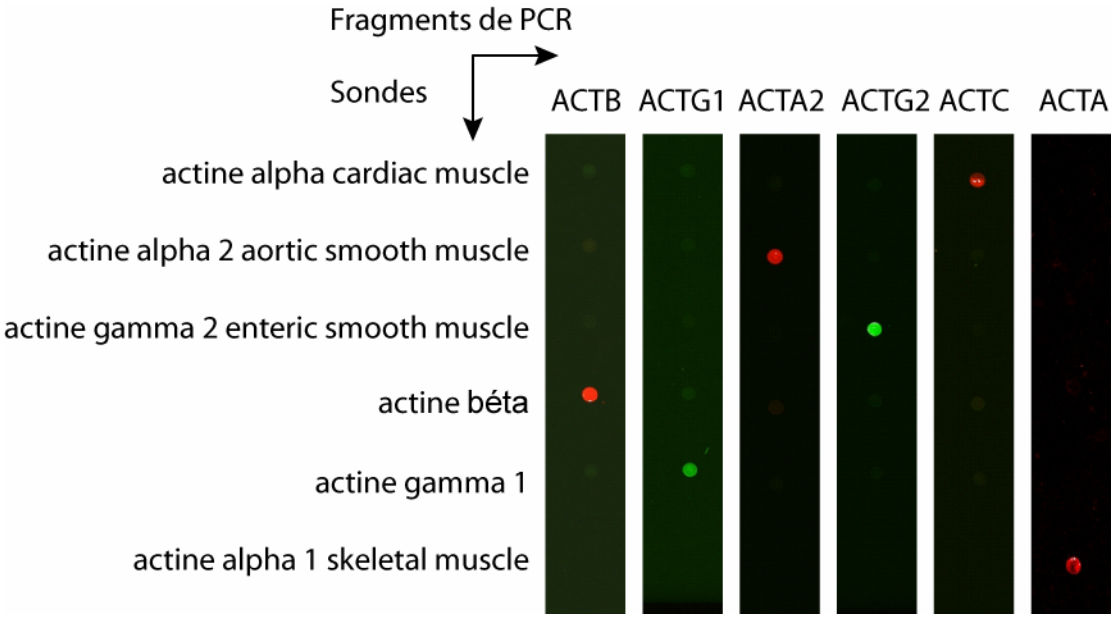


Figure 104 Spécificité des sondes pour les isoformes d'actine.
Des fragments de PCR correspondants aux 6 isoformes d'actine ont été générés à partir d'ADNc et marqués par des composés fluorescents. La figure montre les hybridations indépendantes de chacun de ses produits de PCR sur Actichip.

14.3.3 Une expérience de validation

Le comportement d'Actichip a été étudié avec des échantillons biologiques exprimant un répertoire spécifique de gènes du cytosquelette. Ainsi, 10 expériences de transcriptomique différentielle ont été menées sur du muscle squelettique humain et une lignée cellulaire de carcinome mammaire (MCF-7). Les ADNc cibles préparés à partir des ARNm de ces échantillons ont été marqués par deux fluorochromes différents et ont été hybridés aux sondes présentes sur ActiChip. L'analyse statistique de ces éxpériences a montré les qualités intrinsèques de la puce. Le signal fluorescent enregistré n'est que très peu sujet aux perturbations liées à la taille ou à la forme des spots. Les données sont reproductibles entre les différentes expériences. L'addition aux séquences cibles de quantités bien définies d'ARNm synthétiques (spike RNA) représentés par des sondes sur la puce, a permis d'établir la limite de détection de la puce. Actichip est capable de détecter potentiellement une copie d'ARNm par cellule.

L'étude des gènes du cytosquelette exprimés de façon différentielle dans ces 2 types cellulaires a permis de vérifier la capacité d'Actichip à détecter de façon fiable, différentes isoformes (Figure 105). L'analyse des profils d'expression de ces cellules par ActiChip montre que, par exemple, dans le muscle squelettique, on retrouve l'actine alpha squelettique, les myosines musculaires, les troponines, et la nebuline, alors que pour les

cellules MCF-7, on observe comme attendu l'expression de l'actine béta, des cytokératines 8 et 18, la filamine A, et l'ankyrine 3.

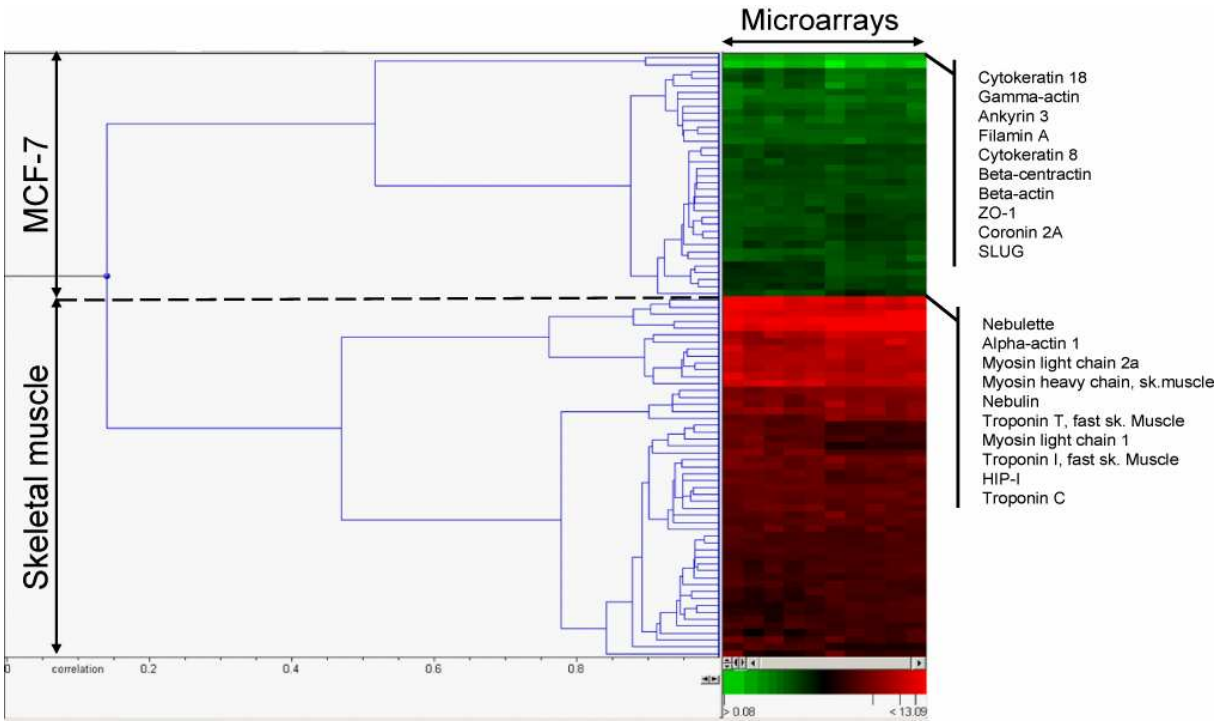


Figure 105 Dendrogramme représentant le résultat du clustering hierarchique.
Le dendrogramme représente le clustering hierarchique des gènes détectés dans au moins 2 expériences sur 3. Les gènes sont regroupés en fonction de la similarité de leurs profils d'expression exprimés par le rapport du log de leurs intensités. A droite, chaque ligne correspond à un gène alors que chaque colonne correspond à une expérience. Les 10 gènes les plus régulés sont indiqués. Les gènes exprimés préférentiellement dans le muscle squelettique ou dans les cellules MCF-7 sont indiqués repectivement en rouge et en vert.

Une expérience complémentaire utilisant les mêmes échantillons a permis de comparer Actichip à 2 autres plateformes de puces à ADN (Affymetrix et Opéron). L'étude des gènes communs aux 3 plateformes (10.3 Comparaison de puces) a montré qu'Actichip est aussi performante que la puce Affymetrix (U133A 2.0) et plus performante que la puce Opéron. Une analyse détaillée des profils d'expression des isoformes d'actine montre des différences entre les 3 plateformes.

	Actichip		Affymetrix		Opéron	
	MCF-7	Muscle	MCF-7	Muscle	MCF-7	Muscle
actine alpha 1 skeletal muscle	-	+	-	+	+	+
actine alpha 2 aortic smooth muscle	-	-	+	+	-	+
actine béta	+	+	+	+	+	+
actine alpha cardiac muscle 1	-	-	-	+	/	/
actine gamma 1	+	+	+	+	-	-
actine gamma 2 enteric smooth muscle	-	-	+	+	-	-

Tableau 19 Expression des isoformes actine dans les 3 plateformes de puces à ADN.
Les profils d'expression des 6 isoformes d'actine dans chaque échantillon et pour les 3 plateformes de puces à ADN sont représentés par « + » (gène exprimé) ou un « - » (gène non exprimé). Une seule isoforme n'existe pas dans le set Opéron (marqué par un « / »).

Les profils d'expression des isoformes d'actine obtenus par Actichip sont conformes à ceux attendus pour les 2 tissus (Khaitlina 2001). Ces résultats sont également cohérents avec la qualité des sondes testées de façon synthétique (14.3.2 Comportement des sondes).

14.3.4 Une première application d'Actichip

Une application sur une lignée cellulaire humaine montre également l'intérêt de l'utilisation d'Actichip.

Cette expérience préliminaire a été réalisée au sein de la plateforme de puces à ADN du Luxembourg dans un projet visant à déterminer les régulations des gènes du cytosquelette d'actine lors de changements de son organisation. La comparaison implique le transcriptome de 2 lignées cellulaires issues d'un clone MCF-7 résistant au TNF-α développé dans le laboratoire de Chouaib *et al* (Cai *et al.* 1997).

L'analyse phénotypique de ces 2 lignées cellulaires montre clairement 2 états typiques du cytosquelette ; l'un correspondant à une cellule de type épithéliale caractérisée par des cellules jointives et polarisées, et l'autre, correspondant à un type cellulaire mésenchymateux caractérisé par des cellules individualisées et des extensions de la membrane comme des filopodes (Figure 106).

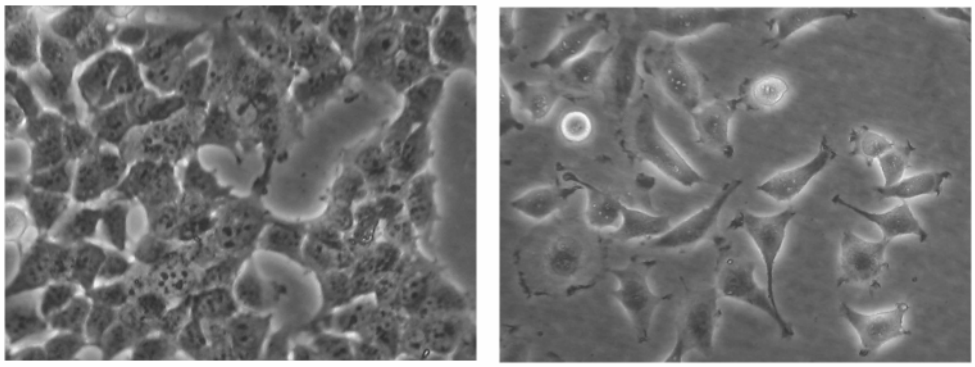


Figure 106 Image en microscopie à contraste de phase des 2 lignées cellulaires.
A gauche la lignée cellulaire MCF7 correspondant à un phénotype de cellule épithéliale (cellules jointives avec peu d'extensions membranaires) et à droite le clone 1001 caractéristique du phénotype mésenchymateux (cellule individualisée avec de grandes et nombreuses extensions membranaires).

Les résultats préliminaires de l'analyse de 3 expériences Actichip ont montrés de grands changements du niveau d'expression des gènes du cytosquelette (Figure 107) avec environ 91 gènes régulés négativement et 112 gènes régulés positivement. L'analyse des gènes les plus dérégulés fait apparaître des marqueurs classiques des 2 tissus et notamment de l'EMT comme la vimentine, snail, la desmoplakine et les cytokératines. L'EMT (*Epithelial-Mesenchymal Transition*) est un processus de différenciation cellulaire responsable du passage d'une cellule de type épithéliale (non motile) à cellule mésenchymateuse (motile) (revue dans (Thiery 2002; Lee *et al.* 2006; Thiery *et al.* 2006)). Cette transition implique la perte des adhérences inter-cellulaires, la perte de la polarité des cellules et l'acquisition du comportement migratoire (impliquant le remodelage du cytosquelette d'actine). Une analyse plus approfondie devrait permettre de mieux comprendre les différences entres les 2 états de ses cellules et peut être de vérifier s'l y a eu une EMT ou pas.

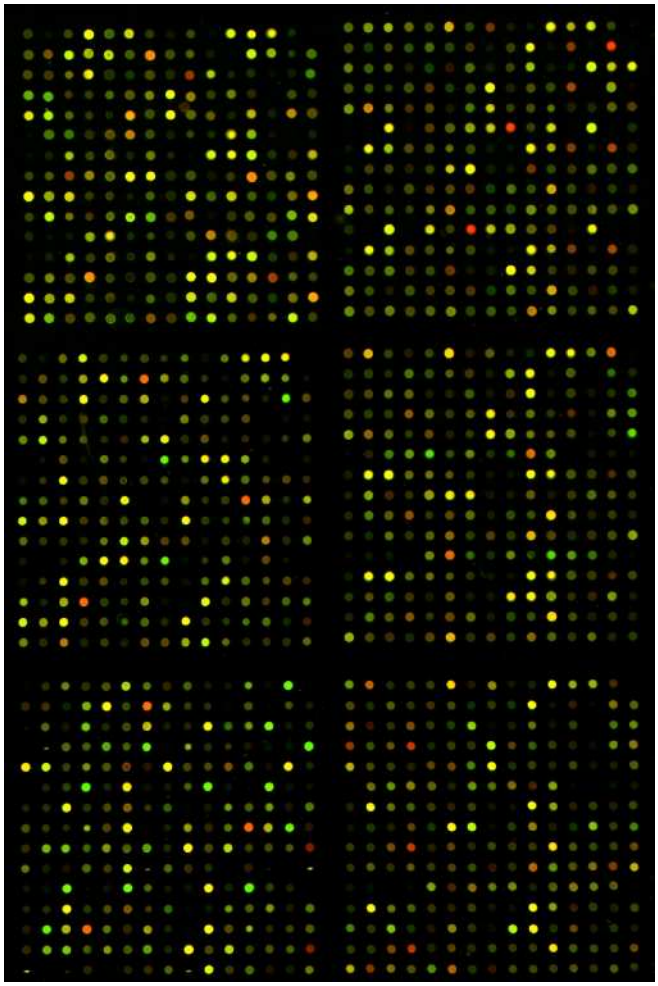


Figure 107 Image d'une expérience d'hybridation avec Actichip.
La puce Actichip est composée de 6 blocs contenant chacun 225 spots. L'échantillon correspondant aux cellules MCF-7 est marqué à la Cy3 (vert) alors que l'échantillon correspondant au clone 1001 est marqué à la Cy5 (rouge).

14.4 Conclusions et perspectives

Ce travail a permis de développer Actichip, une puce à ADN dédiée au cytosquelette d'actine. Au vu des premiers résultats, Actichip semble être une plateforme fonctionnelle capable de discriminer des tissus différents et des isoformes différentes. Son développement est basé sur un set d'oligonucléotides optimisés correspondant aux gènes du cytosquelette d'actine. Les sondes ont été calculées au moyen d'un nouveau logiciel de design de sondes, CADO4MI, que nous avons développé au laboratoire.

L'analyse fonctionnelle du comportement des sondes semble indiquer que la stratégie retenue pour CADO4MI est pleinement satisfaisante. Encore une fois, l'utilisation croisée de plusieurs banques de données permet d'assurer une meilleure qualité des résultats.

L'application d'Actichip à des projets biologiques impliquant le cytosquelette permettra de révéler les régulations des différents éléments.

Enfin, la dernière version de la puce est en cours de construction. Le design du troisième set de sondes qui compte environ 1200 gènes supplémentaires devrait permettre à Actichip 3.0 de couvrir l'ensemble des gènes des 3 cytosquelettes. La fabrication d'Actichip 3.0 est prévue pour la fin 2006 ou au tout début 2007.

Conclusions et perspectives

Chapitre 15 - Conclusions et perspectives

Le travail réalisé au cours de cette thèse a contribué à souligner les forces et les faiblesses des analyses et comparaisons bioinformatiques des données massives générées dans un contexte d'approches à haut débit. Nous avons également mis en avant l'apport de cette discipline dans des études appliquées à la biologie et à la médecine. Ces leçons ont été obtenues en abordant un système biologique complexe bien défini, le cytosquelette, artisan majeur de nombreux processus de la croissance et du développement cellulaire. Les outils conçus visent à mieux comprendre les relations unissant les différents composants du cytosquelette, au travers d'approches combinant des analyses fondées sur l'évolution et sur les variations de l'expression génique.

De façon générale, la bioinformatique s'inscrit souvent dans une alternance entre des phases de développements informatiques et des phases de compréhension et d'analyse du système biologique. La programmation elle-même oscille entre la création d'outils de saisie, de visualisation ou de calculs statistiques visant à obtenir une meilleure intelligence des données et le développement de véritables modèles ou outils de prédiction permettant non seulement de recréer un système ou d'en tester certains paramètres, mais également de tester les limites de l'approche utilisée.

Nous avons ainsi développé pas moins de 3 applications majeures ; ComIcs pour la détermination des profils phylogénétiques, ARPAnno pour identifier les différentes ARPs et CADO4MI pour choisir des sondes spécifiques pour des puces à ADN.

L'étude des profils phylogénétiques des gènes du cytosquelette est un exemple frappant de la quantité et de la complexité des données ainsi que des difficultés auxquelles la bioinformatique et le bio-analyste doit faire face.

Dans ce système, nous avons du faire face au problème des erreurs de prédictions de gènes et de protéines dans les génomes complets. La conséquence directe est une mauvaise définition du protéome des organismes dans les banques de séquences protéiques qui rend très incertaine la notion de présence ou d'absence d'une protéine. Nous avons également du faire face à des problèmes moins habituels liés aux méthodes classiques de détection de similarité. A l'inverse du problème, fréquemment rencontré, de la difficulté à détecter des homologues éloignés, nous avons été confrontés aux complications introduites par l'existence de sous-familles homologues trop proches pour être efficacement distinguées. En

particulier, la superfamille des actines et Actin Related Proteins a été emblématique des limites d'une approche de génomique comparative. Nos efforts se sont alors concentrés sur une caractérisation et une compréhension approfondies des membres appartenant à cette superfamille par la création et l'analyse d'un alignement multiple de l'ensemble de ces séquences. La puissance intégrative de l'alignement multiple de séquences complètes n'est certes plus a démontrée, cependant là encore, il a révélé sa pertinence en permettant de définir les éléments spécifiques et discriminants (insertions, délétions, domaines, zones, résidus…) qui sont la base du serveur ARPAnno. ARPanno représente l'aboutissement du travail fin réalisé sur cette famille de protéines, en proposant un outil d'identification, de discrimination et finalement, d'annotation des séquences homologues de l'actine. L'étude des ARPs est réellement un cas emblématique de l'ensemble des efforts consentis, pour analyser une famille de protéines, allant de la prédiction de gènes dans les génomes à l'analyse structurale des conservations en passant par la définition d'éléments de séquences discriminants et la prédiction d'interactions. Les enseignements sont multiples et concernent aussi bien une meilleure compréhension des fonctions cellulaires, des structures et de l'évolution des complexes multiprotéiques auxquels participent ces protéines que la définition de nouvelles stratégies bioinformatiques afin d'aboutir à des études de génomique comparative réellement efficientes.

Un autre exemple du lien direct qui peut exister entre l'analyse in silico et les applications en biologie et en médecine est illustré par la découverte de 2 nouveaux gènes responsables du syndrome Bardet-Biedl. BBS10 et BBS12 permettent d'identifier à eux seul 25% des patients atteints du BBS. Ces 2 gènes sont formidablement bien adaptés aux méthodes diagnostiques modernes puisque leurs parties codantes ne sont constituées que de 2 exons pour BBS10 et 1 seul pour BBS12. Par delà ces aspects médicaux, notre analyse, basée sur l'alignement multiple et les profils phylogénétiques, a révélé le rôle prépondérant joué par une famille particulière de protéines spécifiques des organismes vertébrés : les chaperonines de type II. Ces gènes sont très certainement impliqués dans la fonction ciliaire et peut être, dans certaines spécialisations apparues chez les organismes supérieurs. A ce jour, notre analyse constitue la caractérisation la plus complète de gènes du BBS et ouvre le champ à des hypothèses dans bon nombre de domaines (structural, fonctionnel…).

Un autre point important récurrent de cette thèse est l'effort constant pour estimer la qualité et la fiabilité des données utilisées. En effet, le signal biologique peut rapidement être masqué par l'accumulation des nombreuses erreurs générées à chacune des étapes des technologies et études à haut débit (séquençage, annotation structurale ou fonctionnelle, génomique fonctionnelle…). Ainsi, nous avons toujours tenté d'utiliser plusieurs sources de

données complémentaires pour croiser les résultats. Ceci est illustré à plusieurs reprises ; aussi bien lors du design de sondes d'Actichip où l'apport de RefSeq et d'UniGene est indéniable que lors des analyses de profils phylogénétiques où l'utilisation de 2 banques protéiques (UniProt et RefSeq) à permis d'optimiser le recouvrement des protéomes des différents organismes. Le développement de CADO4MI a été organisé autour de cette exigence de qualité. Dans le contexte du choix de sondes spécifiques des gènes du cytosquelette, ceci est devenue une nécessité afin de pouvoir distinguer sans ambiguïté des isoformes proches.

Ce travail de thèse aura généré un nombre important de résultats et mis à disposition plusieurs programmes. Il aura également posé un nombre important de questions qui seront à la base d'expériences supplémentaires.

Sur le plan bioinformatique, ce travail nous permet d'envisager l'analyse complète, automatique et validée des profils phylogénétiques des quelques 1800 gènes du cytosquelette afin d'aboutir à une meilleure compréhension des grands événements intervenus dans la constitution et l'évolution du cytosquelette. Cependant, au-delà de l'étude du cytosquelette, cet ensemble d'outils développés autour de COMICS permettra enfin d'exploiter en temps réel les nouveaux génomes d'eucaryotes. Ceci ouvre la perspective d'une génomique comparative à la hauteur des enjeux de l'ère post-génomique et à même d'intégrer les enseignements et messages cachés dans les organismes exotiques en cours de séquençage.

Sur le plan expérimental, un projet est en cours afin d'obtenir des structures 3D de BBS10 et BBS12, il devrait permettre de révéler le repliement spatial de ces protéines et notamment, des insertions spécifiques qui jouent sans doute un rôle majeur dans l'établissement d'un complexe « chaperonine-like » d'un nouveau genre. Des expériences complémentaires de 2-hybride sont également en cours afin de découvrir, et de valider, les éventuels partenaires ou associations.

Enfin et surtout, l'application de la puce à ADN, Actichip permettra de mieux comprendre les régulations fines des éléments du cytosquelette. Des applications majeures sont possibles, on pense notamment à la cancérologie et à l'étude de l'établissement de l'état métastatique qui implique de grands changements du cytosquelette d'actine. Dans le futur, ces puces dédiées, telle Actichip, sont amenées à ouvrir un nouveau chapitre de la médecine. Ces puces seront sans doute au cœur des outils de diagnostique/prognostique de demain susceptibles d'offrir sur une vision plus globale des différents processus biologiques perturbés lors de l'émergence ou de la progression d'une maladie.

Annexes

Annexe 1 Exemple d'une entrée au format EMBL

```
ID   BC026355; SV 1; linear; mRNA; STD; HUM; 3347 BP.
XX
AC   BC026355;
XX
DT   09-APR-2002 (Rel. 71, Created)
DT   13-AUG-2006 (Rel. 88, Last updated, Version 15)
XX
DE   Homo sapiens Bardet-Biedl syndrome 10, mRNA (cDNA clone MGC:26830
DE   IMAGE:4817227), complete cds.
XX
KW   MGC.
XX
OS   Homo sapiens (human)
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC   Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae;
OC   Homo.
XX
RN   [1]
RC   NIH-MGC Project URL: http://mgc.nci.nih.gov
RP   1-3347
RG   NIH MGC Project
RA   ;
RT   ;
RL   Submitted (02-APR-2002) to the EMBL/GenBank/DDBJ databases.
RL   National Institutes of Health, Mammalian Gene Collection (MGC), Bethesda,
RL   MD 20892-2590, USA
XX
DR   H-InvDB; HIT000040216.
DR   RZPD; IMAGp998C2010717.
DR   RZPD; IRAKp961N1132.
XX
CC   Contact: MGC help desk
CC   Email: cgapbs-r@mail.nih.gov
CC   Toshiyuki and Piero Carninci (RIKEN)
CC   cDNA Library Arrayed by: The I.M.A.G.E. Consortium (LLNL)
CC   This clone was selected for full length sequencing because it
CC   passed the following selection criteria: matched mRNA gi: 31377692.
CC   Differences found between this sequence and the human reference
CC   genome (build 36) are described in misc_difference features below
CC   and these differences were also compared to chimpanzee genomic
CC   sequences available as of 09/15/2004.
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..3347
FT                   /organism="Homo sapiens"
FT                   /lab_host="DH10B"
FT                   /mol_type="mRNA"
FT                   /clone_lib="NIH_MGC_95"
FT                   /clone="MGC:26830 IMAGE:4817227"
FT                   /tissue_type="Brain, hippocampus"
FT                   /note="Vector: pBluescriptR"
FT                   /db_xref="taxon:9606"
FT                   /db_xref="RZPD:IMAGp998C2010717"
FT                   /db_xref="RZPD:IRAKp961N1132"
FT   gene            1..3347
FT                   /gene="BBS10"
FT                   /note="synonym: FLJ23560"
FT   CDS             70..2241
FT                   /codon_start=1
FT                   /gene="BBS10"
FT                   /product="Bardet-Biedl syndrome 10"
FT                   /db_xref="GOA:Q8TAM1"
FT                   /db_xref="HGNC:26291"
FT                   /db_xref="UniProtKB/Swiss-Prot:Q8TAM1"
FT                   /protein_id="AAH26355.2"
FT                   /translation="MLSSMAAAGSVKAALQVAEVLEAIVSCCVGPEGRQVLCTKPTGEV
FT                   LLSRNGGRLLEALHLEHPIARMIVDCVSSHLKKTGDGAKTFIIFLCHLLRGLHAITDRE
FT                   KDPLMCENIQTHGRHWKNCSRWKFISQALLTFQTQILDGIMDQYLSRHFLSIFSSAKER
FT                   TLCRSSLELLLEAYFCGRVGRNNHKFISQLMCDYFFKCMTCKSGIGVFELVDDHFVELN
FT                   VGVTGLPVSDSRIIAGLVLQKDFSVYRPADGDMRMVIVTETIQPLFSTSGSEFILNSEA
FT                   QFQTSQFWIMEKTKAIMKHLHSQNVKLLISSVKQPDLVSYYAGVNGISVVECLSSEEVS
FT                   LIRRIIGLSPFVPPQAFSQCEIPNTALVKFCKPLILRSKRYVHLGLISTCAFIPHSIVL
FT                   CGPVHGLIEQHEDALHGALKMLRQLFKDLDLNYMTQTNDQNGTSSLFIYKNSGESYQAP
FT                   DPGNGSIQRPYQDTVAENKDALEKTQTYLKVHSNLVIPDVELETYIPYSTPTLTPTDTF
FT                   QTVETLTCLSLERNRLTDYYEPLLKNNSTAYSTRGNRIEISYENLQVTNITRKGSMLPV
FT                   SCKLPNMGTSQSYLSSSMPAGCVLPVGGNFDILLHYYLLNYAKKCHQSEETMVSMIIAN
FT                   ALLGIPKVLYKSKTGKYSFPHTYIRAVHALQTNQPLVSSQTGLESVMGKYQLLTSVLQC
FT                   LTKILTIDMVITVKRHPQKVHNQDSEDEL"
FT   misc_difference 1449
FT                   /gene="BBS10"
FT                   /note="'G' in cDNA is 'T' in the human genome; no amino
FT                   acid change. The chimpanzee genome agrees with the human
FT                   genomic sequence and not the cDNA."
```

```
FT   misc_difference 1890
FT                   /gene="BBS10"
FT                   /note="'T' in cDNA is 'G' in the human genome; amino acid
FT                   difference: 'D' in cDNA, 'E' in the human genome. The
FT                   chimpanzee genome agrees with the human genomic sequence
FT                   and not the cDNA."
FT   misc_difference 2863
FT                   /gene="BBS10"
FT                   /note="'G' in cDNA is 'A' in the human genome. The
FT                   chimpanzee genome agrees with the human genomic sequence
FT                   and not the cDNA."
FT   misc_difference 3292
FT                   /gene="BBS10"
FT                   /note="'T' in cDNA is 'C' in the human genome."
FT   misc_difference 3326
FT                   /gene="BBS10"
FT                   /note="'A' in cDNA is 'T' in the human genome."
FT   misc_difference 3328
FT                   /gene="BBS10"
FT                   /note="'A' in cDNA is 'C' in the human genome."
FT   misc_difference 3330..3347
FT                   /gene="BBS10"
FT                   /note="polyA tail: 18 bases do not align to the human
FT                   genome."
XX
SQ   Sequence 3347 BP; 1033 A; 603 C; 628 G; 1083 T; 0 other;
     gttcccaccc ctgttttcgg tcggcccggg tgttctgcaa gctggtcaaa aaggggaagc        60
     ggctcagata tgttaagttc tatggccgct gcagggtctg tgaaggcggc gttgcaggtg       120
     gccgaggtgc tggaagccat cgtgagctgc tgcgtggggc ccgagggacg gcaagttttg       180
     tgtacgaagc ccactggcga ggtgcttctc agccggaatg gaggccgcct cctggaggcg       240
     ctacacttag agcatcccat agccaggatg atagtggact gtgtttccag tcatctcaaa       300
     aaaacaggag atggtgcaaa aacatttatt atctttcttt gccatttgct tagaggactt       360
     catgcaatca cagacagaga aaaggatcct ttgatgtgtg aaaacattca aacccatgga       420
     aggcattgga aaaattgttc tcggtggaaa tttatttccc aggctctcct aacgtttcag       480
     acacaaatat tagacggtat tatggaccag tacctaagta gacacttttt gtctatcttt       540
     tcgtctgcta aagagagaac attgtgtagg agctctttag agttgctctt agaagcatac       600
     ttttgtggaa gagtgggaag aaataatcat aaatttattt cacagttgat gtgtgactac       660
     tttttcaagt gtatgacttg taaaagtggg attggtgtat ttgagttagt ggatgaccat       720
     tttgtagagt tgaatgttgg tgtcactggc cttcctgttt cagattccag gatcatagct       780
     ggtcttgtgc ttcagaaaga tttttctgtg taccgcccag cagatggtga catgcgaatg       840
     gtgatagtaa cagaaaccat tcagcctctt ttttccactt ctggatcaga gtttattcta       900
     aattcagaag cacagtttca gacatctcaa ttttggatta tggaaaagac aaaagcaata       960
     atgaaacatc tacatagtca gaatgtaaaa ttgctcatat ctagtgtgaa acaaccagat      1020
     ttagttagtt attatgcagg ggtgaatggc atatcagtgg ttgagtgttt atcatcagaa      1080
     gaagtttctc ttatccggag gatcattggt ctttctccat ttgtaccacc acaggccttt      1140
     tcgcagtgtg aaatacctaa cactgctttg gtgaaatttt gtaaacctct tatccttaga      1200
     tccaaaagat atgttcatct aggcttgata agcacatgtg catttatacc acactctata      1260
     gttctttgtg gaccagtgca tggtctcatt gaacaacatg aggatgcttt acatggagca      1320
     cttaaaatgc ttcggcaatt atttaaagac cttgatctaa attacatgac acaaaccaat      1380
     gaccaaaatg gcacttcaag tctttttatt tataagaaca gtggagaaag ttatcaagca      1440
     ccagatccgg gtaatggctc aatacaaagg ccttatcagg acacagttgc agagaacaaa      1500
     gatgcattgg aaaaaactca aacatattta aaagtacatt ctaatttggt aattccagat      1560
     gtagaattag aaacatatat tccgtattca acccccacac tgacaccaac agatacattc      1620
     caaacagttg aaacgctgac atgtttgtct ttggaaagaa acaggctaac tgattattat      1680
     gaaccattac tcaagaacaa ttccactgct tattcaacaa ggggaaatag aatagaaatt      1740
     tcttacgaaa atttacaggt cacaaatatt actagaaagg gaagcatgtt accagtgagc      1800
     tgtaagttac cgaatatggg tacttcccag agttaccttt cctcatctat gccagctggt      1860
     tgtgttttgc cagtaggtgg taattttgat atcttgttac attactatct tctcaattat      1920
     gccaaaaaat gccatcaatc agaagaaacc atggttagta tgataatagc taatgcactt      1980
     ttaggcattc ccaaagtcct ttataaatct aaaacaggaa agtacagctt tccacataca      2040
     tatataagag ctgtccatgc actgcaaacc aatcaaccct tggtaagcag tcagacaggt      2100
     ttggaatcag taatgggtaa ataccagcta ctaacttcag ttcttcagtg tttgacaaaa      2160
     atattaacca ttgacatggt aatcactgtt aagagacacc ctcagaaagt tcacaatcaa      2220
     gattcagaag atgaactata acatcagaag tttttaatta accaaacttt tcatctaact      2280
     caagccaagt aaagcagtca tgtgaccact ggttctaaag tcagttcagt ctacttagga      2340
     aaatagcgta actttaaaag tctttagaag aagcacacta aggtcaccag accagataca      2400
     aatattaaat tactttatgg aacaaatcta gaggggaagc caagatttgg ctaagtgtgt      2460
     ctgttttttc cctattttat gcctctgtgt ctcagctctg tgttagccta tgtgttagg       2520
     ggagggtttt tctttatagc tccttttttac tctcctgtat ctttttcact ccagccctcc      2580
     ttcatcgtta catgtttagt tcatagaatc atttaatctc tgatttgggt gggcttattc      2640
     taattgtttt taatattgaa tacattattt gcattaattt tccctactca tactttgtaa      2700
     agctgagtaa aaggctcaaa ttatttttttt caaaaagcat aaaattaaat tagcagtgag      2760
     taaaaggctc aaatttttttt tcaaaaagca taaaattaat ttttactttt atgtggtcat      2820
     tggtttactg ccacttcatt tggaaaactt ggatagattt tcgcctttga tatacctttg      2880
     aatatatgtt acctgaaata taactgtgca ttgttaactc tttcatttct gtagtaaaag      2940
     gttaatacta gaaaggatat gcaattaata ctttgatttt ctcctgaccc caagagcttg      3000
     taaggatatg tcatgtattt actggttttt cttgtatctg gtgcatagcc agagttccac      3060
     agtaacaaat aatttgacaa tttttatttc taatgtttat ttctgtttta tttttaattt      3120
     ttatttctaa tttatttcat ttacaaatgt tcatatttta aaaacttgtc aatgtaaata      3180
     atatgatgca tcttatcatg gacaaggaca gtgttttcta cctttatcag ttctctgtaa      3240
     tacccaaaac agtgctgtat ttactccaag tattcagaag tgcttgttga ataaaacagt      3300
     gttatcttta attcattcct ttaaaaaaaa agaaaaaaaa aaaaaaa                    3347
//
```

Annexe 2 Exemple d'une entrée au format GenBank

```
LOCUS       BC026355               3347 bp    mRNA    linear   PRI 11-AUG-2006
DEFINITION  Homo sapiens Bardet-Biedl syndrome 10, mRNA (cDNA clone MGC:26830
            IMAGE:4817227), complete cds.
ACCESSION   BC026355
VERSION     BC026355.1  GI:20072253
KEYWORDS    MGC.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 3347)
  AUTHORS   Strausberg,R.L., Feingold,E.A., Grouse,L.H., Derge,J.G.,
            Klausner,R.D., Collins,F.S., Wagner,L., Shenmen,C.M., Schuler,G.D.,
            Altschul,S.F., Zeeberg,B., Buetow,K.H., Schaefer,C.F., Bhat,N.K.,
            Hopkins,R.F., Jordan,H., Moore,T., Max,S.I., Wang,J., Hsieh,F.,
            Diatchenko,L., Marusina,K., Farmer,A.A., Rubin,G.M., Hong,L.,
            Stapleton,M., Soares,M.B., Bonaldo,M.F., Casavant,T.L.,
            Scheetz,T.E., Brownstein,M.J., Usdin,T.B., Toshiyuki,S.,
            Carninci,P., Prange,C., Raha,S.S., Loquellano,N.A., Peters,G.J.,
            Abramson,R.D., Mullahy,S.J., Bosak,S.A., McEwan,P.J.,
            McKernan,K.J., Malek,J.A., Gunaratne,P.H., Richards,S.,
            Worley,K.C., Hale,S., Garcia,A.M., Gay,L.J., Hulyk,S.W.,
            Villalon,D.K., Muzny,D.M., Sodergren,E.J., Lu,X., Gibbs,R.A.,
            Fahey,J., Helton,E., Ketteman,M., Madan,A., Rodrigues,S.,
            Sanchez,A., Whiting,M., Madan,A., Young,A.C., Shevchenko,Y.,
            Bouffard,G.G., Blakesley,R.W., Touchman,J.W., Green,E.D.,
            Dickson,M.C., Rodriguez,A.C., Grimwood,J., Schmutz,J., Myers,R.M.,
            Butterfield,Y.S., Krzywinski,M.I., Skalska,U., Smailus,D.E.,
            Schnerch,A., Schein,J.E., Jones,S.J. and Marra,M.A.
  CONSRTM   Mammalian Gene Collection Program Team
  TITLE     Generation and initial analysis of more than 15,000 full-length
            human and mouse cDNA sequences
  JOURNAL   Proc. Natl. Acad. Sci. U.S.A. 99 (26), 16899-16903 (2002)
   PUBMED   12477932
REFERENCE   2  (bases 1 to 3347)
  CONSRTM   NIH MGC Project
  TITLE     Direct Submission
  JOURNAL   Submitted (02-APR-2002) National Institutes of Health, Mammalian
            Gene Collection (MGC), Bethesda, MD 20892-2590, USA
  REMARK    NIH-MGC Project URL: http://mgc.nci.nih.gov
COMMENT     Contact: MGC help desk
            Email: cgapbs-r@mail.nih.gov
            Clone distribution: MGC clone distribution information can be found
            through the I.M.A.G.E. Consortium/LLNL at: http://image.llnl.gov
            Series: IRAK Plate: 32 Row: n Column: 11
            This clone was selected for full length sequencing because it
            passed the following selection criteria: matched mRNA gi: 31377692.

            Differences found between this sequence and the human reference
            genome (build 36) are described in misc_difference features below
            and these differences were also compared to chimpanzee genomic
            seqeunces available as of 09/15/2004.
FEATURES             Location/Qualifiers
     source          1..3347
                     /organism="Homo sapiens"
                     /mol_type="mRNA"
                     /db_xref="taxon:9606"
                     /clone="MGC:26830 IMAGE:4817227"
                     /tissue_type="Brain, hippocampus"
                     /clone_lib="NIH_MGC_95"
                     /lab_host="DH10B"
                     /note="Vector: pBluescriptR"
     gene            1..3347
                     /gene="BBS10"
                     /note="synonym: FLJ23560"
                     /db_xref="GeneID:79738"
                     /db_xref="HGNC:26291"
     CDS             70..2241
                     /gene="BBS10"
                     /codon_start=1
                     /product="Bardet-Biedl syndrome 10"
                     /protein_id="AAH26355.2"
                     /db_xref="GI:112180700"
                     /db_xref="GeneID:79738"
                     /db_xref="HGNC:26291"
                     /translation="MLSSMAAAGSVKAALQVAEVLEAIVSCCVGPEGRQVLCTKPTGE
                     VLLSRNGGRLLEALHLEHPIARMIVDCVSSHLKKTGDGAKTFIIFLCHLLRGLHAITD
                     REKDPLMCENIQTHGRHWKNCSRWKFISQALLTFQTQILDGIMDQYLSRHFLSIFSSA
                     KERTLCRSSLELLLEAYFCGRVGRNNHKFISQLMCDYFFKCMTCKSGIGVFELVDDHF
                     VELNVGVTGLPVSDSRIIAGLVLQKDFSVYRPADGDMRMVIVTETIQPLFSTSGSEFI
                     LNSEAQFQTSQFWIMEKTKAIMKHLHSQNVKLLISSVKQPDLVSYYAGVNGISVVECL
                     SSEEVSLIRRIIGLSPFVPPQAFSQCEIPNTALVKFCKPLILRSKRYVHLGLISTCAF
                     IPHSIVLCGPVHGLIEQHEDALHGALKMLRQLFKDLDLNYMTQTNDQNGTSSLFIYKN
```

```
                     SGESYQAPDPGNGSIQRPYQDTVAENKDALEKTQTYLKVHSNLVIPDVELETYIPYST
                     PTLTPTDTFQTVETLTCLSLERNRLTDYYEPLLKNNSTAYSTRGNRIEISYENLQVTN
                     ITRKGSMLPVSCKLPNMGTSQSYLSSSMPAGCVLPVGGNFDILLHYYLLNYAKKCHQS
                     EETMVSMIIANALLGIPKVLYKSKTGKYSFPHTYIRAVHALQTNQPLVSSQTGLESVM
                     GKYQLLTSVLQCLTKILTIDMVITVKRHPQKVHNQDSEDEL"
     misc_difference 1449
                     /gene="BBS10"
                     /note="'G' in cDNA is 'T' in the human genome; no amino
                     acid change.  The chimpanzee genome agrees with the human
                     genomic sequence and not the cDNA."
     misc_difference 1890
                     /gene="BBS10"
                     /note="'T' in cDNA is 'G' in the human genome; amino acid
                     difference: 'D' in cDNA, 'E' in the human genome.  The
                     chimpanzee genome agrees with the human genomic sequence
                     and not the cDNA."
     misc_difference 3328
                     /gene="BBS10"
                     /note="'A' in cDNA is 'C' in the human genome."
     misc_difference 3330..3347
                     /gene="BBS10"
                     /note="polyA tail: 18 bases do not align to the human
                     genome."
ORIGIN
        1 gttcccaccc ctgttttcgg tcggcccggg tgttctgcaa gctggtcaaa aaggggaagc
       61 ggctcagata tgttaagttc tatggccgct gcagggtctg tgaaggcggc gttgcaggtg
      121 gccgaggtgc tggaagccat cgtgagctgc tgcgtggggc ccgagggacg gcaagttttg
      181 tgtacgaagc ccactggcga ggtgcttctc agccggaatg gaggccgcct cctggaggcg
      241 ctacacttag agcatcccat agccaggatg atagtggact gtgtttccag tcatctcaaa
      301 aaaacaggag atggtgcaaa aacatttatt atctttcttt gccatttgct tagaggactt
      361 catgcaatca cagacagaga aaaggatcct ttgatgtgtg aaaacattca aacccatgga
      421 aggcattgga aaaattgttc tcggtggaaa tttatttccc aggctctcct aacgtttcag
      481 acacaaatat tagacggtat tatggaccag tacctaagta gacacttttt gtctatcttt
      541 tcgtctgcta aagagagaac attgtgtagg agctctttag agttgctctt agaagcatac
      601 ttttgtggaa gagtgggaag aaataatcat aaatttattt cacagttgat gtgtgactac
      661 tttttcaagt gtatgacttg taaaagtggg attggtgtat ttgagttagt ggatgaccat
      721 tttgtagagt tgaatgttgg tgtcactggc cttcctgttt cagattccag gatcatagct
      781 ggtcttgtgc ttcagaaaga tttttctgtg taccgcccag cagatggtga catgcgaatg
      841 gtgatagtaa cagaaaccat tcagcctctt ttttccactt ctggatcaga gtttattcta
      901 aattcagaag cacagtttca gacatctcaa ttttggatta tggaaaagac aaaagcaata
      961 atgaaacatc tacatagtca gaatgtaaaa ttgctcatat ctagtgtgaa acaaccagat
     1021 ttagttagtt attatgcagg ggtgaatggc atatcagtgg ttgagtgttt atcatcagaa
     1081 gaagtttctc ttatccggag gatcattggt ctttctccat ttgtaccacc acaggccttt
     1141 tcgcagtgtg aaatacctaa cactgctttg gtgaaatttt gtaaacctct tatccttaga
     1201 tccaaaagat atgttcatct aggcttgata agcacatgtg catttatacc acactctata
     1261 gttctttgtg gaccagtgca tggtctcatt gaacaacatg aggatgcttt acatggagca
     1321 cttaaaatgc ttcggcaatt atttaaagac cttgatctaa attacatgac acaaaccaat
     1381 gaccaaaatg gcacttcaag tctttttatt tataagaaca gtggagaaag ttatcaagca
     1441 ccagatccgg gtaatggctc aatacaaagg ccttatcagg acacagttgc agagaacaaa
     1501 gatgcattgg aaaaaactca aacatattta aaagtacatt ctaatttggt aattccagat
     1561 gtagaattag aaacatatat tccgtattca acccccacac tgacaccaac agatacattc
     1621 caaacagttg aaacgctgac atgtttgtct ttggaaagaa acaggctaac tgattattat
     1681 gaaccattac tcaagaacaa ttccactgct tattcaacaa ggggaaatag aatagaaatt
     1741 tcttacgaaa atttacaggt cacaaatatt actagaaagg gaagcatgtt accagtgagc
     1801 tgtaagttac cgaatatggg tacttcccag agttaccttt cctcatctat gccagctggt
     1861 tgtgttttgc cagtaggtgg taattttgat atcttgttac attactatct tctcaattat
     1921 gccaaaaaat gccatcaatc agaagaaacc atggttagta tgataatagc taatgcactt
     1981 ttaggcattc ccaaagtcct ttataaatct aaaacaggaa agtacagctt tccacataca
     2041 tatataagag ctgtccatgc actgcaaacc aatcaaccct tggtaagcag tcagacaggt
     2101 ttggaatcag taatgggtaa ataccagcta ctaacttcag ttcttcagtg tttgacaaaa
     2161 atattaacca ttgacatggt aatcactgtt aagagacacc ctcagaaagt tcacaatcaa
     2221 gattcagaag atgaactata acatcagaag tttttaatta accaaacttt tcatctaact
     2281 caagccaagt aaagcagtca tgtgaccact ggttctaaag tcagttcagt ctacttagga
     2341 aaatagcgta actttaaaag tctttagaag aagcacacta aggtcaccag accagataca
     2401 aatattaaat tactttatgg aacaaatcta gaggggaagc caagatttgg ctaagtgtgt
     2461 ctgtttttttc cctattttat gcctctgtgt ctcagctctg tgttagccta tgtgtttagg
     2521 ggagggtttt tctttatagc tccttttac tctcctgtat ctttttcact ccagccctcc
     2581 ttcatcgtta catgtttagt tcatagaatc atttaatctc tgatttgggt gggcttattc
     2641 taattgtttt taatattgaa tacattattt gcattaattt tcctactca tactttgtaa
     2701 agctgagtaa aaggctcaaa ttatttttt caaaaagcat aaaattaaat tagcagtgag
     2761 taaaaggctc aaatttttt tcaaaaagca taaaattaat ttttactttt atgtggtcat
     2821 tggtttactg ccacttcatt tggaaaactt ggatagattt tcgcctttga tataccttg
     2881 aatatatgtt acctgaaata taactgtgca ttgttaactc tttcatttct gtagtaaaag
     2941 gttaatacta gaaaggatat gcaattaata ctttgatttt ctcctgaccc caagagcttg
     3001 taaggatatg tcatgtattt actggttttt cttgtatctg gtgcatagcc agagttccac
     3061 agtaacaaat aatttgacaa tttttatttc taatgtttat ttctgtttta tttttaattt
     3121 ttatttctaa tttatttcat ttacaaatgt tcatatttta aaaacttgtc aatgtaaata
     3181 atatgatgca tcttatcatg gacaaggaca gtgttttcta cctttatcag ttctctgtaa
     3241 tacccaaaac agtgctgtat ttactccaag tattcagaag tgcttgttga ataaaacagt
     3301 gttatcttta attcattcct ttaaaaaaaa agaaaaaaaa aaaaaaa
//
```

Annexe 3 Exemple d'une entrée au format UniProt

```
ID   BBS10_HUMAN    STANDARD;      PRT;   723 AA.
AC   Q8TAM1; Q96CW2; Q9H5D2;
DT   16-MAY-2006, integrated into UniProtKB/Swiss-Prot.
DT   16-MAY-2006, sequence version 2.
DT   13-JUN-2006, entry version 31.
DE   Bardet-Biedl syndrome 10 protein.
GN   Name=BBS10; Synonyms=C12orf58;
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
OC   Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   NUCLEOTIDE SEQUENCE [LARGE SCALE MRNA].
RC   TISSUE=Brain, and Skin;
RX   MEDLINE=22388257; PubMed=12477932; DOI=10.1073/pnas.242603899;
RA   Strausberg R.L., Feingold E.A., Grouse L.H., Derge J.G.,
RA   Klausner R.D., Collins F.S., Wagner L., Shenmen C.M., Schuler G.D.,
RA   Altschul S.F., Zeeberg B., Buetow K.H., Schaefer C.F., Bhat N.K.,
RA   Hopkins R.F., Jordan H., Moore T., Max S.I., Wang J., Hsieh F.,
RA   Diatchenko L., Marusina K., Farmer A.A., Rubin G.M., Hong L.,
RA   Stapleton M., Soares M.B., Bonaldo M.F., Casavant T.L., Scheetz T.E.,
RA   Brownstein M.J., Usdin T.B., Toshiyuki S., Carninci P., Prange C.,
RA   Raha S.S., Loquellano N.A., Peters G.J., Abramson R.D., Mullahy S.J.,
RA   Bosak S.A., McEwan P.J., McKernan K.J., Malek J.A., Gunaratne P.H.,
RA   Richards S., Worley K.C., Hale S., Garcia A.M., Gay L.J., Hulyk S.W.,
RA   Villalon D.K., Muzny D.M., Sodergren E.J., Lu X., Gibbs R.A.,
RA   Fahey J., Helton E., Ketteman M., Madan A., Rodrigues S., Sanchez A.,
RA   Whiting M., Madan A., Young A.C., Shevchenko Y., Bouffard G.G.,
RA   Blakesley R.W., Touchman J.W., Green E.D., Dickson M.C.,
RA   Rodriguez A.C., Grimwood J., Schmutz J., Myers R.M.,
RA   Butterfield Y.S.N., Krzywinski M.I., Skalska U., Smailus D.E.,
RA   Schnerch A., Schein J.E., Jones S.J.M., Marra M.A.;
RT   "Generation and initial analysis of more than 15,000 full-length human
RT   and mouse cDNA sequences.";
RL   Proc. Natl. Acad. Sci. U.S.A. 99:16899-16903(2002).
RN   [2]
RP   NUCLEOTIDE SEQUENCE [LARGE SCALE MRNA] OF 512-723.
RC   TISSUE=Lung;
RX   PubMed=14702039; DOI=10.1038/ng1285;
RA   Ota T., Suzuki Y., Nishikawa T., Otsuki T., Sugiyama T., Irie R.,
RA   Wakamatsu A., Hayashi K., Sato H., Nagai K., Kimura K., Makita H.,
RA   Sekine M., Obayashi M., Nishi T., Shibahara T., Tanaka T., Ishii S.,
RA   Yamamoto J., Saito K., Kawai Y., Isono Y., Nakamura Y., Nagahari K.,
RA   Murakami K., Yasuda T., Iwayanagi T., Wagatsuma M., Shiratori A.,
RA   Sudo H., Hosoiri T., Kaku Y., Kodaira H., Kondo H., Sugawara M.,
RA   Takahashi M., Kanda K., Yokoi T., Furuya T., Kikkawa E., Omura Y.,
RA   Abe K., Kamihara K., Katsuta N., Sato K., Tanikawa M., Yamazaki M.,
RA   Ninomiya K., Ishibashi T., Yamashita H., Murakawa K., Fujimori K.,
RA   Tanai H., Kimata M., Watanabe M., Hiraoka S., Chiba Y., Ishida S.,
RA   Ono Y., Takiguchi S., Watanabe S., Yosida M., Hotuta T., Kusano J.,
RA   Kanehori K., Takahashi-Fujii A., Hara H., Tanase T.-O., Nomura Y.,
RA   Togiya S., Komai F., Hara R., Takeuchi K., Arita M., Imose N.,
RA   Musashino K., Yuuki H., Oshima A., Sasaki N., Aotsuka S.,
RA   Yoshikawa Y., Matsunawa H., Ichihara T., Shiohata N., Sano S.,
RA   Moriya S., Momiyama H., Satoh N., Takami S., Terashima Y., Suzuki O.,
RA   Nakagawa S., Senoh A., Mizoguchi H., Goto Y., Shimizu F., Wakebe H.,
RA   Hishigaki H., Watanabe T., Sugiyama A., Takemoto M., Kawakami B.,
RA   Yamazaki M., Watanabe K., Kumagai A., Itakura S., Fukuzumi Y.,
RA   Fujimori Y., Komiyama M., Tashiro H., Tanigami A., Fujiwara T.,
RA   Ono T., Yamada K., Fujii Y., Ozaki K., Hirao M., Ohmori Y.,
RA   Kawabata A., Hikiji T., Kobatake N., Inagaki H., Ikema Y., Okamoto S.,
RA   Okitani R., Kawakami T., Noguchi S., Itoh T., Shigeta K., Senba T.,
RA   Matsumura K., Nakajima Y., Mizuno T., Morinaga M., Sasaki M.,
RA   Togashi T., Oyama M., Hata H., Watanabe M., Komatsu T.,
RA   Mizushima-Sugano J., Satoh T., Shirai Y., Takahashi Y., Nakagawa K.,
RA   Okumura K., Nagase T., Nomura N., Kikuchi H., Masuho Y., Yamashita R.,
RA   Nakai K., Yada T., Nakamura Y., Ohara O., Isogai T., Sugano S.;
RT   "Complete sequencing and characterization of 21,243 full-length human
RT   cDNAs.";
RL   Nat. Genet. 36:40-45(2004).
RN   [3]
RP   VARIANTS BBS10 PRO-34; TRP-49; TRP-91; SER-170; TRP-195; CYS-197;
RP   GLY-240; PHE-308; ALA-311; LEU-329; LEU-363; SER-414; ARG-579;
RP   HIS-613; CYS-613; VAL-677 AND PRO-689.
RX   PubMed=16582908; DOI=10.1038/ng1771;
RA   Stoetzel C., Laurier V., Davis E.E., Muller J., Rix S., Badano J.L.,
RA   Leitch C.C., Salem N., Chouery E., Corbani S., Jalk N., Vicaire S.,
RA   Sarda P., Hamel C., Lacombe D., Holder M., Odent S., Holder S.,
RA   Brooks A.S., Elcioglu N.H., Da Silva E., Rossillion B., Sigaudy S.,
RA   de Ravel T.J., Alan Lewis R., Leheup B., Verloes A., Amati-Bonneau P.,
RA   Megarbane A., Poch O., Bonneau D., Beales P.L., Mandel J.-L.,
RA   Katsanis N., Dollfus H.;
RT   "BBS10 encodes a vertebrate-specific chaperonin-like protein and is a
```

```
RT   major BBS locus.";
RL   Nat. Genet. 38:521-524(2006).
CC   -!- FUNCTION: Probable molecular chaperone; assist the folding of
CC       proteins upon ATP hydrolysis (By similarity).
CC   -!- DISEASE: Defects in BBS10 are the cause of Bardet-Biedl syndrome
CC       type 10 (BBS10) [MIM:209900]. Bardet-Biedl syndrome (BBS) is a
CC       genetically heterogeneous, autosomal recessive disorder
CC       characterized by usually severe pigmentary retinopathy, early
CC       onset obesity, polydactyly, hypogenitalism, renal malformation and
CC       mental retardation.
CC   -!- SIMILARITY: Belongs to the TCP-1 chaperonin family.
CC   ---------------------------------------------------------------------
CC   Copyrighted by the UniProt Consortium, see http://www.uniprot.org/terms
CC   Distributed under the Creative Commons Attribution-NoDerivs License
CC   ---------------------------------------------------------------------
DR   EMBL; BC013795; AAH13795.1; ALT_INIT; mRNA.
DR   EMBL; BC026355; AAH26355.1; ALT_INIT; mRNA.
DR   EMBL; AK027213; BAB15695.1; ALT_INIT; mRNA.
DR   UniGene; Hs.96322; -.
DR   Ensembl; ENSG00000179941; Homo sapiens.
DR   HGNC; HGNC:26291; BBS10.
DR   MIM; 209900; phenotype.
DR   RZPD-ProtExp; IOH10848; -.
DR   RZPD-ProtExp; IOH11405; -.
DR   RZPD-ProtExp; W2661; -.
DR   InterPro; IPR008950; GroEL-ATPase.
KW   ATP-binding; Bardet-Biedl syndrome; Chaperone; Disease mutation;
KW   Nucleotide-binding; Obesity; Polymorphism; Sensory transduction;
KW   Vision.
FT   CHAIN         1    723       Bardet-Biedl syndrome 10 protein.
FT                                /FTId=PRO_0000235272.
FT   VARIANT      34     34       R -> P (in BBS10).
FT                                /FTId=VAR_026391.
FT   VARIANT      49     49       R -> W (in BBS10).
FT                                /FTId=VAR_026392.
FT   VARIANT      91     91       C -> W (in BBS10).
FT                                /FTId=VAR_026393.
FT   VARIANT     170    170       L -> S (in BBS10).
FT                                /FTId=VAR_026394.
FT   VARIANT     195    195       C -> W (in BBS10).
FT                                /FTId=VAR_026395.
FT   VARIANT     197    197       Y -> C (in BBS10).
FT                                /FTId=VAR_026396.
FT   VARIANT     240    240       V -> G (in BBS10).
FT                                /FTId=VAR_026397.
FT   VARIANT     308    308       L -> F (in BBS10).
FT                                /FTId=VAR_026398.
FT   VARIANT     311    311       S -> A (in BBS10).
FT                                /FTId=VAR_026399.
FT   VARIANT     329    329       S -> L (in BBS10).
FT                                /FTId=VAR_026400.
FT   VARIANT     363    363       P -> L (in BBS10).
FT                                /FTId=VAR_026401.
FT   VARIANT     376    376       L -> F (in dbSNP:11109474).
FT                                /FTId=VAR_026402.
FT   VARIANT     414    414       L -> S (in BBS10).
FT                                /FTId=VAR_026403.
FT   VARIANT     579    579       K -> R (in BBS10).
FT                                /FTId=VAR_026404.
FT   VARIANT     613    613       Y -> C (in BBS10).
FT                                /FTId=VAR_026405.
FT   VARIANT     613    613       Y -> H (in BBS10).
FT                                /FTId=VAR_026406.
FT   VARIANT     677    677       G -> V (in BBS10).
FT                                /FTId=VAR_026407.
FT   VARIANT     689    689       T -> P (in BBS10).
FT                                /FTId=VAR_026408.
FT   CONFLICT    514    514       T -> S (in Ref. 2; BAB15695).
FT   CONFLICT    586    586       S -> Y (in Ref. 2; BAB15695).
FT   CONFLICT    607    607       E -> D (in Ref. 1; AAH26355).
SQ   SEQUENCE   723 AA;  80838 MW;  558143FFA5F191DD CRC64;
     MLSSMAAAGS VKAALQVAEV LEAIVSCCVG PEGRQVLCTK PTGEVLLSRN GGRLLEALHL
     EHPIARMIVD CVSSHLKKTG DGAKTFIIFL CHLLRGLHAI TDREKDPLMC ENIQTHGRHW
     KNCSRWKFIS QALLTFQTQI LDGIMDQYLS RHFLSIFSSA KERTLCRSSL ELLLEAYFCG
     RVGRNNHKFI SQLMCDYFFK CMTCKSGIGV FELVDDHFVE LNVGVTGLPV SDSRIIAGLV
     LQKDFSVYRP ADGDMRMVIV TETIQPLFST SGSEFILNSE AQFQTSQFWI MEKTKAIMKH
     LHSQNVKLLI SSVKQPDLVS YYAGVNGISV VECLSSEEVS LIRRIIGLSP FVPPQAFSQC
     EIPNTALVKF CKPLILRSKR YVHLGLISTC AFIPHSIVLC GPVHGLIEQH EDALHGALKM
     LRQLFKDLDL NYMTQTNDQN GTSSLFIYKN SGESYQAPDP GNGSIQRPYQ DTVAENKDAL
     EKTQTYLKVH SNLVIPDVEL ETYIPYSTPT LTPTDTFQTV ETLTCLSLER NRLTDYYEPL
     LKNNSTAYST RGNRIEISYE NLQVTNITRK GSMLPVSCKL PNMGTSQSYL SSSMPAGCVL
     PVGGNFEILL HYYLLNYAKK CHQSEETMVS MIIANALLGI PKVLYKSKTG KYSFPHTYIR
     AVHALQTNQP LVSSQTGLES VMGKYQLLTS VLQCLTKILT IDMVITVKRH PQKVHNQDSE
     DEL
//
```

Annexe 4 Exemple de séquence au format FASTA

```
>BBS10_HUMAN Bardet-Biedl syndrome 10 protein.
MLSSMAAAGSVKAALQVAEVLEAIVSCCVGPEGRQVLCTKPTGEVLLSRNGGRLLEALHL
EHPIARMIVDCVSSHLKKTGDGAKTFIIFLCHLLRGLHAITDREKDPLMCENIQTHGRHW
KNCSRWKFISQALLTFQTQILDGIMDQYLSRHFLSIFSSAKERTLCRSSLELLLEAYFCG
RVGRNNHKFISQLMCDYFFKCMTCKSGIGVFELVDDHFVELNVGVTGLPVSDSRIIAGLV
LQKDFSVYRPADGDMRMVIVTETIQPLFSTSGSEFILNSEAQFQTSQFWIMEKTKAIMKH
LHSQNVKLLISSVKQPDLVSYYAGVNGISVVECLSSEEVSLIRRIIGLSPFVPPQAFSQC
EIPNTALVKFCKPLILRSKRYVHLGLISTCAFIPHSIVLCGPVHGLIEQHEDALHGALKM
LRQLFKDLDLNYMTQTNDQNGTSSLFIYKNSGESYQAPDPGNGSIQRPYQDTVAENKDAL
EKTQTYLKVHSNLVIPDVELETYIPYSTPTLTPTDTFQTVETLTCLSLERNRLTDYYEPL
LKNNSTAYSTRGNRIEISYENLQVTNITRKGSMLPVSCKLPNMGTSQSYLSSSMPAGCVL
PVGGNFEILLHYYLLNYAKKCHQSEETMVSMIIANALLGIPKVLYKSKTGKYSFPHTYIR
AVHALQTNQPLVSSQTGLESVMGKYQLLTSVLQCLTKILTIDMVITVKRHPQKVHNQDSE
DEL
```

Annexe 5 Exemple de fichier au format GAL (type ATF)

```
ATF   1.0
18    5
"Type=GenePix ArrayList V1.0"
"BlockCount=12"
"BlockType=0"
"Block1=8000, 12080, 150, 15, 280, 15, 280"
"Block2=12500, 12080, 150, 15, 280, 15, 280"
"Block3=8000, 16580, 150, 15, 280, 15, 280"
"Block4=12500, 16580, 150, 15, 280, 15, 280"
"Block5=8000, 21080, 150, 15, 280, 15, 280"
"Block6=12500, 21080, 150, 15, 280, 15, 280"
"Block7=8000, 52080, 150, 15, 280, 15, 280"
"Block8=12500, 52080, 150, 15, 280, 15, 280"
"Block9=8000, 56580, 150, 15, 280, 15, 280"
"Block10=12500, 56580, 150, 15, 280, 15, 280"
"Block11=8000, 61080, 150, 15, 280, 15, 280"
"Block12=12500, 61080, 150, 15, 280, 15, 280"
"Supplier=BioRobotics"
"ArrayerSoftwareName=TAS Application Suite (MicroGrid II)"
"ArrayerSoftwareVersion=2.4.0.2"
"Block"   "Column"          "Row"    "ID"    "Name"
1     1   1     ACT0456      Beta adducin (Erythrocyte adducin beta subunit).
1     1   2     ACT0390      Wiskott-Aldrich syndrome protein family member 2.
1     1   3     ACT0309      Dystrophin (Muscular dystrophy, Duchenne and Becker types).
1     1   4     ACT0151      Destrin (Actin-depolymerizing factor) (ADF).
1     1   5     ACT0036      Actin-related protein 3-beta.
1     1   6     buffer       buffer
1     1   8     ACT1671      Actin-related protein 8.
1     1   9     ACT0765      Elongation factor 1-gamma (EF-1-gamma).
1     1   10    ACT0615      Catenin delta-2 (Delta-catenin)
1     1   11    RCA          RCA
```

Références bibliographiques

(1998). "Genome sequence of the nematode C. elegans: a platform for investigating biology. The C. elegans Sequencing Consortium." Science **282**(5396): 2012-8.

Abrahamsen, M. S., T. J. Templeton, et al. (2004). "Complete genome sequence of the apicomplexan, Cryptosporidium parvum." Science **304**(5669): 441-5.

Adams, M. D., S. E. Celniker, et al. (2000). "The genome sequence of Drosophila melanogaster." Science **287**(5461): 2185-95.

Adams, M. D., J. M. Kelley, et al. (1991). "Complementary DNA sequencing: expressed sequence tags and human genome project." Science **252**(5013): 1651-6.

Akman, L., A. Yamashita, et al. (2002). "Genome sequence of the endocellular obligate symbiont of tsetse flies, Wigglesworthia glossinidia." Nat Genet **32**(3): 402-7.

Al-Lazikani, B., J. Jung, et al. (2001). "Protein structure prediction." Curr Opin Chem Biol **5**(1): 51-6.

Allawi, H. T. and J. SantaLucia, Jr. (1997). "Thermodynamics and NMR of internal G.T mismatches in DNA." Biochemistry **36**(34): 10581-94.

Altschul, S. F., W. Gish, et al. (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-10.

Altschul, S. F., T. L. Madden, et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res **25**(17): 3389-402.

Andrieux, A. and D. Job (2003). "[A role for the cytoskeleton in mental diseases?]." Med Sci (Paris) **19**(2): 135-7.

Anfinsen, C. B. and R. R. Redfield (1956). "Protein structure in relation to function and biosynthesis." Adv Protein Chem **48**(11): 1-100.

Ansley, S. J., J. L. Badano, et al. (2003). "Basal body dysfunction is a likely cause of pleiotropic Bardet-Biedl syndrome." Nature **425**(6958): 628-33.

Aparicio, S., J. Chapman, et al. (2002). "Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes." Science **297**(5585): 1301-10.

Apweiler, R., T. K. Attwood, et al. (2001). "The InterPro database, an integrated documentation resource for protein families, domains and functional sites." Nucleic Acids Res **29**(1): 37-40.

Arabidospis, I. (2000). "Analysis of the genome sequence of the flowering plant Arabidopsis thaliana." Nature **408**(6814): 796-815.

Archibald, J. M., J. M. Logsdon, Jr., et al. (2000). "Origin and evolution of eukaryotic chaperonins: phylogenetic evidence for ancient duplications in CCT genes." Mol Biol Evol **17**(10): 1456-66.

Armbrust, E. V., J. A. Berges, et al. (2004). "The genome of the diatom Thalassiosira pseudonana: ecology, evolution, and metabolism." Science **306**(5693): 79-86.

Arnold, K., L. Bordoli, et al. (2006). "The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling." Bioinformatics **22**(2): 195-201.

Ashburner, M., C. A. Ball, et al. (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet **25**(1): 25-9.

Ausmees, N., J. R. Kuhn, et al. (2003). "The bacterial cytoskeleton: an intermediate filament-like function in cell shape." Cell **115**(6): 705-13.

Badano, J. L., S. J. Ansley, et al. (2003). "Identification of a novel Bardet-Biedl syndrome protein, BBS7, that shares structural features with BBS1 and BBS2." Am J Hum Genet **72**(3): 650-8.

Badano, J. L., N. Mitsuma, et al. (2006). "The Ciliopathies: An Emerging Class of Human Genetic Disorders." Annu Rev Genomics Hum Genet.

Baker, W., A. van den Broek, et al. (2000). "The EMBL nucleotide sequence database." Nucleic Acids Res **28**(1): 19-23.

Bardet, G. (1920). "Sur un syndrome d'obesite infantile avec polydactylie et retinite pigmentaire (contribution a l'etude des formes cliniques de l'obesite hypophysaire)." Thèse.

Barker, W. C., F. Pfeiffer, et al. (1996). "Superfamily classification in PIR-International Protein Sequence Database." Methods Enzymol **266**: 59-71.

Barrans, J. D., D. Stamatiou, et al. (2001). "Construction of a human cardiovascular cDNA microarray: portrait of the failing heart." Biochem Biophys Res Commun **280**(4): 964-9.

Beales, P. L. (2005). "Lifting the lid on Pandora's box: the Bardet-Biedl syndrome." Curr Opin Genet Dev **15**(3): 315-23.

Beales, P. L., J. L. Badano, et al. (2003). "Genetic interaction of BBS1 mutations with alleles at other BBS loci can result in non-Mendelian Bardet-Biedl syndrome." Am J Hum Genet **72**(5): 1187-99.

Beales, P. L., N. Elcioglu, et al. (1999). "New criteria for improved diagnosis of Bardet-Biedl syndrome: results of a population survey." J Med Genet **36**(6): 437-46.

Bellis, M. and P. Casellas (1997). "La puce ADN : un multi-réacteur de paillasse." Médecine Sciences **13**: 1317-24.

Ben Hamida, C., N. Doerflinger, et al. (1993). "Localization of Friedreich ataxia phenotype with selective vitamin E deficiency to chromosome 8q by homozygosity mapping." Nat Genet **5**(2): 195-200.

Benson, D. A., I. Karsch-Mizrachi, et al. (2000). "GenBank." Nucleic Acids Res **28**(1): 15-8.

Benson, D. A., I. Karsch-Mizrachi, et al. (2006). "GenBank." Nucleic Acids Res **34**(Database issue): D16-20.

Bergen, L. G. and G. G. Borisy (1980). "Head-to-tail polymerization of microtubules in vitro. Electron microscope analysis of seeded assembly." J Cell Biol **84**(1): 141-50.

Berget, S. M., C. Moore, et al. (1977). "Spliced segments at the 5' terminus of adenovirus 2 late mRNA." Proc Natl Acad Sci U S A **74**(8): 3171-5.

Berriman, M., E. Ghedin, et al. (2005). "The genome of the African trypanosome Trypanosoma brucei." Science **309**(5733): 416-22.

Biedl, A. (1922). "Ein Geschwisterpaar mit adipose-genitaler Dystrophie." Dtsch Med Wochenschr **48:1630**.

Blacque, O. E., M. J. Reardon, et al. (2004). "Loss of C. elegans BBS-7 and BBS-8 protein function results in cilia defects and compromised intraflagellar transport." Genes Dev **18**(13): 1630-42.

Blattner, F. R., G. Plunkett, 3rd, et al. (1997). "The complete genome sequence of Escherichia coli K-12." Science **277**(5331): 1453-74.

Blessing, C. A., G. T. Ugrinova, et al. (2004). "Actin and ARPs: action in the nucleus." Trends Cell Biol **14**(8): 435-42.

Blumenthal, T., D. Evans, et al. (2002). "A global analysis of Caenorhabditis elegans operons." Nature **417**(6891): 851-4.

Bodkin, D. K. and D. L. Knudson (1985). "Assessment of sequence relatedness of double-stranded RNA genes by RNA-RNA blot hybridization." J Virol Methods **10**(1): 45-52.

Boguski, M. S., T. M. Lowe, et al. (1993). "dbEST--database for "expressed sequence tags"." Nat Genet **4**(4): 332-3.

Boldogh, I. R. and L. A. Pon (2006). "Interactions of mitochondria with the actin cytoskeleton." Biochim Biophys Acta **1763**(5-6): 450-62.

Bolton, E. T. and C. B. Mc (1962). "A general method for the isolation of RNA complementary to DNA." Proc Natl Acad Sci U S A **48**: 1390-7.

Bork, P., C. Sander, et al. (1992). "An ATPase domain common to prokaryotic cell cycle proteins, sugar kinases, actin, and hsp70 heat shock proteins." Proc Natl Acad Sci U S A **89**(16): 7290-4.

Borkovich, K. A., L. A. Alex, et al. (2004). "Lessons from the genome sequence of Neurospora crassa: tracing the path from genomic blueprint to multicellular organism." Microbiol Mol Biol Rev **68**(1): 1-108, table of contents.

Breathnach, R., J. L. Mandel, et al. (1977). "Ovalbumin gene is split in chicken DNA." Nature **270**(5635): 314-9.

Breslauer, K. J., R. Frank, et al. (1986). "Predicting DNA duplex stability from the base sequence." Proc Natl Acad Sci U S A **83**(11): 3746-50.

Bretscher, A. and K. Weber (1979). "Villin: the major microfilament-associated protein of the intestinal microvillus." Proc Natl Acad Sci U S A **76**(5): 2321-5.

Bretscher, A. and K. Weber (1980). "Fimbrin, a new microfilament-associated protein present in microvilli and other cell surface structures." J Cell Biol **86**(1): 335-40.

Brosius, J. (1999). "RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements." Gene **238**(1): 115-34.

Burke, D. T., G. F. Carle, et al. (1987). "Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors." Science **236**(4803): 806-12.

Cai, Y., J. Jin, et al. (2005). "The mammalian YL1 protein is a shared subunit of the TRRAP/TIP60 histone acetyltransferase and SRCAP complexes." J Biol Chem **280**(14): 13665-70.

Cai, Z., A. Bettaieb, et al. (1997). "Alteration of the sphingomyelin/ceramide pathway is associated with resistance of human breast carcinoma MCF7 cells to tumor necrosis factor-alpha-mediated cytotoxicity." J Biol Chem **272**(11): 6918-26.

Cairns, B. R., H. Erdjument-Bromage, et al. (1998). "Two actin-related proteins are shared functional components of the chromatin-remodeling complexes RSC and SWI/SNF." Mol Cell **2**(5): 639-51.

Carlton, J. M., S. V. Angiuoli, et al. (2002). "Genome sequence and comparative analysis of the model rodent malaria parasite Plasmodium yoelii yoelii." Nature **419**(6906): 512-9.

Casey, J. and N. Davidson (1977). "Rates of formation and thermal stabilities of RNA:DNA and DNA:DNA duplexes at high concentrations of formamide." Nucleic Acids Res **4**(5): 1539-52.

Cavalier-Smith, T. (2002). "The phagotrophic origin of eukaryotes and phylogenetic classification of Protozoa." Int J Syst Evol Microbiol **52**(Pt 2): 297-354.

Chadwick, B. P., J. Mull, et al. (1999). "Cloning, mapping, and expression of two novel actin genes, actin-like-7A (ACTL7A) and actin-like-7B (ACTL7B), from the familial dysautonomia candidate region on 9q31." Genomics **58**(3): 302-9.

Chang, L. and R. D. Goldman (2004). "Intermediate filaments mediate cytoskeletal crosstalk." Nat Rev Mol Cell Biol **5**(8): 601-13.

Chiang, A. P., J. S. Beck, et al. (2006). "Homozygosity mapping with SNP arrays identifies TRIM32, an E3 ubiquitin ligase, as a Bardet-Biedl syndrome gene (BBS11)." Proc Natl Acad Sci U S A **103**(16): 6287-92.

Chiang, A. P., D. Nishimura, et al. (2004). "Comparative genomic analysis identifies an ADP-ribosylation factor-like gene as the cause of Bardet-Biedl syndrome (BBS3)." Am J Hum Genet **75**(3): 475-84.

Chou, H. H., A. P. Hsia, et al. (2004). "Picky: oligo microarray design for large genomes." Bioinformatics **20**(17): 2893-902.

Chou, P. Y. and G. D. Fasman (1974). "Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins." Biochemistry **13**(2): 211-22.

Chow, L. T., R. E. Gelinas, et al. (1977). "An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA." Cell **12**(1): 1-8.

Clamp, M., J. Cuff, et al. (2004). "The Jalview Java alignment editor." Bioinformatics **20**(3): 426-7.

Clark, S. W. and M. D. Rose (2006). "Arp10p is a pointed-end-associated component of yeast dynactin." Mol Biol Cell **17**(2): 738-48.

Cohen, D., I. Chumakov, et al. (1993). "A first-generation physical map of the human genome." Nature **366**(6456): 698-701.

Collins, F. S. (1990). "Identifying human disease genes by positional cloning." Harvey Lect **86**: 149-64.

Consortium (1998). "Genome sequence of the nematode C. elegans: a platform for investigating biology. The C. elegans Sequencing Consortium." Science **282**(5396): 2012-8.

Coulombe, P. A., O. Bousquet, et al. (2000). "The 'ins' and 'outs' of intermediate filament organization." Trends Cell Biol **10**(10): 420-8.

Coulombe, P. A., L. Ma, et al. (2001). "Intermediate filaments at a glance." J Cell Sci **114**(Pt 24): 4345-7.

Crick, F. (1970). "Central dogma of molecular biology." Nature **227**(5258): 561-3.

Crick, F. H. (1958). "On protein synthesis." Symp Soc Exp Biol **12**: 138-63.

Dandekar, T., B. Snel, et al. (1998). "Conservation of gene order: a fingerprint of proteins that physically interact." Trends Biochem Sci **23**(9): 324-8.

Davis, E. E., M. Brueckner, et al. (2006). "The emerging complexity of the vertebrate cilium: new functional roles for an ancient organelle." Dev Cell **11**(1): 9-19.

Dayhoff, M. O. (1965). "Computer aids to protein sequence determination." J Theor Biol **8**(1): 97-112.

Dehal, P., Y. Satou, et al. (2002). "The draft genome of Ciona intestinalis: insights into chordate and vertebrate origins." Science **298**(5601): 2157-67.

del Sol Mesa, A., F. Pazos, et al. (2003). "Automatic methods for predicting functionally important residues." J Mol Biol **326**(4): 1289-302.

Delcher, A. L., D. Harmon, et al. (1999). "Improved microbial gene identification with GLIMMER." Nucleic Acids Res **27**(23): 4636-41.

Delcourt, S. G. and R. D. Blake (1991). "Stacking energies in DNA." J Biol Chem **266**(23): 15160-9.

Delgehyr, N., J. Sillibourne, et al. (2005). "Microtubule nucleation and anchoring at the centrosome are independent processes linked by ninein function." J Cell Sci **118**(Pt 8): 1565-75.

Dennis, C. and C. Surridge (2000). "Arabidopsis thaliana genome. Introduction." Nature **408**(6814): 791.

DeRisi, J., L. Penland, et al. (1996). "Use of a cDNA microarray to analyse gene expression patterns in human cancer." Nat Genet **14**(4): 457-60.

DeRisi, J. L., V. R. Iyer, et al. (1997). "Exploring the metabolic and genetic control of gene expression on a genomic scale." Science **278**(5338): 680-6.

Devos, D. and A. Valencia (2001). "Intrinsic errors in genome annotation." Trends Genet **17**(8): 429-31.

Diatchenko, L., Y. F. Lau, et al. (1996). "Suppression subtractive hybridization: a method for generating differentially regulated or tissue-specific cDNA probes and libraries." Proc Natl Acad Sci U S A **93**(12): 6025-30.

Dietrich, F. S., S. Voegeli, et al. (2004). "The Ashbya gossypii genome as a tool for mapping the ancient Saccharomyces cerevisiae genome." Science **304**(5668): 304-7.

Ditzel, L., J. Lowe, et al. (1998). "Crystal structure of the thermosome, the archaeal chaperonin and homolog of CCT." Cell **93**(1): 125-38.

Do, C. B., M. S. Mahabhashyam, et al. (2005). "ProbCons: Probabilistic consistency-based multiple sequence alignment." Genome Res **15**(2): 330-40.

Doktycz, M. J., R. F. Goldstein, et al. (1992). "Studies of DNA dumbbells. I. Melting curves of 17 DNA dumbbells with different duplex stem sequences linked by T4 endloops: evaluation of the nearest-neighbor stacking interactions in DNA." Biopolymers **32**(7): 849-64.

Doktycz, M. J., M. D. Morris, et al. (1995). "Optical melting of 128 octamer DNA duplexes. Effects of base pair location and nearest neighbors on thermal stability." J Biol Chem **270**(15): 8439-45.

Dominguez, R. (2004). "Actin-binding proteins--a unifying hypothesis." Trends Biochem Sci **29**(11): 572-8.

dos Remedios, C. G., D. Chhabra, et al. (2003). "Actin binding proteins: regulation of cytoskeletal microfilaments." Physiol Rev **83**(2): 433-73.

Douglas, S., S. Zauner, et al. (2001). "The highly reduced genome of an enslaved algal nucleus." Nature **410**(6832): 1091-6.

Dujon, B., D. Sherman, et al. (2004). "Genome evolution in yeasts." Nature **430**(6995): 35-44.

Dunham, I., N. Shimizu, et al. (1999). "The DNA sequence of human chromosome 22." Nature **402**(6761): 489-95.

Eckley, D. M., S. R. Gill, et al. (1999). "Analysis of dynactin subcomplexes reveals a novel actin-related protein associated with the arp1 minifilament pointed end." J Cell Biol **147**(2): 307-20.

Eckley, D. M. and T. A. Schroer (2003). "Interactions between the evolutionarily conserved, actin-related protein, Arp11, actin, and Arp1." Mol Biol Cell **14**(7): 2645-54.

Egea, G., F. Lazaro-Dieguez, et al. (2006). "Actin dynamics at the Golgi complex in mammalian cells." Curr Opin Cell Biol **18**(2): 168-78.

Ehrhardt, D. W. and S. L. Shaw (2006). "Microtubule dynamics and organization in the plant cortical array." Annu Rev Plant Biol **57**: 859-75.

Eichinger, L., J. A. Pachebat, et al. (2005). "The genome of the social amoeba Dictyostelium discoideum." Nature **435**(7038): 43-57.

Eisenberg, D., E. M. Marcotte, et al. (2000). "Protein function in the post-genomic era." Nature **405**(6788): 823-6.

El-Sayed, N. M., P. J. Myler, et al. (2005). "The genome sequence of Trypanosoma cruzi, etiologic agent of Chagas disease." Science **309**(5733): 409-15.

Enright, A. J., I. Iliopoulos, et al. (1999). "Protein interaction maps for complete genomes based on gene fusion events." Nature **402**(6757): 86-90.

Etzold, T. and P. Argos (1993). "SRS--an indexing and retrieval tool for flat file data libraries." Comput Appl Biosci **9**(1): 49-57.

Faix, J. and R. Grosse (2006). "Staying in shape with formins." Dev Cell **10**(6): 693-706.

Farag, T. I. and A. S. Teebi (1988). "Bardet-Biedl and Laurence-Moon syndromes in a mixed Arab population." Clin Genet **33**(2): 78-82.

Farag, T. I. and A. S. Teebi (1989). "High incidence of Bardet Biedl syndrome among the Bedouin." Clin Genet **36**(6): 463-4.

Fitch, W. M. (1970). "Distinguishing homologous from analogous proteins." Syst Zool **19**(2): 99-113.

Fitch, W. M. and E. Margoliash (1967). "Construction of phylogenetic trees." Science **155**(760): 279-84.

Fleischmann, R. D., M. D. Adams, et al. (1995). "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd." Science **269**(5223): 496-512.

Fraser, C. M., S. Casjens, et al. (1997). "Genomic sequence of a Lyme disease spirochaete, Borrelia burgdorferi." Nature **390**(6660): 580-6.

Freeman, W. M., D. J. Robertson, et al. (2000). "Fundamentals of DNA hybridization arrays for gene expression analysis." Biotechniques **29**(5): 1042-6, 1048-55.

Friederich, E., C. Huet, et al. (1989). "Villin induces microvilli growth and actin redistribution in transfected fibroblasts." Cell **59**(3): 461-75.

Frixione, E. (2000). "Recurring views on the structure and function of the cytoskeleton: a 300-year epic." Cell Motil Cytoskeleton **46**(2): 73-94.

Fuchs, E. and D. W. Cleveland (1998). "A structural scaffolding of intermediate filaments in health and disease." Science **279**(5350): 514-9.

Fuchs, M., J. Gerber, et al. (2001). "The p400 complex is an essential E1A transformation target." Cell **106**(3): 297-307.

Galagan, J. E., S. E. Calvo, et al. (2003). "The genome sequence of the filamentous fungus Neurospora crassa." Nature **422**(6934): 859-68.

Galarneau, L., A. Nourani, et al. (2000). "Multiple links between the NuA4 histone acetyltransferase complex and epigenetic control of transcription." Cell **5**(6): 927-37.

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

Galtier, N., M. Gouy, et al. (1996). "SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny." Comput Appl Biosci **12**(6): 543-8.

Gao, Y., J. O. Thomas, et al. (1992). "A cytoplasmic chaperonin that catalyzes beta-actin folding." Cell **69**(6): 1043-50.

Gardner, M. J., R. Bishop, et al. (2005). "Genome sequence of Theileria parva, a bovine pathogen that transforms lymphocytes." Science **309**(5731): 134-7.

Gardner, M. J., N. Hall, et al. (2002). "Genome sequence of the human malaria parasite Plasmodium falciparum." Nature **419**(6906): 498-511.

Gerdes, J. M. and N. Katsanis (2005). "Microtubule transport defects in neurological and ciliary disease." Cell Mol Life Sci **62**(14): 1556-70.

Gerhold, D., T. Rushmore, et al. (1999). "DNA chips: promising toys have become powerful tools." Trends Biochem Sci **24**(5): 168-73.

Gherman, A., E. E. Davis, et al. (2006). "The ciliary proteome database: an integrated community resource for the genetic and functional dissection of cilia." Nat Genet **38**(9): 961-2.

Gibbs, R. A., G. M. Weinstock, et al. (2004). "Genome sequence of the Brown Norway rat yields insights into mammalian evolution." Nature **428**(6982): 493-521.

Gibson, F., J. Walsh, et al. (1995). "A type VII myosin encoded by the mouse deafness gene shaker-1." Nature **374**(6517): 62-4.

Giganti, A., Friederich, E. (2003). "The actin cytoskeleton as a therapeutic target : state of the art and future directions." Progress in cell cycle research edited by L. Reiziger, A. Jezequel and M.Roberge: 511-25.

Goff, S. A., D. Ricke, et al. (2002). "A draft sequence of the rice genome (Oryza sativa L. ssp. japonica)." Science **296**(5565): 92-100.

Goffeau, A., B. G. Barrell, et al. (1996). "Life with 6000 genes." Science **274**(5287): 546, 563-7.

Goldman, R. D., Y. H. Chou, et al. (1999). "Intermediate filaments: dynamic processes regulating their assembly, motility, and interactions with other cytoskeletal systems." Faseb J **13 Suppl 2**: S261-5.

Goldman, R. D., Y. Gruenbaum, et al. (2002). "Nuclear lamins: building blocks of nuclear architecture." Genes Dev **16**(5): 533-47.

Goldman, R. D., S. Khuon, et al. (1996). "The function of intermediate filaments in cell shape and cytoskeletal integrity." J Cell Biol **134**(4): 971-83.

Goley, E. D. and M. D. Welch (2006). "The ARP2/3 complex: an actin nucleator comes of age." Nat Rev Mol Cell Biol **7**(10): 713-26.

Goodner, B., G. Hinkle, et al. (2001). "Genome sequence of the plant pathogen and biotechnology agent Agrobacterium tumefaciens C58." Science **294**(5550): 2323-8.

Goodson, H. V. and W. F. Hawse (2002). "Molecular evolution of the actin family." J Cell Sci **115**(Pt 13): 2619-22.

Gordon, J. L. and L. D. Sibley (2005). "Comparative genome analysis reveals a conserved family of actin-like proteins in apicomplexan parasites." BMC Genomics **6**: 179.

Gorzer, I., C. Schuller, et al. (2003). "The nuclear actin-related protein Act3p/Arp4p of Saccharomyces cerevisiae is involved in transcription regulation of stress genes." Mol Microbiol **50**(4): 1155-71.

Gotoh, O. and Y. Tagashira (1981). "Locations of frequently opening regions on natural DNAs and their relation to functional loci." Biopolymers **20**(5): 1043-58.

Gouy, M., F. Milleret, et al. (1984). "ACNUC: a nucleic acid sequence data base and analysis system." Nucleic Acids Res **12**(1 Pt 1): 121-7.

Grava, S., P. Dumoulin, et al. (2000). "Functional analysis of six genes from chromosomes XIV and XV of Saccharomyces cerevisiae reveals YOR145c as an essential gene and YNL059c/ARP5 as a strain-dependent essential gene encoding nuclear proteins." Yeast **16**(11): 1025-33.

Griffin, T. J. and L. M. Smith (1998). "An approach to predicting the stabilities of peptide nucleic acid:DNA duplexes." Anal Biochem **260**(1): 56-63.

Halliburton, W. D. (1887). "On Muscle-Plasma." J Physiol **8**(3-4): 133-202.

Hamm, G. H. and G. N. Cameron (1986). "The EMBL data library." Nucleic Acids Res **14**(1): 5-9.

Hamosh, A., A. F. Scott, et al. (2005). "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders." Nucleic Acids Res **33**(Database issue): D514-7.

Harata, M., K. Nishimori, et al. (2001). "Identification of two cDNAs for human actin-related proteins (Arps) that have remarkable similarity to conventional actin." Biochim Biophys Acta **1522**(2): 130-3.

Harata, M., Y. Oma, et al. (2000). "Multiple actin-related proteins of Saccharomyces cerevisiae are present in the nucleus." J Biochem (Tokyo) **128**(4): 665-71.

Hardison, R. C. (2003). "Comparative genomics." PLoS Biol **1**(2): E58.

Hartman, H. and A. Fedorov (2002). "The origin of the eukaryotic cell: a genomic investigation." Proc Natl Acad Sci U S A **99**(3): 1420-5.

Heidelberg, J. F., J. A. Eisen, et al. (2000). "DNA sequence of both chromosomes of the cholera pathogen Vibrio cholerae." Nature **406**(6795): 477-83.

Helfand, B. T., L. Chang, et al. (2004). "Intermediate filaments are dynamic and motile elements of cellular architecture." J Cell Sci **117**(Pt 2): 133-41.

Higgins, D. G. and P. M. Sharp (1988). "CLUSTAL: a package for performing multiple sequence alignment on a microcomputer." Gene **73**(1): 237-44.

Hildebrandt, F. and E. Otto (2005). "Cilia and centrosomes: a unifying pathogenic concept for cystic kidney disease?" Nat Rev Genet **6**(12): 928-40.

Hillier, L. W., W. Miller, et al. (2004). "Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution." Nature **432**(7018): 695-716.

Hirokawa, N. (1998). "Kinesin and dynein superfamily proteins and the mechanism of organelle transport." Science **279**(5350): 519-26.

Holloway, A. J., R. K. van Laar, et al. (2002). "Options available--from start to finish--for obtaining data from DNA microarrays II." Nat Genet **32 Suppl**: 481-9.

Holmes, K. C., C. Sander, et al. (1993). "A new ATP-binding fold in actin, hexokinase and Hsc70." Trends Cell Biol **3**(2): 53-9.

Holt, R. A., G. M. Subramanian, et al. (2002). "The genome sequence of the malaria mosquito Anopheles gambiae." Science **298**(5591): 129-49.

Howard, J. and A. A. Hyman (2003). "Dynamics and mechanics of the microtubule plus end." Nature **422**(6933): 753-8.

Howley, P. M., M. A. Israel, et al. (1979). "A rapid method for detecting and mapping homology between heterologous DNAs. Evaluation of polyomavirus genomes." J Biol Chem **254**(11): 4876-83.

Hughes, T. R., M. Mao, et al. (2001). "Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer." Nat Biotechnol **19**(4): 342-7.

Huynen, M., T. Dandekar, et al. (1998). "Differential genome analysis applied to the species-specific features of Helicobacter pylori." FEBS Lett **426**(1): 1-5.

Huynen, M. A., Y. Diaz-Lazcoz, et al. (1997). "Differential genome display." Trends Genet **13**(10): 389-90.

Ikura, T., V. V. Ogryzko, et al. (2000). "Involvement of the TIP60 histone acetylase complex in DNA repair and apoptosis." Cell **102**(4): 463-73.

Inglis, P. N., K. A. Boroevich, et al. (2006). "Piecing together a ciliome." Trends Genet **22**(9): 491-500.

Ishikawa, H., R. Bischoff, et al. (1968). "Mitosis and intermediate-sized filaments in developing skeletal muscle." J Cell Biol **38**(3): 538-55.

Ivens, A. C., C. S. Peacock, et al. (2005). "The genome of the kinetoplastid parasite, Leishmania major." Science **309**(5733): 436-42.

Jacob, F. and J. Monod (1961). "Genetic regulatory mechanisms in the synthesis of proteins." J Mol Biol **3**: 318-56.

Jacq, C., J. R. Miller, et al. (1977). "A pseudogene structure in 5S DNA of Xenopus laevis." Cell **12**(1): 109-20.

Jaillon, O., J. M. Aury, et al. (2004). "Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype." Nature **431**(7011): 946-57.

Janmey, P. A. and C. Chaponnier (1995). "Medical aspects of the actin cytoskeleton." Curr Opin Cell Biol **7**(1): 111-7.

Johnson, J. M., J. Castle, et al. (2003). "Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays." Science **302**(5653): 2141-4.

Jones, L. J., R. Carballido-Lopez, et al. (2001). "Control of cell shape in bacteria: helical, actin-like filaments in Bacillus subtilis." Cell **104**(6): 913-22.

Kabsch, W. and K. C. Holmes (1995). "The actin fold." Faseb J **9**(2): 167-74.

Kampa, D., J. Cheng, et al. (2004). "Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22." Genome Res **14**(3): 331-42.

Kane, M. D., T. A. Jatkoe, et al. (2000). "Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays." Nucleic Acids Res **28**(22): 4552-7.

Katinka, M. D., S. Duprat, et al. (2001). "Genome sequence and gene compaction of the eukaryote parasite Encephalitozoon cuniculi." Nature **414**(6862): 450-3.

Katsanis, N., P. L. Beales, et al. (2000). "Mutations in MKKS cause obesity, retinal dystrophy and renal malformations associated with Bardet-Biedl syndrome." Nat Genet **26**(1): 67-70.

Kenney, D., L. Cairns, et al. (1986). "Morphological abnormalities in the lymphocytes of patients with the Wiskott-Aldrich syndrome." Blood **68**(6): 1329-32.

Kent, W. J. (2002). "BLAT--the BLAST-like alignment tool." Genome Res **12**(4): 656-64.

Kent, W. J., C. W. Sugnet, et al. (2002). "The human genome browser at UCSC." Genome Res **12**(6): 996-1006.

Khaitlina, S. Y. (2001). "Functional specificity of actin isoforms." Int Rev Cytol **202**: 35-98.

Kim, J. C., J. L. Badano, et al. (2004). "The Bardet-Biedl protein BBS4 targets cargo to the pericentriolar region and is required for microtubule anchoring and cell cycle progression." Nat Genet **36**(5): 462-70.

Kim, J. C., Y. Y. Ou, et al. (2005). "MKKS/BBS6, a divergent chaperonin-like protein linked to the obesity disorder Bardet-Biedl syndrome, is a novel centrosomal component required for cytokinesis." J Cell Sci **118**(Pt 5): 1007-20.

Kitagawa, H., R. Fujiki, et al. (2003). "The chromatin-remodeling complex WINAC targets a nuclear receptor to promoters and is impaired in Williams syndrome." Cell **113**(7): 905-17.

Kothapalli, R., S. J. Yoder, et al. (2002). "Microarray results: how accurate are they?" BMC Bioinformatics **3**: 22.

Kouranov, A., L. Xie, et al. (2006). "The RCSB PDB information portal for structural genomics." Nucleic Acids Res **34**(Database issue): D302-5.

Kubota, H., G. Hynes, et al. (1994). "Identification of six Tcp-1-related genes encoding divergent subunits of the TCP-1-containing chaperonin." Curr Biol **4**(2): 89-99.

Kuribara, S., M. Kato, et al. (2006). "Identification of a novel actin-related protein in Tetrahymena cilia." Cell Motil Cytoskeleton **63**(7): 437-46.

Kuroda, Y., Y. Oma, et al. (2002). "Brain-specific expression of the nuclear actin-related protein ArpNalpha and its involvement in mammalian SWI/SNF chromatin remodeling complex." Biochem Biophys Res Commun **299**(2): 328-34.

Lambrechts, A., M. Van Troys, et al. (2004). "The actin cytoskeleton in normal and pathological cell motility." Int J Biochem Cell Biol **36**(10): 1890-909.

Lander, E. S. and D. Botstein (1987). "Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children." Science **236**(4808): 1567-70.

Lander, E. S., L. M. Linton, et al. (2001). "Initial sequencing and analysis of the human genome." Nature **409**(6822): 860-921.

Laurier, V., C. Stoetzel, et al. (2006). "Pitfalls of homozygosity mapping: an extended consanguineous Bardet-Biedl syndrome family with two mutant genes (BBS2, BBS10), three mutations, but no triallelism." Eur J Hum Genet.

Lecompte, O., R. Ripp, et al. (2001). "Genome evolution at the genus level: comparison of three complete genomes of hyperthermophilic archaea." Genome Res **11**(6): 981-93.

Lecompte, O., J. D. Thompson, et al. (2001). "Multiple alignment of complete sequences (MACS) in the post-genomic era." Gene **270**(1-2): 17-30.

Lee, J. M., S. Dedhar, et al. (2006). "The epithelial-mesenchymal transition: new insights in signaling, development, and disease." J Cell Biol **172**(7): 973-81.

Lee, S., B. C. Lee, et al. (2006). "Prediction of protein secondary structure content using amino acid composition and evolutionary information." Proteins **62**(4): 1107-14.

Li, F. and G. D. Stormo (2001). "Selection of optimal DNA oligos for gene expression arrays." Bioinformatics **17**(11): 1067-76.

Li, J. B., J. M. Gerdes, et al. (2004). "Comparative genomics identifies a flagellar and basal body proteome that includes the BBS5 human disease gene." Cell **117**(4): 541-52.

Liolios, K., N. Tavernarakis, et al. (2006). "The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide." Nucleic Acids Res **34**(Database issue): D332-4.

Lipman, D. J. and W. R. Pearson (1985). "Rapid and sensitive protein similarity searches." Science **227**(4693): 1435-41.

Lipshutz, R. J., S. P. Fodor, et al. (1999). "High density synthetic oligonucleotide arrays." Nat Genet **21**(1 Suppl): 20-4.

Loftus, B., I. Anderson, et al. (2005). "The genome of the protist parasite Entamoeba histolytica." Nature **433**(7028): 865-8.

Loftus, B. J., E. Fung, et al. (2005). "The genome of the basidiomycetous yeast and human pathogen Cryptococcus neoformans." Science **307**(5713): 1321-4.

Lowe, J., F. van den Ent, et al. (2004). "Molecules of the bacterial cytoskeleton." Annu Rev Biophys Biomol Struct **33**: 177-98.

Luebke, K. J., R. P. Balog, et al. (2003). "Prioritized selection of oligodeoxyribonucleotide probes for efficient hybridization to RNA transcripts." Nucleic Acids Res **31**(2): 750-8.

Machesky, L. M., S. J. Atkinson, et al. (1994). "Purification of a cortical complex containing two unconventional actins from Acanthamoeba by affinity chromatography on profilin-agarose." J Cell Biol **127**(1): 107-15.

Machesky, L. M. and R. C. May (2001). "Arps: actin-related proteins." Results Probl Cell Differ **32**: 213-29.

Machesky, L. M. and M. Schliwa (2000). "Cell dynamics: a new look at the cytoskeleton." Nat Cell Biol **2**(1): E17-8.

Magnusson, N. E., A. K. Cardozo, et al. (2005). "Construction and validation of the APOCHIP, a spotted oligo-microarray for the study of beta-cell apoptosis." BMC Bioinformatics **6**: 311.

Maitra, A., Y. Cohen, et al. (2004). "The Human MitoChip: a high-throughput sequencing microarray for mitochondrial mutation detection." Genome Res **14**(5): 812-9.

Manton, I., K. Kowallik, et al. (1970). "Observations on the fine structure and development of the spindle at mitosis and meiosis in a marine centric diatom (Lithodesmium undulatum). IV. The second meiotic division and conclusion." J Cell Sci **7**(2): 407-43.

Marcotte, E. M. (2000). "Computational genetics: finding protein function by nonhomology methods." Curr Opin Struct Biol **10**(3): 359-65.

Marcotte, E. M., M. Pellegrini, et al. (1999). "Detecting protein function and protein-protein interactions from genome sequences." Science **285**(5428): 751-3.

Martys, J. L., C. L. Ho, et al. (1999). "Intermediate filaments in motion: observations of intermediate filaments in cells using green fluorescent protein-vimentin." Mol Biol Cell **10**(5): 1289-95.

Matsuzaki, M., O. Misumi, et al. (2004). "Genome sequence of the ultrasmall unicellular red alga Cyanidioschyzon merolae 10D." Nature **428**(6983): 653-7.

Matthaei, J. H., O. W. Jones, et al. (1962). "Characteristics and composition of RNA coding units." Proc Natl Acad Sci U S A **48**: 666-77.

McKinney, E. C., M. K. Kandasamy, et al. (2002). "Arabidopsis contains ancient classes of differentially expressed actin-related protein genes." Plant Physiol **128**(3): 997-1007.

McKusick, V. A. and F. H. Ruddle (1987). "Toward a complete map of the human genome." Genomics **1**(2): 103-6.

Meinkoth, J. and G. Wahl (1984). "Hybridization of nucleic acids immobilized on solid supports." Anal Biochem **138**(2): 267-84.

Mermall, V., P. L. Post, et al. (1998). "Unconventional myosins in cell movement, membrane traffic, and signal transduction." Science **279**(5350): 527-33.

Mighell, A. J., N. R. Smith, et al. (2000). "Vertebrate pseudogenes." FEBS Lett **468**(2-3): 109-14.

Mikesell, P., B. E. Ivins, et al. (1983). "Evidence for plasmid-mediated toxin production in Bacillus anthracis." Infect Immun **39**(1): 371-6.

Miralles, F. and N. Visa (2006). "Actin in transcription and transcription regulation." Curr Opin Cell Biol **18**(3): 261-6.

Mitchison, T. and M. Kirschner (1984). "Dynamic instability of microtubule growth." Nature **312**(5991): 237-42.

Mizuguchi, G., X. Shen, et al. (2004). "ATP-driven exchange of histone H2AZ variant catalyzed by SWR1 chromatin remodeling complex." Science **303**(5656): 343-8.

Modrek, B. and C. Lee (2002). "A genomic view of alternative splicing." Nat Genet **30**(1): 13-9.

Mohrmann, L., K. Langenberg, et al. (2004). "Differential targeting of two distinct SWI/SNF-related Drosophila chromatin-remodeling complexes." Mol Cell Biol **24**(8): 3077-88.

Moir, R. D. and T. P. Spann (2001). "The structure and function of nuclear lamins: implications for disease." Cell Mol Life Sci **58**(12-13): 1748-57.

Moore, S. J., J. S. Green, et al. (2005). "Clinical and genetic epidemiology of Bardet-Biedl syndrome in Newfoundland: a 22-year prospective, population-based, cohort study." Am J Med Genet A **132**(4): 352-60.

Mukherjee, A., K. Dai, et al. (1993). "Escherichia coli cell division protein FtsZ is a guanine nucleotide binding protein." Proc Natl Acad Sci U S A **90**(3): 1053-7.

Muller, J., Y. Oma, et al. (2005). "Sequence and Comparative Genomic Analysis of Actin-related Proteins." Mol Biol Cell.

Mykytyn, K., T. Braun, et al. (2001). "Identification of the gene that, when mutated, causes the human obesity syndrome BBS4." Nat Genet **28**(2): 188-91.

Mykytyn, K., R. F. Mullins, et al. (2004). "Bardet-Biedl syndrome type 4 (BBS4)-null mice implicate Bbs4 in flagella formation but not global cilia assembly." Proc Natl Acad Sci U S A **101**(23): 8664-9.

Mykytyn, K., D. Y. Nishimura, et al. (2003). "Evaluation of complex inheritance involving the most common Bardet-Biedl syndrome locus (BBS1)." Am J Hum Genet **72**(2): 429-37.

Mykytyn, K., D. Y. Nishimura, et al. (2002). "Identification of the gene (BBS1) most commonly involved in Bardet-Biedl syndrome, a complex human obesity syndrome." Nat Genet **31**(4): 435-8.

Needleman, S. B. and C. D. Wunsch (1970). "A general method applicable to the search for similarities in the amino acid sequence of two proteins." J Mol Biol **48**(3): 443-53.

Nelson, W. J. (2003). "Adaptation of core mechanisms to generate cell polarity." Nature **422**(6933): 766-74.

Nie, Z., Y. Xue, et al. (2000). "A specificity and targeting subunit of a human SWI/SNF family-related chromatin-remodeling complex." Mol Cell Biol **20**(23): 8879-88.

Nielsen, H. B., R. Wernersson, et al. (2003). "Design of oligonucleotides for microarrays and perspectives for design of multi-transcriptome arrays." Nucleic Acids Res **31**(13): 3491-6.

Nierman, W. C., A. Pain, et al. (2005). "Genomic sequence of the pathogenic and allergenic filamentous fungus Aspergillus fumigatus." Nature **438**(7071): 1151-6.

Nishimura, D. Y., C. C. Searby, et al. (2001). "Positional cloning of a novel gene on chromosome 16q causing Bardet-Biedl syndrome (BBS2)." Hum Mol Genet **10**(8): 865-74.

Nishimura, D. Y., R. E. Swiderski, et al. (2005). "Comparative genomics and gene expression analysis identifies BBS9, a new Bardet-Biedl syndrome gene." Am J Hum Genet **77**(6): 1021-33.

Nogales, E., K. H. Downing, et al. (1998). "Tubulin and FtsZ form a distinct family of GTPases." Nat Struct Biol **5**(6): 451-8.

Nolen, B. J., R. S. Littlefield, et al. (2004). "Crystal structures of actin-related protein 2/3 complex with bound ATP or ADP." Proc Natl Acad Sci U S A **101**(44): 15627-32.

Nonaka, S., H. Shiratori, et al. (2002). "Determination of left-right patterning of the mouse embryo by artificial nodal flow." Nature **418**(6893): 96-9.

Nordberg, E. K. (2005). "YODA: selecting signature oligonucleotides." Bioinformatics **21**(8): 1365-70.

Okubo, K., H. Sugawara, et al. (2006). "DDBJ in preparation for overview of research activities behind data submissions." Nucleic Acids Res **34**(Database issue): D6-9.

Olave, I. A., S. L. Reck-Peterson, et al. (2002). "Nuclear actin and actin-related proteins in chromatin remodeling." Annu Rev Biochem **71**: 755-81.

Ou, Y. and J. B. Rattner (2004). "The centrosome in higher organisms: structure, composition, and duplication." Int Rev Cytol **238**: 119-82.

Overbeek, R., M. Fonstein, et al. (1999). "The use of gene clusters to infer functional coupling." Proc Natl Acad Sci U S A **96**(6): 2896-901.

Page, R. D. (1996). "TreeView: an application to display phylogenetic trees on personal computers." Comput Appl Biosci **12**(4): 357-8.

Papoulas, O., S. J. Beek, et al. (1998). "The Drosophila trithorax group proteins BRM, ASH1 and ASH2 are subunits of distinct protein complexes." Development **125**(20): 3955-66.

Pazour, G. J. and J. L. Rosenbaum (2002). "Intraflagellar transport and cilia-dependent diseases." Trends Cell Biol **12**(12): 551-5.

Pederson, T. and U. Aebi (2002). "Actin in the nucleus: what form and what for?" J Struct Biol **140**(1-3): 3-9.

Pederson, T. and U. Aebi (2005). "Nuclear actin extends, with no contraction in sight." Mol Biol Cell **16**(11): 5055-60.

Pellegrini, M., E. M. Marcotte, et al. (1999). "Assigning protein functions by comparative genome analysis: protein phylogenetic profiles." Proc Natl Acad Sci U S A **96**(8): 4285-8.

Peterson, C. L., Y. Zhao, et al. (1998). "Subunits of the yeast SWI/SNF complex are members of the actin-related protein (ARP) family." J Biol Chem **273**(37): 23641-4.

Peterson, M. G. (1988). "DNA sequencing using Taq polymerase." Nucleic Acids Res **16**(22): 10915.

Petit, C. (2001). "Usher syndrome: from genetics to pathogenesis." Annu Rev Genomics Hum Genet **2**: 271-97.

Phillips, A., D. Janies, et al. (2000). "Multiple sequence alignment in phylogenetic analysis." Mol Phylogenet Evol **16**(3): 317-30.

Plewniak, F., L. Bianchetti, et al. (2003). "PipeAlign: A new toolkit for protein family analysis." Nucleic Acids Res **31**(13): 3829-32.

Poch, O. and B. Winsor (1997). "Who's who among the Saccharomyces cerevisiae actin-related proteins? A classification and nomenclature proposal for a large family." Yeast **13**(11): 1053-8.

Pollak, M. R., Y. H. Chou, et al. (1993). "Homozygosity mapping of the gene for alkaptonuria to chromosome 3q2." Nat Genet **5**(2): 201-4.

Pollard, T. D. (2003). "The cytoskeleton, cellular motility and the reductionist agenda." Nature **422**(6933): 741-5.

Pollard, T. D. and G. G. Borisy (2003). "Cellular motility driven by assembly and disassembly of actin filaments." Cell **112**(4): 453-65.

Pollard, T. D. and J. A. Cooper (1986). "Actin and actin-binding proteins. A critical evaluation of mechanisms and functions." Annu Rev Biochem **55**: 987-1035.

Praetorius, H. A. and K. R. Spring (2001). "Bending the MDCK cell primary cilium increases intracellular calcium." J Membr Biol **184**(1): 71-9.

Prigent, V., J. C. Thierry, et al. (2005). "DbW: automatic update of a functional family-specific multiple alignment." Bioinformatics **21**(8): 1437-42.

Prince, V. E. and F. B. Pickett (2002). "Splitting pairs: the diverging fates of duplicated genes." Nat Rev Genet **3**(11): 827-37.

Pruitt, K. D., T. Tatusova, et al. (2005). "NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins." Nucleic Acids Res **33 Database Issue**: D501-4.

Quinlan, M. E., J. E. Heuser, et al. (2005). "Drosophila Spire is an actin nucleation factor." Nature **433**(7024): 382-8.

Rao, M. V., L. J. Engle, et al. (2002). "Myosin Va binding to neurofilaments is essential for correct myosin Va distribution and transport and neurofilament density." J Cell Biol **159**(2): 279-90.

Revenu, C., R. Athman, et al. (2004). "The co-workers of actin filaments: from cell structures to signals." Nat Rev Mol Cell Biol **5**(8): 635-46.

Reymond, N., H. Charles, et al. (2004). "ROSO: optimizing oligonucleotide probes for microarrays." Bioinformatics **20**(2): 271-3.

Rice, P., I. Longden, et al. (2000). "EMBOSS: the European Molecular Biology Open Software Suite." Trends Genet **16**(6): 276-7.

Rieder, C. L., S. Faruki, et al. (2001). "The centrosome in vertebrates: more than a microtubule-organizing center." Trends Cell Biol **11**(10): 413-9.

Rimour, S., D. Hill, et al. (2005). "GoArrays: highly dynamic and efficient microarray probe design." Bioinformatics **21**(7): 1094-103.

Robinson, R. C., K. Turbedsky, et al. (2001). "Crystal structure of Arp2/3 complex." Science **294**(5547): 1679-84.

Rodal, A. A., O. Sokolova, et al. (2005). "Conformational changes in the Arp2/3 complex leading to actin nucleation." Nat Struct Mol Biol **12**(1): 26-31.

Rosenbaum, J. L. and G. B. Witman (2002). "Intraflagellar transport." Nat Rev Mol Cell Biol **3**(11): 813-25.

Ross, A. J., H. May-Simera, et al. (2005). "Disruption of Bardet-Biedl syndrome ciliary proteins perturbs planar cell polarity in vertebrates." Nat Genet **37**(10): 1135-40.

Rouillard, J. M., C. J. Herbert, et al. (2002). "OligoArray: genome-scale oligonucleotide design for microarrays." Bioinformatics **18**(3): 486-7.

Rouillard, J. M., M. Zuker, et al. (2003). "OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach." Nucleic Acids Res **31**(12): 3057-62.

Salzberg, S. L., A. L. Delcher, et al. (1998). "Microbial gene identification using interpolated Markov models." Nucleic Acids Res **26**(2): 544-8.

Sanger, F., A. R. Coulson, et al. (1978). "The nucleotide sequence of bacteriophage phiX174." J Mol Biol **125**(2): 225-46.

Sanger, F., S. Nicklen, et al. (1977). "DNA sequencing with chain-terminating inhibitors." Proc Natl Acad Sci U S A **74**(12): 5463-7.

Sanger, F. and E. O. Thompson (1953). "The amino-acid sequence in the glycyl chain of insulin. I. The identification of lower peptides from partial hydrolysates." Biochem J **53**(3): 353-66.

Sanger, F. and E. O. Thompson (1953). "The amino-acid sequence in the glycyl chain of insulin. II. The investigation of peptides from enzymic hydrolysates." Biochem J **53**(3): 366-74.

SantaLucia, J., Jr. (1998). "A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics." Proc Natl Acad Sci U S A **95**(4): 1460-5.

SantaLucia, J., Jr., H. T. Allawi, et al. (1996). "Improved nearest-neighbor parameters for predicting DNA duplex stability." Biochemistry **35**(11): 3555-62.

Sathananthan, A. H., W. D. Ratnasooriya, et al. (2006). "Rediscovering Boveri's centrosome in Ascaris (1888): its impact on human fertility and development." Reprod Biomed Online **12**(2): 254-70.

Schafer, D. A. and T. A. Schroer (1999). "Actin-related proteins." Annu Rev Cell Dev Biol **15**: 341-63.

Schliwa, M. and G. Woehlke (2003). "Molecular motors." Nature **422**(6933): 759-65.

Schnackenberg, B. J. and R. E. Palazzo (1999). "Identification and function of the centrosome centromatrix." Biol Cell **91**(6): 429-38.

Schneider, M. E., I. A. Belyantseva, et al. (2002). "Rapid renewal of auditory hair bundles." Nature **418**(6900): 837-8.

Scholey, J. M., I. Brust-Mascher, et al. (2003). "Cell division." Nature **422**(6933): 746-52.

Schroer, T. A., E. Fyrberg, et al. (1994). "Actin-related protein nomenclature and classification." J Cell Biol **127**(6 Pt 2): 1777-8.

Sekerkova, G., L. Zheng, et al. (2004). "Espins are multifunctional actin cytoskeletal regulatory proteins in the microvilli of chemosensory and mechanosensory cells." J Neurosci **24**(23): 5445-56.

Sekerkova, G., L. Zheng, et al. (2006). "Espins and the actin cytoskeleton of hair cell stereocilia and sensory cell microvilli." Cell Mol Life Sci.

Selinger, D. W., K. J. Cheung, et al. (2000). "RNA expression analysis using a 30 base pair resolution Escherichia coli genome array." Nat Biotechnol **18**(12): 1262-8.

Shaw, S. L., R. Kamyar, et al. (2003). "Sustained microtubule treadmilling in Arabidopsis cortical arrays." Science **300**(5626): 1715-8.

Shen, X., G. Mizuguchi, et al. (2000). "A chromatin remodelling complex involved in transcription and DNA processing." Nature **406**(6795): 541-4.

Shen, X., R. Ranallo, et al. (2003). "Involvement of actin-related proteins in ATP-dependent chromatin remodeling." Mol Cell **12**(1): 147-55.

Shibata, Y., G. K. Voeltz, et al. (2006). "Rough sheets and smooth tubules." Cell **126**(3): 435-9.

Shih, Y. L. and L. Rothfield (2006). "The bacterial cytoskeleton." Microbiol Mol Biol Rev **70**(3): 729-54.

Silflow, C. D. and P. A. Lefebvre (2001). "Assembly and motility of eukaryotic cilia and flagella. Lessons from Chlamydomonas reinhardtii." Plant Physiol **127**(4): 1500-7.

Slautterback, D. B. (1963). "Cytoplasmic Microtubules. I. Hydra." J Cell Biol **18**: 367-88.

Slavotinek, A. M., E. M. Stone, et al. (2000). "Mutations in MKKS cause Bardet-Biedl syndrome." Nat Genet **26**(1): 15-6.

Small, J. V., T. Stradal, et al. (2002). "The lamellipodium: where motility begins." Trends Cell Biol **12**(3): 112-20.

Smith, T. F. and M. S. Waterman (1981). "Identification of common molecular subsequences." J Mol Biol **147**(1): 195-7.

Snell, W. J., J. Pan, et al. (2004). "Cilia and flagella revealed: from flagellar assembly in Chlamydomonas to human obesity disorders." Cell **117**(6): 693-7.

Snitkin, E. S., A. M. Gustafson, et al. (2006). "Comparative assessment of performance and genome dependence among phylogenetic profiling methods." BMC Bioinformatics **7**: 420.

Southern, E. M. (1974). "An improved method for transferring nucleotides from electrophoresis strips to thin layers of ion-exchange cellulose." Anal Biochem **62**(1): 317-8.

Southern, E. M. (1975). "Detection of specific sequences among DNA fragments separated by gel electrophoresis." J Mol Biol **98**(3): 503-17.

Spiess, C., A. S. Meyer, et al. (2004). "Mechanism of the eukaryotic chaperonin: protein folding in the chamber of secrets." Trends Cell Biol **14**(11): 598-604.

Stein, L. D., Z. Bao, et al. (2003). "The genome sequence of Caenorhabditis briggsae: a platform for comparative genomics." PLoS Biol **1**(2): E45.

Sternlicht, H., G. W. Farr, et al. (1993). "The t-complex polypeptide 1 complex is a chaperonin for tubulin and actin in vivo." Proc Natl Acad Sci U S A **90**(20): 9422-6.

Stidwill, R. P., T. Wysolmerski, et al. (1984). "The brush border cytoskeleton is not static: in vivo turnover of proteins." J Cell Biol **98**(2): 641-5.

Stoetzel, C., V. Laurier, et al. (2006). "BBS10 encodes a vertebrate-specific chaperonin-like protein and is a major BBS locus." Nat Genet **38**(5): 521-4.

Stone, D. L., A. Slavotinek, et al. (2000). "Mutation of a gene encoding a putative chaperonin causes McKusick-Kaufman syndrome." Nat Genet **25**(1): 79-82.

Stover, C. K., X. Q. Pham, et al. (2000). "Complete genome sequence of Pseudomonas aeruginosa PA01, an opportunistic pathogen." Nature **406**(6799): 959-64.

Sugimoto, N., S. Nakano, et al. (1996). "Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes." Nucleic Acids Res **24**(22): 4501-5.

Svitkina, T. M. and G. G. Borisy (1999). "Arp2/3 complex and actin depolymerizing factor/cofilin in dendritic organization and treadmilling of actin filament array in lamellipodia." J Cell Biol **145**(5): 1009-26.

Svitkina, T. M., A. B. Verkhovsky, et al. (1996). "Plectin sidearms mediate interaction of intermediate filaments with microtubules and other components of the cytoskeleton." J Cell Biol **135**(4): 991-1007.

Szerlong, H., A. Saha, et al. (2003). "The nuclear actin-related proteins Arp7 and Arp9: a dimeric module that cooperates with architectural proteins for chromatin remodeling." Embo J **22**(12): 3175-87.

Tabin, C. J. (2006). "The key to left-right asymmetry." Cell **127**(1): 27-32.

Talla, E., F. Tekaia, et al. (2003). "A novel design of whole-genome microarray probes for Saccharomyces cerevisiae which minimizes cross-hybridization." BMC Genomics **4**(1): 38.

Tamas, I., L. Klasson, et al. (2002). "50 million years of genomic stasis in endosymbiotic bacteria." Science **296**(5577): 2376-9.

Tateno, Y., S. Miyazaki, et al. (2000). "DNA data bank of Japan (DDBJ) in collaboration with mass sequencing teams." Nucleic Acids Res **28**(1): 24-6.

Thiery, J. P. (2002). "Epithelial-mesenchymal transitions in tumour progression." Nat Rev Cancer **2**(6): 442-54.

Thiery, J. P. and J. P. Sleeman (2006). "Complex networks orchestrate epithelial-mesenchymal transitions." Nat Rev Mol Cell Biol **7**(2): 131-42.

Thompson, J. D., A. Muller, et al. (2006). "MACSIMS: multiple alignment of complete sequences information management system." BMC Bioinformatics **7**: 318.

Tilney, L. G., M. S. Tilney, et al. (1992). "Actin filaments, stereocilia, and hair cells: how cells count and measure." Annu Rev Cell Biol **8**: 257-74.

Tilney, L. G., M. S. Tilney, et al. (1995). "F actin bundles in Drosophila bristles. I. Two filament cross-links are involved in bundling." J Cell Biol **130**(3): 629-38.

Tilney, M. S., L. G. Tilney, et al. (1989). "Preliminary biochemical characterization of the stereocilia and cuticular plate of hair cells of the chick cochlea." J Cell Biol **109**(4 Pt 1): 1711-23.

Torrents, D., M. Suyama, et al. (2003). "A genome-wide survey of human pseudogenes." Genome Res **13**(12): 2559-67.

Tsoka, S. and C. A. Ouzounis (2000). "Recent developments and future directions in computational genomics." FEBS Lett **480**(1): 42-8.

Tyska, M. J. and M. S. Mooseker (2002). "MYO1A (brush border myosin I) dynamics in the brush border of LLC-PK1-CL4 cells." Biophys J **82**(4): 1869-83.

Valpuesta, J. M., J. Martin-Benito, et al. (2002). "Structure and function of a protein folding machine: the eukaryotic cytosolic chaperonin CCT." FEBS Lett **529**(1): 11-6.

van den Ent, F., L. A. Amos, et al. (2001). "Prokaryotic origin of the actin cytoskeleton." Nature **413**(6851): 39-44.

van den Ent, F., J. Moller-Jensen, et al. (2002). "F-actin-like filaments formed by plasmid segregation protein ParM." Embo J **21**(24): 6935-43.

Van Sluys, M. A., M. C. de Oliveira, et al. (2003). "Comparative analyses of the complete genome sequences of Pierce's disease and citrus variegated chlorosis strains of Xylella fastidiosa." J Bacteriol **185**(3): 1018-26.

Van Troys, M., J. Vandekerckhove, et al. (1999). "Structural modules in actin-binding proteins: towards a new classification." Biochim Biophys Acta **1448**(3): 323-48.

Vanin, E. F. (1985). "Processed pseudogenes: characteristics and evolution." Annu Rev Genet **19**: 253-72.

Venter, J. C., M. D. Adams, et al. (2001). "The sequence of the human genome." Science **291**(5507): 1304-51.

Volkmann, N., K. J. Amann, et al. (2001). "Structure of Arp2/3 complex in its activated state and in actin filament branch junctions." Science **293**(5539): 2456-9.

Vologodskii, A. V., B. R. Amirikyan, et al. (1984). "Allowance for heterogeneous stacking in the DNA helix-coil transition theory." J Biomol Struct Dyn **2**(1): 131-48.

von Ahsen, N., M. Oellerich, et al. (1999). "Application of a thermodynamic nearest-neighbor model to estimate nucleic acid stability and optimize probe design: prediction of melting points of multiple mutations of apolipoprotein B-3500 and factor V with a hybridization probe genotyping assay on the LightCycler." Clin Chem **45**(12): 2094-101.

Wallace, R. B., J. Shaffer, et al. (1979). "Hybridization of synthetic oligodeoxyribonucleotides to phi chi 174 DNA: the effect of single base pair mismatch." Nucleic Acids Res **6**(11): 3543-57.

Walsh, S. and B. Barrell (1996). "The Saccharomyces cerevisiae genome on the World Wide Web." Trends Genet **12**(7): 276-7.

Walter, G., K. Bussow, et al. (2000). "Protein arrays for gene expression and molecular interaction screening." Curr Opin Microbiol **3**(3): 298-302.

Wang, H. Y., R. L. Malek, et al. (2003). "Assessing unmodified 70-mer oligonucleotide probe performance on glass-slide microarrays." Genome Biol **4**(1): R5.

Waters, E., M. J. Hohn, et al. (2003). "The genome of Nanoarchaeum equitans: insights into early archaeal evolution and derived parasitism." Proc Natl Acad Sci U S A **100**(22): 12984-8.

Waterston, R. H., K. Lindblad-Toh, et al. (2002). "Initial sequencing and comparative analysis of the mouse genome." Nature **420**(6915): 520-62.

Watson, J. D. and F. H. Crick (1953). "The structure of DNA." Cold Spring Harb Symp Quant Biol **18**: 123-31.

Webber, W. A. and J. Lee (1975). "Fine structure of mammalian renal cilia." Anat Rec **182**(3): 339-43.

Weber, V., M. Harata, et al. (1995). "The actin-related protein Act3p of Saccharomyces cerevisiae is located in the nucleus." Mol Biol Cell **6**(10): 1263-70.

Wegner, A. (1976). "Head to tail polymerization of actin." J Mol Biol **108**(1): 139-50.

Weil, D., S. Blanchard, et al. (1995). "Defective myosin VIIA gene responsible for Usher syndrome type 1B." Nature **374**(6517): 60-1.

Wicker, N., D. Dembele, et al. (2002). "Density of points clustering, application to transcriptomic data analysis." Nucleic Acids Res **30**(18): 3992-4000.

Winder, S. J. and K. R. Ayscough (2005). "Actin-binding proteins." J Cell Sci **118**(Pt 4): 651-4.

Woese, C. R., O. Kandler, et al. (1990). "Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya." Proc Natl Acad Sci U S A **87**(12): 4576-9.

Wood, V., R. Gwilliam, et al. (2002). "The genome sequence of Schizosaccharomyces pombe." Nature **415**(6874): 871-80.

Wootton, J. C. and S. Federhen (1996). "Analysis of compositionally biased regions in sequence databases." Methods Enzymol **266**: 554-71.

Wu, C. H., R. Apweiler, et al. (2006). "The Universal Protein Resource (UniProt): an expanding universe of protein information." Nucleic Acids Res **34**(Database issue): D187-91.

Wu, C. H., L. S. Yeh, et al. (2003). "The Protein Information Resource." Nucleic Acids Res **31**(1): 345-7.

Xu, P., G. Widmer, et al. (2004). "The genome of Cryptosporidium hominis." Nature **431**(7012): 1107-12.

Yang, A. S. and B. Honig (2000). "An integrated approach to the analysis and modeling of protein sequences and structures. III. A comparative study of sequence conservation in protein structural families using multiple structural alignments." J Mol Biol **301**(3): 691-711.

Yates, J. R., 3rd, A. Gilchrist, et al. (2005). "Proteomics of organelles and large cellular structures." Nat Rev Mol Cell Biol **6**(9): 702-14.

Yu, J., S. Hu, et al. (2002). "A draft sequence of the rice genome (Oryza sativa L. ssp. indica)." Science **296**(5565): 79-92.

Zhang, Z. and M. Gerstein (2004). "Large-scale analysis of pseudogenes in the human genome." Curr Opin Genet Dev **14**(4): 328-35.

Zhang, Z., P. M. Harrison, et al. (2003). "Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome." Genome Res **13**(12): 2541-58.

Zhao, K., W. Wang, et al. (1998). "Rapid and phosphoinositol-dependent binding of the SWI/SNF-like BAF complex to chromatin after T lymphocyte receptor signaling." Cell **95**(5): 625-36.

Zheng, L., G. Sekerkova, et al. (2000). "The deaf jerker mouse has a mutation in the gene encoding the espin actin-bundling proteins of hair cell stereocilia and lacks espins." Cell **102**(3): 377-85.

Zhu, B., G. Ping, et al. (2005). "Comparison of gene expression measurements from cDNA and 60-mer oligonucleotide microarrays." Genomics **85**(6): 657-65.

Zmasek, C. M. and S. R. Eddy (2001). "ATV: display and manipulation of annotated phylogenetic trees." Bioinformatics **17**(4): 383-4.

Publications

Publication N°4

Publication N°5

Résumé

Le travail présenté dans cette thèse concerne l'étude du cytosquelette par des techniques de bioinformatique et de biologie à haut débit. Il décrit également le développement de plusieurs outils, à même de traiter un système aussi complexe que le cytosquelette.
La première partie concerne la génomique comparative et plus particulièrement la méthode des profils phylogénétiques. Un outil, ComIcs, a été implémenté et le calcul en automatique des profils phylogénétiques de l'ensemble des gènes du cytosquelette dans 41 organismes eucaryotes a été effectué. Leur analyse a révélé des problèmes majeurs liés aux familles de protéines très conservées au cours de l'évolution et à la couverture imparfaite des protéomes de certains organismes. Certains de ses problèmes ont été adressés par l'étude de la famille des actines et des Actin-Related Proteins, dont la caractérisation profonde a permis le développement d'ARPAnno, un serveur d'annotation et d'identification des séquences proches de l'actine.
Dans le cadre d'une collaboration avec le laboratoire de Diagnostique Médical, l'application de cette approche a permis d'identifier, BBS10 et BBS12, deux nouveaux gènes majeurs responsables du syndrome de Bardet Biedl.
La seconde partie aborde la transcriptomique et décrit le développement d'Actichip, une puce à oligonucléotide dédiée aux gènes du cytosquelette. Cet outil d'analyse intégrée du cytosquelette a nécessité le développement de CADO4MI, un logiciel de sélection de sondes spécifiques. La stratégie de validation des sondes spécifiques ainsi qu'une première application d'Actichip sont présentées dans cette partie.

Abstract

The work of my PhD focuses on applications of bioinformatics methodologies and high throughput analysis techniques to study the cytoskeleton. My work also highlights several novel bioinformatics developments allowing the study of highly complex biological systems like the cytoskeleton.
In the first part, comparative genomics and particularly phylogenetic profiling methods are discussed. ComIcs, a new tool, was developed to automatically establish the phylogenetic profiles for the complete set of cytoskeleton genes in 41 eukaryotic organisms. Results revealed several major limitations of the method linked either to highly similar protein families or the lack of complete proteomes for some organisms. Some of these issues were addressed by an in depth analysis of actin and the Actin-Related Proteins family. This led to the implementation of ARPAnno, a web server dedicated to the identification of protein sequences similar to actin.
In the second part, I described the development of a new dedicated microarray, named Actichip, to monitor the expression profiles of cytoskeleton genes. The development of this dedicated tool required the implementation of CADO4MI, a new program for the design of specific oligonucleotide probes for microarrays. The strategy and a first application of Actichip are presented in this part.

Mots clés/Keywords

Cytosquelette/cytoskeleton, génomique comparative/comparative genomics, Actin-Related Protein, analyse de séquence/sequence analysis, syndrome de Bardet-Biedl/Bardet-Biedl Syndrom, transcriptomique/transcriptomics, puces à ADN/microarray.

Laboratoires/Laboratories

Laboratoire de Bioinformatique et de Génomique Intégratives (UMR7104), IGBMC, 1 rue Laurent Fries, 67404 Illkirch-Graffenstaden cedex, France.
Laboratoire de Biologie Moléculaire, d'Analyse Génique et de Modélisation, Centre de Recherche Public-Santé du Luxembourg, 84 Val Fleuri, L-1526 Luxembourg, Luxembourg.