Caractérisation et analyse évolutive des répétitions intragéniques : une étude au niveau des gènes, des séquences protéiques et des structures tridimensionnelles

Anne-Laure Abraham

THESE DE DOCTORAT DE
L'UNIVERSITE PIERRE ET MARIE CURIE

Spécialité

Analyse des Génomes et Modélisation Moléculaire

Ecole doctorale Logique du Vivant

Présentée par

Anne-Laure Abraham

Pour obtenir le grade de

DOCTEUR de l'UNIVERSITÉ PIERRE ET MARIE CURIE

Caractérisation et analyse évolutive des répétitions intragéniques : une étude au niveau des gènes, des séquences protéiques et des structures tridimensionnelles

Soutenue le 15 décembre 2008
Devant le jury composé de :

Dr. Daniel KAHN	Rapporteur
Dr. Jean-François GIBRAT	Rapporteur
Dr. Marie-France SAGOT	Examinatrice
Pr. Pierre NETTER	Examinateur
Dr. Eduardo ROCHA	Directeur de thèse
Dr. Joël POTHIER	Directeur de thèse

Université Pierre & Marie Curie - Paris 6
Bureau d'accueil, inscription des doctorants et base de données
Esc G, 2ème étage
15 rue de l'école de médecine
75270-PARIS CEDEX 06

Tél. Secrétariat : 01 42 34 68 35
Fax : 01 42 34 68 40
Tél. pour les étudiants de A à EL : 01 42 34 69 54
Tél. pour les étudiants de EM à ME : 01 42 34 68 41
Tél. pour les étudiants de MF à Z : 01 42 34 68 51
E-mail : scolarite.doctorat@upmc.fr

Résumé

Titre : Caractérisation et analyse évolutive des répétitions intragéniques : une étude au niveau des gènes, des séquences protéiques et des structures tridimensionnelles

Résumé : Les duplications jouent un rôle important dans l'évolution des protéines et sont à l'origine des répétitions intragéniques présentes dans environ 14% des séquences protéiques. Nous avons choisi d'étudier ces répétitions d'un point de vue évolutif. Pour cela, nous avons développé un programme, Swelfe, qui cherche les répétitions à la fois dans les gènes, les séquences d'acides aminés et les structures tridimensionnelles des protéines. Ce programme utilise le même algorithme de programmation dynamique à tous les niveaux et une représentation séquentielle des structures 3D. Les scores et les tests de significativité des répétitions obtenues ont été adaptés pour chaque niveau. Nous avons créé une banque contenant les séquences d'ADN et d'acides aminés correspondant aux structures de la PDB, et comparé Swelfe à DALI pour valider la méthode au niveau des répétitions structurales. Enfin, ce programme est disponible pour la communauté à l'adresse http://bioserv.rpbs.jussieu.fr/swelfe et peut être téléchargé ou utilisé en ligne.

Swelfe a trouvé un nombre important de répétitions dans un ensemble non redondant de séquences nucléiques, séquences protéiques et structures tridimensionnelles, et environ 10% des protéines contiennent des répétitions à au moins un niveau. Cependant, le recouvrement des répétitions aux trois niveaux est assez faible et beaucoup de répétitions ne sont trouvées qu'à un seul niveau, ce qui confirme l'intérêt de cette étude sur les trois niveaux en parallèle Les causes de ce recouvrement faible sont discutées dans la présente thèse. L'étude des répétitions structurales longues montre qu'environ 30% de ces répétitions sont pseudo-symétriques à 180°, comme le sont les deux éléments d'un homo-dimère. L'analyse de ces protéines indique que certaines pourraient effectivement remplacer des dimères.

Mots clés : répétitions, gènes, séquences protéiques, structures, évolution, structure quaternaire.

Title : Characterisation and evolutionary analysis of intragenic repeats : a study in genes, protein sequences and three-dimensional structures of proteins

Abstract : Duplications play a major role in protein evolution and result in intragenic repeats found in about 14% of protein sequences. We chose to study these repeats from an evolutionary point of view. For this, we developed a program, Swelfe, that looks for repeats in genes, amino acid sequences and three dimensional structures of proteins. This program uses the same dynamic programming algorithm at all levels and a sequential representation of 3D structures. Repeat scores and significance tests are adapted at each level. We created a bank containing DNA sequences and amino acid sequences corresponding to PDB structures, and Swelfe was compared to DALI to validate our method concerning structural repeats. Finally this program is available for the community at http://bioserv.rpbs.jussieu.fr/swelfe and can be downloaded or used online.

Swelfe found an important number of repeats in a non redundant data set of DNA sequences, amino acid sequences and 3D structures and about 10% of proteins contain repeats at last at one level. However the repeat overlap at the three level is weak and most repeats are only found at one level, this confirm the relevance of studying repeats at three levels at the same time. Reasons of the weak overlap are discussed in this thesis. The study of long structural repeats shows that about 30% of these repeats are symmetrical at 180°, as are the two elements of a homodimer. The analysis of these proteins shows that some of them could effectively replace homodimers.

Key words : repeats, genes, protein sequences, structures, evolution, quaternary structure.

Cette thèse a été préparée à l'Atelier de BioInformatique, Maison de la Pédagogie, bâtiment C RdC, Boite courrier 1202, 4 place Jussieu, 75252 Paris cedex 05

Sommaire

Table des figures

Table des tableaux

Abréviations et acronymes

3D : Trois Dimensions

ADN : Acide DésoxyriboNucléique

ADNc : Acide DésoxyriboNucléique complémentaire

AFP : Aligned Fragment Pair

ARN : Acide RiboNucléique

ARNm : Acide RiboNucléique messager

CATH : Class Architecture Topology Homology

EST : Expressed Sequence Tag

HMM : Hidden Markov Models

LUCA : Last Universal Common Ancestor

LTR : Long Terminal Repeat

PDB : Protein Data Bank

PCR : Polymerase Chain Reaction

PSSM : Position-Specific Scoring Matrix,

PQS : Protein Quaternary Structure

SCOP : Structural Classification of Proteins

SHSP : Structural High Scoring Pair

SSE : Secondary Structure Element

RMSD : Root Mean Square Deviation

RRMSD : Relative Root Mean Square Deviation

SSR : Simple Sequence Repeat

TPR : Tetratrico Peptide Repeat

TrEMBL : Translated EMBL

VNTR : Variable Number of Tandem Repeat

Introduction générale

Le plus récent ancêtre universel commun à tous les êtres vivants, LUCA (Last Universal Common Ancestor), a vécu il y a environ 4 milliards d'années. Aujourd'hui, entre 5 et 30 millions d'espèces peuplent la planète d'après l'institut français de la biodiversité. Entre temps, les organismes vivants ont évolué et présentent une diversité étonnante. Ils se sont adaptés à des conditions très variables de température, pression, ressources naturelles, etc. Cette diversité est inscrite dans l'ADN de ces organismes. Cette molécule contient les informations pour synthétiser les protéines nécessaires au fonctionnement des cellules des êtres vivants : structure, catalyse de réactions chimiques, transport, communication, signalisation etc.

L'évolution du génome des organismes est guidée par deux contraintes contradictoires : stabilité et variabilité. La conservation du génome est possible grâce à un mécanisme précis de copie de l'ADN et à des systèmes de réparation des lésions accidentelles. Inversement, les variations progressives du génome vont modifier l'organisme et certains changements lui permettent de mieux s'adapter à un environnement qui évolue en permanence.

Plusieurs types de modifications peuvent affecter l'ADN : mutation ponctuelle, insertion ou délétion de paires de bases, duplication, recombinaison, réarrangement de fragments d'ADN. Tous ces changements modifient le génome de façon aléatoire, et un certain nombre ont lieu dans des régions non codantes du génome. Cependant, d'autres sont situés dans des gènes. Ils vont se répercuter sur la séquence d'acides aminés puis sur le repliement de la protéine. Finalement, sa fonction peut également être altérée. En général, les modifications n'affectent pas, ou peu, la structure et la fonction. En effet, le code génétique est redondant, et plusieurs codons de l'ADN codent pour le même acide aminé. De plus, certaines substitutions d'acides aminés modifient peu le repliement. Cependant, d'autres modifications peuvent avoir des conséquences importantes sur la protéine. Certaines sont délétères et peuvent aboutir à la perte de la fonction de la protéine, ce qui peut entraîner une diminution de l'espérance de vie de l'organisme, voire sa mort. D'autres modifications peuvent entraîner une amélioration de la fonction ou l'acquisition d'une nouvelle fonction avantageuse pour l'organisme, ce qui augmente la probabilité qu'elle soit conservée au cours des générations. Ainsi, les modifications les moins avantageuses sont moins fréquemment transmises, et les modifications avantageuses sont conservées en moyenne plus longtemps dans l'espèce : c'est le

modèle néo-darwiniste proposé par Fisher, Haldane et Wright (Fisher, 1930 ; Haldane, 1932 ; Wright, 1931)

L'objet de cette thèse est l'étude des duplications qui ont lieu à l'intérieur des gènes. La majorité des répétitions créées sont délétères pour la protéine puisque la probabilité de synthétiser une protéine non fonctionnelle est élevé : cela peut entraîner un décalage du cadre de lecture de la traduction ou modifier de façon importante le repliement. Cependant, un certain nombre des répétitions peuvent être observées et ont donc été conservées par les organismes. Plusieurs questions se posent : combien y a-t-il de répétitions intragéniques ? Pourquoi ces répétitions sont-elles conservées ? Apportent-elles un avantage à l'organisme ? Comment la structure des protéines s'adapte-t-elle à ces répétitions ? Y a-t-il des conséquences sur la fonction de la protéine ?

Plusieurs études ont essayé de mieux comprendre ces répétitions, en les étudiant au niveau des séquences nucléiques ou protéiques, des structures, ou des domaines. Elles ont permis de mettre en évidence qu'environ 14% des séquences protéiques contiennent des répétitions (Marcotte et al., 1999b). Les répétitions sont plus fréquentes chez les eucaryotes que chez les procaryotes, et plus fréquentes chez les pluricellulaires que chez les unicellulaires (Lavorgna et al., 2001). Il existe également des protéines contenant de nombreux éléments répétés, et dont le nombre de répétition évolue rapidement (Andrade et al., 2001). Elles sont particulièrement impliquées dans des fonctions de liaison à l'ADN ou à d'autres protéines. Les répétitions internes peuvent poser des problèmes de repliement pour les protéines, et sont donc évitées.

A notre connaissance, aucune étude n'a encore été faite à la fois au niveau des séquences nucléiques, des séquences protéiques et des structures tridimensionnelle des protéines. Cette approche est intéressante dans la mesure où ces trois niveaux n'évoluent pas à la même vitesse mais sont intimement liés : les événements de duplication ont lieu dans les gènes mais ont des conséquences sur la séquence protéique et sur la structure des protéines. Les séquences d'ADN évoluent plus vite que les séquences protéiques : du fait de la redondance du code génétique, certaines mutations n'ont pas de conséquence sur la chaîne d'acides aminés. Les séquences protéiques

évoluent plus vite que les structures : certains changements d'acides aminés aboutissent à la même structure, et les contraintes fonctionnelles sur les structures vont défavoriser les mutations trop délétères. L'analyse des répétitions à ces trois niveaux devrait donc permettre de trouver des résultats différents des analyses faites à un seul niveau, et d'apporter des données évolutives sur ces répétitions.

Le premier chapitre de cette thèse est une introduction générale sur l'évolution des protéines, et sur l'étude des protéines contenant des répétitions. Cette présentation est faite au niveau des gènes, des séquences d'acides aminés, des structures tertiaires et quaternaires des protéines, ainsi qu'au niveau des domaines. Ensuite, les objectifs sont présentés dans la deuxième partie.

La troisième partie présente Swelfe, le programme que nous avons conçu pour identifier les répétitions dans les séquences d'ADN, d'acides aminés et dans les structures tridimensionnelles des protéines. Tout d'abord, plusieurs algorithmes de comparaison de séquences ou de structures existant sont présentés, puis l'algorithme de programmation dynamique utilisé pour Swelfe est détaillé. La validation du programme ainsi que les scores et les tests de significativité des répétitions obtenues sont également présentés dans cette partie, ainsi que la création de la banque contenant les séquences d'ADN et d'acides aminés correspondant aux structures.

La quatrième partie concerne les répétitions trouvées par Swelfe dans les séquences et les structures d'un ensemble non redondant de protéines. Les résultats obtenus montrent qu'un grand nombre de répétitions a été trouvé aux trois niveaux mais que les répétitions trouvées sont assez différentes. Dans cette partie, nous avons essayé d'expliquer ces observations.

La cinquième partie est consacrée aux protéines contenant des répétitions dont les copies sont symétriques à 180°. Environ 30% des protéines contenant des répétitions longues présentent une telle symétrie. Les protéines contenant ces répétitions ressemblent à des homodimères formés d'une seule protéine. Nous les avons analysées afin de les quantifier et comprendre leur rôle. Il pourrait s'agir d'un mécanisme évolutif qui permet de remplacer certains dimères par une seule protéine ayant une symétrie interne. Un modèle est proposé pour l'évolution des structures grâce à une répétition pseudo-symétrique.

Enfin, une conclusion et des perspectives clôtureront ce document.

I Les duplications, acteurs de l'évolution des protéines

I.A L'évolution des protéines : du gène à la structure quaternaire

La séquence nucléique des gènes détermine la séquence d'acides aminés de la protéine issue de ce gène, séquence qui elle-même détermine la structure tridimensionnelle. Les modifications qui peuvent survenir au niveau des gènes vont donc être répercutées sur les protéines au niveau de leurs structures primaire, secondaire, tertiaire et quaternaire (assemblage des monomères). Si ces modifications ont des conséquences délétères sur la fonction de la protéine, elles seront contre sélectionnées. L'évolution des gènes résulte d'un équilibre entre d'une part des processus génétiques qui vont créer des modification dans l'ADN, et d'autre part par une pression de sélection sur les fonctions des protéines ou des ARNs qui va contre sélectionner les mutations délétères. Les répétitions dans les ARNs ne seront pas du tout abordées ici. L'évolution des protéines est très étudiée au niveau des domaines, qui peuvent être considérés comme des unités d'évolution à part entière, et seront traités à la fin de cette partie.

I.A.1 La protéine

I.A.1.a Le gène

L'acide désoxyribonucléique (ADN) est le support de l'information génétique. Il contient toute l'information permettant la synthèse des protéines nécessaires à la cellule et est transmis en totalité ou en partie aux cellules filles lorsque la cellule divise (Figure 1).

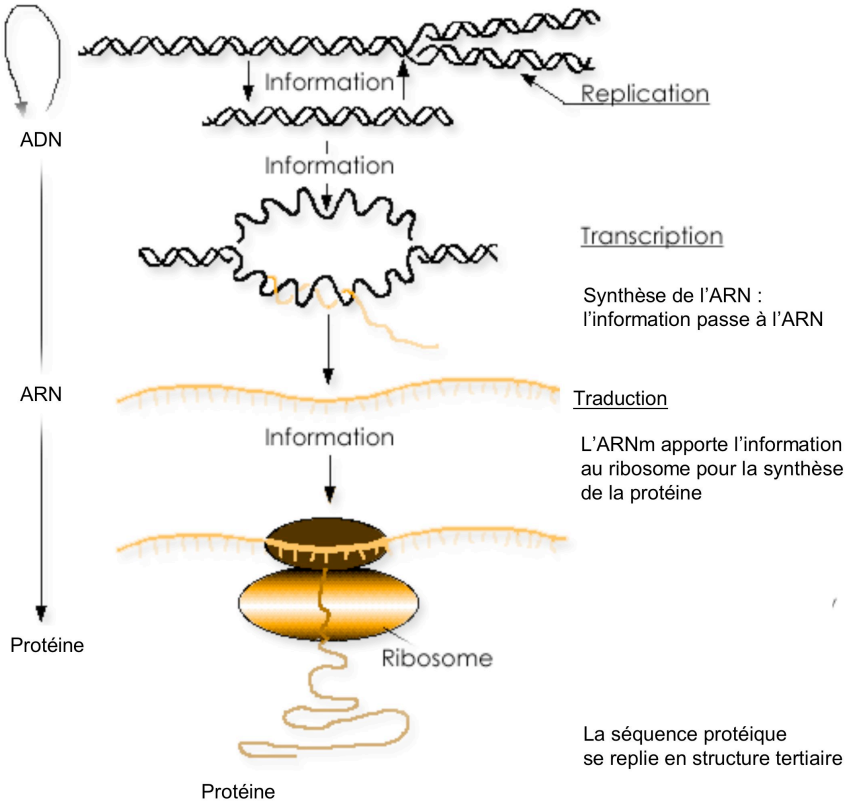


Figure 1 : Réplication de l'ADN et synthèse des protéines.
(tiré de http://www.ict-science-to-society.org/Pathogenomics/genome.aspx).

Un gène est une séquence d'ADN, composée des nucléotides adénine (A), cytosine (C), guanine (G), et thymine (T), et qui code pour un ARN (acide ribonucléique). S'il s'agit d'un ARN messager (ARNm), il contient l'information pour synthétiser une protéine. La plupart du temps, le gène est précédé d'une séquence promotrice qui permet d'initier et de réguler la transcription de l'ADN en ARN. Certains gènes sont constitués de séquences codantes (exons) et de séquences non codantes (introns) qui seront supprimées lors de l'épissage, avant la synthèse de la protéine.

Au cours des générations, plusieurs modifications peuvent affecter la molécule d'ADN. La plupart surviennent lors de la réplication, il peut s'agir de mutations[1], insertions ou délétions de base(s). Il peut également survenir des duplications au sein des chromosomes ou des duplications de chromosomes entiers, ces mécanismes créent des répétitions dans l'ADN et seront détaillés ultérieurement. Chez les eucaryotes, des

[1] Une mutation est une substitution ponctuelle d'un nucléotide sur la séquence d'ADN, une insertion est l'ajout d'un ou plusieurs nucléotide(s) et une délétion est la perte d'un ou plusieurs nucléotide(s).

événements de recombinaison inégaux entre chromosomes peuvent survenir : les deux chromosomes échangent un fragment de chromosome ; si la recombinaison est inégale, les deux fragments de chromosomes n'ont pas la même taille, ce qui entraîne le gain de matériel génétique pour un chromosome et sa perte pour l'autre. Chez les procaryotes, des processus de recombinaisons peuvent avoir lieu entre le génome et un ADN étranger, ou à l'intérieur d'un chromosome. Les réarrangements[2] sont particulièrement fréquents aux endroits du génome riches en répétitions (les « breakpoints ») (Bourque et al., 2005). De plus, dans certaines populations bactériennes, il a été observé qu'en condition de stress, la mutagénèse augmente (Tenaillon et al., 2004). La mutagenèse induite par le stress pourrait être le résultat d'une sélection car elle engendre des mutations potentiellement bénéfiques et elle pourrait jouer un rôle important dans l'évolution des bactéries. De façon intéressante, la perte de certains mécanismes de réparation entraîne une augmentation de la recombinaison entre les répétitions.

[2] Les réarrangements sont des modifications du génome à large échelle telle que des inversions ou transposition de segments d'ADN, des translocations entre chromosomes non homologues, des fusions ou fissions de chromosomes, des délétions ou duplications de petites ou grandes portions de génome.

I.A.1.b La séquence d'acides aminés, la structure primaire

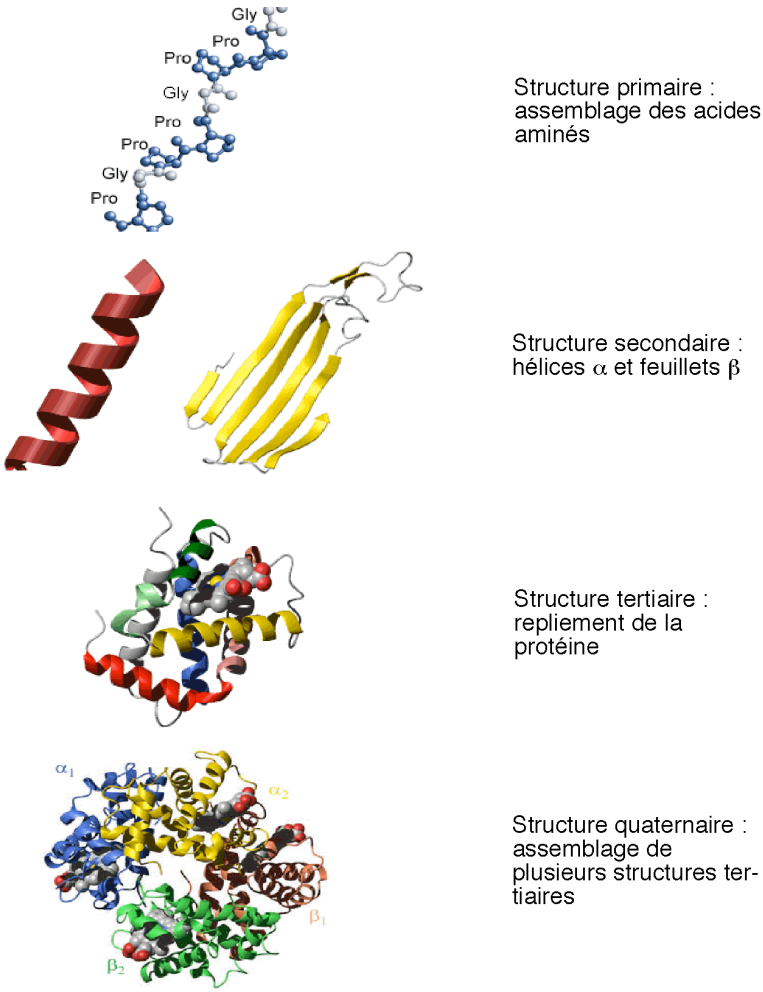


Figure 2 : De la structure primaire à la structure quaternaire des protéines.
(D'après http://www.ac-orleans-tours.fr/svt/mol3d/3d/module3/html/3page.htm).

Le gène est transcrit en ARNm, puis cet ARNm est lu par groupes de trois nucléotides, ou codon, par le ribosome. Chaque codon correspond à un acide aminé. La séquence primaire est l'enchaînement des acides aminés (Figure 3) qui sont assemblés de l'extrémité N-terminale (à gauche sur la Figure 3) à l'extrémité C-terminale de la chaîne (à droite sur la Figure 3). Les acides aminés sont liés entre eux par des liaisons peptidiques. Les chaînes latérales des acides aminés ont des propriétés physico-chimiques (Figure 4) qui sont à l'origine de la diversité fonctionnelle des protéines. Ces chaînes latérales ainsi que les atomes du squelette peptidique vont déterminer le repliement de la protéine.

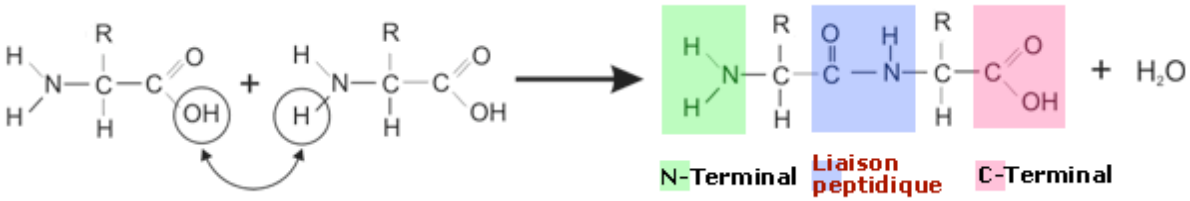


Figure 3 : La liaison peptidique.
(Issu de http://www.azaquar.com/iaa/chimie/ca_images/ca_proteine3.gif).

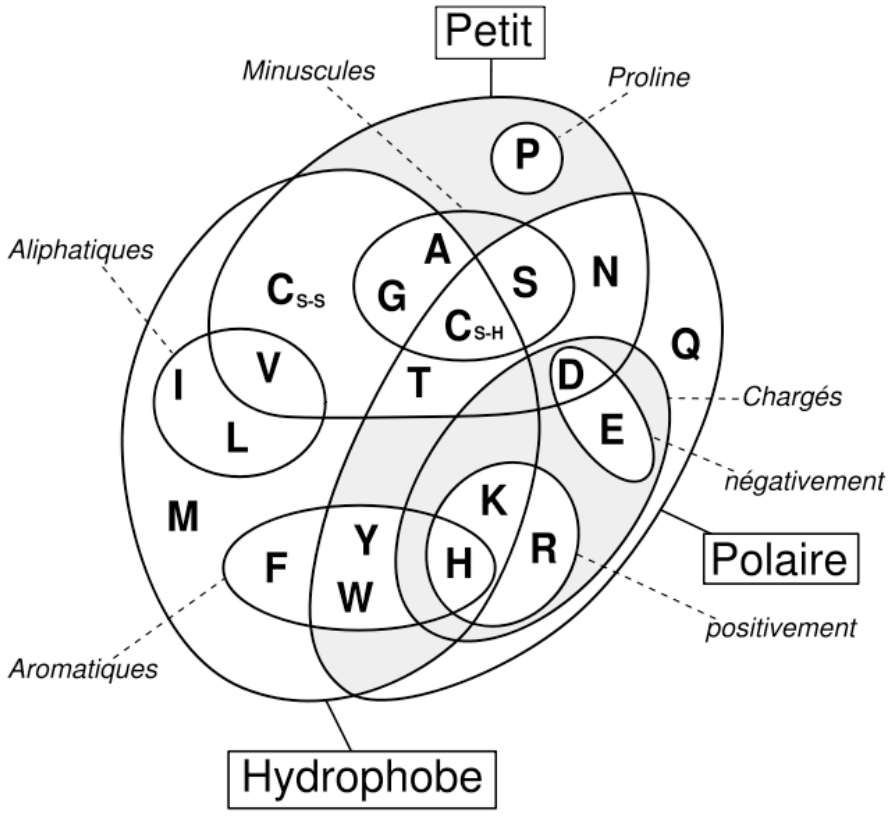


Figure 4 : Diagramme de Venn des propriétés des acides aminés.
D'après (Taylor, 1986).

Lorsque les mutations, insertions et délétions surviennent dans un gène, elles vont avoir des conséquences sur la protéine pour laquelle celui-ci code. Certaines mutations peuvent avoir des effets restreints : grâce à la redondance du code génétique, la plupart des acides aminés sont codés par plusieurs codons, et il est donc possible que la mutation ne modifie pas l'acide aminé, ou que celui-ci soit remplacé par un acide aminé proche (même polarité, même encombrement stérique par exemple). Par contre, certaines mutations peuvent modifier le codon d'initiation ou de terminaison de la transcription ou de la traduction ou les sites d'épissage, ce qui entraîne des modifications importantes de la protéine. Les insertions ou délétions peuvent modifier le cadre de lecture, ce qui modifie généralement les acides aminés codés et la position

du codon stop. De plus, des modifications du site d'initiation ou de régulation de la transcription peuvent modifier l'expression de la protéine. Si les mutations sont trop importantes, la protéine pourra perdre sa fonction et si cette protéine est utile ou vitale, la mutation sera contre sélectionnée.

I.A.1.c La structure secondaire

La structure secondaire désigne la conformation adoptée par la chaîne peptidique au niveau local. Les deux principales structures régulières observées sont les hélices α et les feuillets β (Figure 2) (Pauling and Corey, 1951 ; Pauling et al., 1951). L'hélice α est une hélice droite stabilisée par des liaisons hydrogènes entre le résidu i et le résidu i+4, et a une périodicité de 3,6 résidus par tour d'hélice. Elle concerne environ 30% des résidus (Martin, 2005). Le feuillet β est stabilisé par des liaisons hydrogènes entre des résidus éloignés le long de la séquence, dans des portions de chaîne en conformation étendue, les brins β. Ils concernent environ 20% des résidus protéiques. Les structures apériodiques, ou boucles sont par définition ni α, ni β. Elles contiennent quelques motifs réguliers comme le coude β, l'hélice 3-10, ou l'hélice π, et des structures non régulières (« random coil »).

I.A.1.d La structure tridimensionnelle ou structure tertiaire

La structure tridimensionnelle est le repliement qu'adopte la chaîne d'acides aminés qui compose la protéine. Des expériences montrent que les protéines repliées aléatoirement sont inactives et qu'après dénaturation modérée, une protéine peut retrouver son repliement natif. Cela indique que la chaîne d'acides aminés contient en général l'information suffisante pour déterminer la conformation de la structure de la protéine (Alberts et al., 1994; Anfinsen, 1973). Les liaisons peptidiques des protéines autorisent la libre rotation des résidus qu'elles relient et permettent donc théoriquement une grande variabilité du squelette peptidique. Cependant, dans les cellules, les protéines vont souvent adopter une seule conformation préférentielle. Le repliement et la stabilité des protéines sont guidés par plusieurs forces : les liaisons hydrogènes, les forces de Van der Waals, les forces électrostatiques et les interactions hydrophobes.

Les liaisons hydrogènes sont formées lorsqu'un atome d'hydrogène est partagé entre un donneur d'hydrogène (l'atome lié covalemment à l'hydrogène) et un accepteur

d'hydrogène. Elles sont présentes par exemple dans la formation des hélices α et des feuillets β. Les forces de Van der Waals sont des forces faibles et à courte portée mais efficaces en grand nombre. Une asymétrie passagère de charge autour d'un atome, créée par un dipôle permanent ou parce que la distribution électronique de charge autour d'un atome varie au cours du temps, induit une asymétrie opposée dans un atome adjacent : ils s'attirent alors mutuellement, mais peuvent se repousser s'ils sont trop proches. La distance à laquelle la force d'attraction et la force de répulsion s'équilibrent est appelée distance de Van der Waals. Les interactions électrostatiques, ou ponts salins, sont formés entre deux atomes de charges opposées. Les interactions hydrophobes sont des forces indirectes qui favorisent le regroupement des molécules non polaires. En effet, lorsque la protéine est déstructurée, les molécules d'eau qui l'entourent vont former des « cages » autour des résidus non polaires, ce qui va ordonner les molécules d'eau et diminuer leur entropie, mais l'entropie de la protéine augmente. Inversement, lorsque la protéine est repliée et que les résidus hydrophobes sont situés à l'intérieur, l'entropie des molécules d'eau augmente, mais l'entropie de la protéine diminue.

Les protéines globulaires se replient en une forme globulaire compacte où les résidus hydrophobes se retrouvent enfouis à l'intérieur de la protéine et les résidus polaires sont à l'extérieur où ils peuvent interagir entre eux et avec les molécules d'eau. Les protéines membranaires se replient différemment : les résidus hydrophobes sont à l'extérieur de la protéine, en contact avec la membrane. Les protéines de surface ou excrétées peuvent contenir des liaisons disulfures entre des résidus cystéines qui stabilisent la protéine. Ces interactions sont défavorisées dans le cytoplasme à cause de la présence du glutathion qui rompt les ponts disulfures.

Le repliement des protéines est rapide, de l'ordre de la milliseconde, et la protéine ne peut pas explorer toutes les conformations possibles (Robson, 1999). Une hypothèse est que les structures secondaires, hélices α et feuilles β, se formeraient rapidement, suivies des interactions longue portée, puis de réarrangements locaux, avant d'aboutir à la conformation finale.

Le nombre de séquences protéiques possibles est presque illimité (théoriquement 20^N, N étant la taille de la protéine), par contre le nombre de repliements différents adoptés par les structures est en nombre limité, et relativement faible. Mille repliements différents suffiraient à décrire la majorité des protéines (Koonin et al., 2002). Ces protéines, qui adoptent le même repliement protéique, proviennent-elles

d'un ancêtre commun ou s'agit-il d'un phénomène de convergence ? Koonin *et al*. ont observé que la plupart du temps, les interactions physicochimiques et des motifs de séquences sont conservés. De plus, il y a plusieurs exemples de protéines qui n'ont pas le même repliement et effectuent la même fonction dans un organisme différent ou dans des cellules différentes du même organisme, ce qui indique qu'il n'est pas nécessaire d'avoir le même repliement pour avoir la même fonction. Ces arguments pencheraient en faveur d'une origine monophylétique. Cependant, certains repliements simples et symétriques tels que les tonneaux TIM ou les hélices β pourraient avoir acquis le même repliement par convergence, car les contraintes physico-chimiques favorisent certaines topologies de chaînes (Chothia and Gerstein, 1997).

I.A.1.e Les contraintes évolutives des protéines à tous les niveaux

Les structures s'adaptent aux modifications créées au niveau de leur gène. Si les modifications des acides aminés ne sont pas trop importantes, les protéines peuvent s'accommoder, comme le montrent quelques exemples ci-dessous. Cependant, dans le cas inverse, ces modifications pourront être contre sélectionnées. En conséquence les changements trop radicaux ne seront pas conservés, sauf s'ils permettent à la protéine d'améliorer sa fonction ou d'en acquérir une nouvelle.

- *Les structures s'adaptent aux petites modifications de séquence*

La structure d'une protéine dépend de la séquence de son gène. Dans quelle mesure l'évolution de la séquence va-t-elle modifier la structure ?

Les protéines homologues ont des régions qui conservent le même repliement et des régions pour lesquelles le repliement diffère. Plusieurs études se sont intéressées aux relations entre la séquence, la structure et la fonction. . Wilson *et al*. (Wilson et al., 2000) ont étudié les domaines SCOP et ont observé que pour une paire de domaines qui ont le même repliement, la fonction est conservée jusqu'à environ 40% d'identité de séquence, et la classe fonctionnelle générale est conservée jusqu'à environ 25% d'identité de séquence.Todd *et al*. (Todd et al., 2001) se sont intéressés aux enzymes de la PDB. Pour les protéines avec un ou plusieurs domaines, la variation de numéro EC est rare au dessus de 40% d'identité de séquence, et au delà de 30% d'identité de

séquence, les 3 premiers chiffres sont prédits avec 90% de réussite. En dessous de 20% d'identité de séquence, les régions avec le même repliement peuvent représenter moins de la moitié de chaque protéine et les fonctions peuvent être très différentes (Todd et al., 2001; Wilson et al., 2000). Chothia et Lesk (Chothia and Lesk, 1986) ont calculé que la proportion de repliement similaire entre deux structures est proportionnelle à leur similarité de séquence :

$$\Delta = 0,40e^{1,87H}$$

Équation 1

H est la fraction de résidus mutés et Δ est le RMSD[3] (Root Mean Square Deviation) entre les carbones α des deux chaînes. Les structures évoluent moins vite que les séquences et gardent donc plus longtemps les traces de leur origine.

Les variations de séquence peuvent entraîner des variations structurales importantes. Ces modifications sont acceptables si elles maintiennent la stabilité de la protéine et n'affectent pas négativement sa fonction. C'est le cas en particulier si ces modifications sont distantes du site actif, c'est à dire du site qui fixe le substrat et catalyse la réaction chimique, ou proches du site actif mais couplées de façon à conserver la fonction (Chothia and Gerstein, 1997). Il est aussi nécessaire que ces modifications ne modifient pas la stabilité de la protéine ou sa propension à s'agréger (DePristo et al., 2005).

Gassner *et al*. (Gassner et al., 1996) ont testé l'effet de plusieurs mutations du lysozyme du bactériophage T4 sur son activité. Ce lysozyme a un site actif dans un sillon entre deux domaines. En ne mutant qu'un seul résidu, l'activité du lysozyme est légèrement modifiée, mais à partir de sept résidus mutés, l'activité est réduite de 43%, ce qui indique que plus de la moitié de l'activité est conservée. La différence globale de RMSD entre le fragment muté et le fragment d'origine n'est que de 0.2 Å, et la protéine s'est adaptée à plusieurs des positions mutées pour garder une forme similaire. Chothia et Gerstein (Chothia and Gerstein, 1997), quant à eux, ont remarqué plusieurs changements locaux significatifs dans les atomes de la chaîne principale dans les régions où les mutations ont eu lieu, bien que la forme globale soit la même. De plus, les six leucines mutées ont été remplacées par six méthionines qui ont un encombrement stérique similaire. Donc grâce à plusieurs changements subtils de

[3] Le RMSD est une mesure de comparaison de structures, calculé après superposition optimale des deux structures. Il sera détaillé ultérieurement.

conformation, le lysozyme peut s'adapter à certaines mutations et garder un repliement qui conserve une grande partie de son activité.

Une autre expérience a été réalisée en remplaçant treize résidus du coeur hydrophobe de la barnase par d'autres résidus hydrophobes. Axe *et al*. (Axe et al., 1996) ont montré qu'en ne mutant qu'un résidu à la fois, près d'un quart des mutants avaient une activité enzymatique. Même en mutant les treize sites en même temps, certaines combinaisons de mutations n'empêchent pas quelques mutants d'avoir une activité élevée. Les changements produits sont importants mais spatialement opposés au site actif de la protéine.

La structure d'une protéine peut donc souvent s'adapter pour garder son activité malgré plusieurs mutations, et plusieurs combinaisons d'acides aminés hydrophobes dans le cœur permettent de maintenir son intégrité structurale.

Il est également important que les mutations n'affectent pas la stabilité et l'agrégation de la protéine. La plupart des mutations non synonymes ont des effets physico-chimiques importants (Blundell and Wood, 1975). Il a été observé que la plupart des substitutions d'acides aminés ont subi une pression évolutive positive pour limiter les effets délétères (Sawyer et al., 2003). Les agrégats de protéines sont non fonctionnels, insolubles, cytotoxiques et contiennent de nombreux peptides ou protéines. La propension à s'agréger dépend de la stabilité de la protéine, de sa charge et de sa tendance à former des brins β (Dobson, 2003). La plupart des mutations, même conservatives, changent la stabilité de la protéine, et des mutations importantes comme l'introduction de résidus polaires dans le cœur hydrophobe de la protéine, peuvent entraîner un mauvais repliement et la rendre inactive (Matthews, 1995). Par exemple, les 2/3 des mutations non synonymes de la protéine Cro du bactériophage λ affectent significativement sa stabilité (Pakula et al., 1986). Beaucoup d'études de stabilité après mutations ont été menées et montrent que la plupart des substitutions d'acides aminés entraînent un effet sur la stabilité de la protéine (DePristo et al., 2005). L'agrégation est également très sensible aux mutations : la diminution de la stabilité augmente la propension à former des agrégats (Chiti et al., 2000). Les mutations qui affectent la stabilité ou l'agrégation sont distribuées sur toute la protéine, contrairement aux mutations qui affectent le site actif et qui sont localisées. Cependant il existe certains cas où les protéines moins stables sont favorisées : dans les environnements à basse température, les enzymes psychrophiles sont moins thermostables que leurs homologues

thermophiles, ce qui leur permet de se replier à basse température, mais aussi de conserver leur flexibilité, essentielle pour la catalyse (Somero, 1995).

Certaines protéines, comme les histones ou l'actine, dont la fonction implique des interactions extensives avec plusieurs protéines ou l'ADN, ont de fortes contraintes évolutives et les modifications de séquence ne sont que de 10% par milliard d'année (Doolittle, 1992). Par contre, pour les domaines de la superfamille des immunoglobulines, dont les gènes dupliqués ont des fonctions différentes, moins d'un tiers de la structure des domaines est conservée et seuls le caractère hydrophobe et la taille approximative de douze résidus sont maintenus dans la séquence. Donc la divergence de séquence des protéines est inversement liée aux contraintes dues à leurs fonctions (Chothia and Gerstein, 1997).

- *L'évolution de la séquence est contrainte par la structure*

Les gènes n'évoluent pas uniformément, certaines positions sont plus modifiées que d'autres.

Sasidharan *et al*. (Sasidharan and Chothia, 2007) ont comparé des orthologues entre homme et souris, homme et poulet, *Escherichia coli* et *Salmonella enterica*, et ont observé que les modifications observées montrent le même schéma. Il y a 190 types de mutations possibles et tous les événements de substitution d'acide aminé possibles sont observés. Cependant, les fréquences des modifications suivent une distribution exponentielle. Cette distribution décroît avec la divergence : 75% des mutations entre deux séquences qui ont divergé de moins de 10% correspondent à 30 types de mutations. De plus ces types de mutations sont conservatifs. Pour les paires de protéines qui ont divergé de 50 à 60%, 65 types de mutations représentent les 3/4 des mutations. Ces mutations conduisent à des acides aminés moins conservés dans leur forme ou leurs propriétés physico-chimiques. De plus, pour des séquences ayant une forte similarité, les substitutions d'acides aminés sont peu nombreuses à l'intérieur de la protéine et quand il y en a, elles sont conservatives. À l'inverse, des séquences plus divergentes entraînent des modifications plus nombreuses à l'intérieur de la protéine, y compris non conservatives. Donc l'augmentation du nombre de changements de résidus enfouis est à l'origine de la relation exponentielle entre divergence de séquence et de structure présentée équation 1. Les modifications sont plus ou moins nombreuses selon l'emplacement du résidu dans la protéine.

De plus, les résidus à l'interface des protéines et les résidus de liaison à un ligand ou à un site actif sont également davantage conservés. Teichmann (Teichmann, 2002) a comparé les protéines orthologues de *Saccharomyces cerevisiae* et de *Saccharomyces pombe* et a montré que les protéines qui n'ont pas d'interaction connue ont une identité de séquence de 38% contre 46% pour les protéines impliquées dans des complexes stables, et 41% pour les protéines impliquées dans des interactions transitoires.

L'analyse d'une banque d'alignements multiples basés sur les structures 3D par Pascarella *et al*. (Pascarella and Argos, 1992) montre que les indels[4] sont généralement courts (1 à 5 résidus) et prennent souvent une structure de tour/boucle. Les interruptions dans les hélices et les feuillets sont très rares. Les structures semblent tolérer peu d'indels.

I.A.2 La structure quaternaire : les complexes protéiques

Les protéines ont une taille variable, de quelques dizaines àà plus de 34 000 acides aminés. Mais la plupart des protéines sont composées de plusieurs chaînes, d'une taille de 100 à 750 acides aminés, avec une majorité comprise entre 150 et 450 acides aminés (Goodsell and Olson, 2000).

La majorité des protéines est oligomérique. Goodsell (Goodsell, 1991) a calculé que la structure quaternaire moyenne des protéines est le tétramère. Seulement 20% des protéines d'*E. coli* seraient monomériques (Goodsell and Olson, 2000). Les dimères sont les plus fréquents et représentent plus de 38% des protéines d'*E. coli*. La plupart des oligomères ont un nombre d'unités pair (Klotz et al., 1970). Les monomères sont dominants dans la PDB[5], mais c'est dû au fait qu'ils sont plus faciles à cristalliser.

I.A.2.a Pourquoi construire des protéines oligomériques ?

L'évolution des protéines est conduite par deux forces contradictoires : leur fonction favorise les grosses protéines, et les mécanismes de synthèse favorisent les petites. Les raisons sont présentées ci-après. Un compromis entre les deux serait de construire des grosses protéines oligomériques assemblées à partir de plusieurs monomères de petite taille.

[4] Indel : insertion ou délétion
[5] Protein Data Bank

Quels seraient les avantages des grosses protéines ? Kohsland (Koshland, 1976) suggère que pour les cellules primitives qui avaient des membranes perméables, les grosses protéines avaient moins de chance d'être perdues. Goodsell et Olson (Goodsell and Olson, 1993 ; Goodsell and Olson, 2000) proposent plusieurs avantages présentés par les grosses protéines :

1. *La fonction morphologique* : les grosses protéines sont plus stables, et certaines fonctions nécessitent une grande stabilité, comme certaines protéines de structure ou de capside. Srere (Srere, 1984) ajoute que pour avoir des interactions spécifiques entre protéines, par exemple pour la régulation ou la localisation, il faut que la protéine soit suffisamment grosse. Cette stabilité pourrait également permettre d'orienter les résidus actifs des protéines (Monod et al., 1965 ; Srere, 1984). D'autre part, Payens (Payens, 1983) propose qu'une grande surface accessible au solvant, pour les grosses protéines, permette de former une sorte d'entonnoir pour guider le substrat vers le site actif.
2. *La fonction coopérative* : l'allostérie et les associations multivalentes sélectionnent les grosses protéines avec plusieurs sites actifs identiques plutôt que les protéines avec un seul site actif (Schultz and Schirmer, 1979).
3. *La stabilité contre la dénaturation* : les grosses protéines ont un repliement plus stable que les petites protéines car elles compensent la faiblesse des interactions internes par leur nombre. Les petites protéines doivent compenser le faible nombre d'interactions par des interactions plus fortes telles que les ponts disulfures ou les sites métalliques spécifiques.
4. *La réduction de la surface accessible au solvant* : la surface accessible est plus faible pour une grosse protéine que pour plusieurs petites. Réduire la surface accessible au solvant revient à réduire la quantité de solvant nécessaire pour hydrater la protéine, et donc la place occupée par une protéine et son solvant. La réduction de la surface accessible au solvant protège également de la dégradation. Pauling *et al*. (Pauling, 1953)

suggèrent que cela pourrait aussi diminuer le risque de s'agréger au hasard.

De plus, Goodsell *et al*. (Goodsell and Olson, 1993) ont remarqué que la taille de la protéine était approximativement inversement proportionnelle à la taille de son substrat. Cela permet aux petites protéines qui ont un gros substrat de diffuser plus facilement vers leur substrat et inversement,les petits substrats peuvent diffuser facilement vers les grosses protéines. .

Pour ces multiples raisons, il serait donc intéressant de construire des grosses protéines. Cependant la construction de ces protéines d'un seul tenant présente plusieurs inconvénients et il est donc avantageux de synthétiser des petites protéines qui s'assembleront en oligomère. Plusieurs explications sont proposées par Klotz (Klotz et al., 1970) :

1. *Contrôle des erreurs* : les erreurs de traduction sont réduites : les erreurs sont en moyenne de $5.\ 10^{-4}$ par codon. Cela représente une espérance de 0,25 pour une protéine de 500 acides aminés alors qu'une protéine de 2000 acides aminés aura une espérance de 1 de contenir une erreur. La majorité des erreurs, cependant, a de faibles conséquences et diminue seulement faiblement la fonction de la protéine. Cette hypothèse nécessite que les protéines bien formées soient séparées des mal formées avant l'assemblage.
2. *Efficacité de codage* : le gène qui code pour un monomère qui s'assemble en oligomère est évidemment plus court qu'un gène qui coderait pour la même protéine en une seule chaîne. 85% à 90% des protéines oligomériques sont composées de plusieurs chaînes identiques (Pereira-Leal et al., 2006).
3. *Régulation de l'assemblage* :. Certaines protéines composées de nombreuses sous-unités identiques, comme par exemple l'actine ou les microtubules, ont un système de contrôle de l'assemblage. Des protéines se lient à l'actine ou aux microtubules afin de réguler chaque étape de l'assemblage et de la séparation de ces protéines qui polymérisent spontanément.

4. *Régulation de la fonction* : la fonction des protéines peut être régulée via l'association ou la dissociation des oligomères.

5. *Pression de sélection* : Monod *et al*. (Monod et al., 1965) suggèrent que les protéines oligomériques subiraient une plus forte pression de sélection et de ce fait seraient davantage conservées. Fornasari *et al*. (Fornasari et al., 2007) ont montré que l'évolution des séquences était contrainte non seulement par la structure 3D de la protéine mais également par sa structure quaternaire. La conservation de la structure quaternaire impose donc des contraintes supplémentaires. André *et al*. (Andre et al., 2008) ont étudié l'énergie d'interaction entre oligomères symétriques et non symétriques, et ils ont calculé que l'écart type de l'énergie d'interaction associée à une mutation à l'interface est plus grande pour les complexes symétriques que pour les complexes asymétriques. Ils expliquent ceci par le fait que la mutation sera répercutée sur les deux monomères.

I.A.2.b La symétrie dans les complexes protéiques

Les monomères sont le plus souvent asymétriques (Chothia, 1991). En effet, l'asymétrie des acides aminés L[6] entraîne une préférence pour les hélices α et les feuillets β. Les cas de symétrie interne sont assez rares, c'est le cas des tonneaux α–β. Ils sont souvent le résultat d'une duplication de gène et doivent être vus comme une sorte de structure quaternaire liée. Par contre, la symétrie est la règle plutôt que l'exception chez les protéines oligomériques (Goodsell and Olson, 2000).

La plupart des protéines solubles ou liées à la membrane forment des complexes protéiques d'au moins deux sous-unités, et presque toutes les protéines de structure sont des polymères symétriques composés de plusieurs centaines à plusieurs millions de sous-unités.

Il existe plusieurs formes de symétrie (Figure 5) :

[6] Les acides aminés ne sont pas symétriques et la totalité des acides aminés présents dans les protéines sont de la forme L, c'est à dire que le groupe -NH2 est situé à gauche en représentation de Fisher, par opposition aux acides aminés D où le groupe –NH2 est situé à droite.

1. *Les groupes cycliques* : les monomères présentant une symétrie cyclique CN sont superposables par une rotation de 360°/N autour d'un axe. La symétrie C2 (dimères) est fréquente parmi les protéines.

2. *Les groupes dihédraux* : deux rotations autour de deux axes perpendiculaires permettent de superposer les monomères. Ce type de symétrie est très répandu parmi les enzymes cytoplasmiques solubles, en particulier les tétramères qui ont des symétries D2. Les oligomères ayant une symétrie dihédrale ont plusieurs types d'interface possibles : les interfaces entre oligomères permises par la symétrie de rotation principale et les interfaces dimériques permises par l'axe de symétrie C2 perpendiculaire. Cela fournit une infrastructure riche à partir de laquelle il est possible de construire des contrôles allostériques. Les possibilités de stabilité et d'interaction sont donc plus intéressantes qu'avec une symétrie cyclique.

3. *Les groupes cubiques* : ils contiennent une symétrie C3 combinée avec une autre symétrie de rotation autour d'un axe non perpendiculaire. Il y a trois possibilités : tétrahédral avec des axes C2 et C3, octahédral avec des axes C3 et C4, et icosahédral avec des axes C3 et C5. Les protéines qui ont ce type de symétrie jouent généralement un rôle dans le stockage et le transport.

4. *Les groupes de symétrie linéaire, planaire et spatiale* : l'addition d'une symétrie de translation à des symétries de rotation forme des structures hélicoïdales, des plans symétriques et des cristaux qui remplissent l'espace. Ces symétries sont indépendantes et elles peuvent être étendues indéfiniment, jusqu'à ce que l'organisme manque de place, de sous-unités ou arrête la croissance mécaniquement.

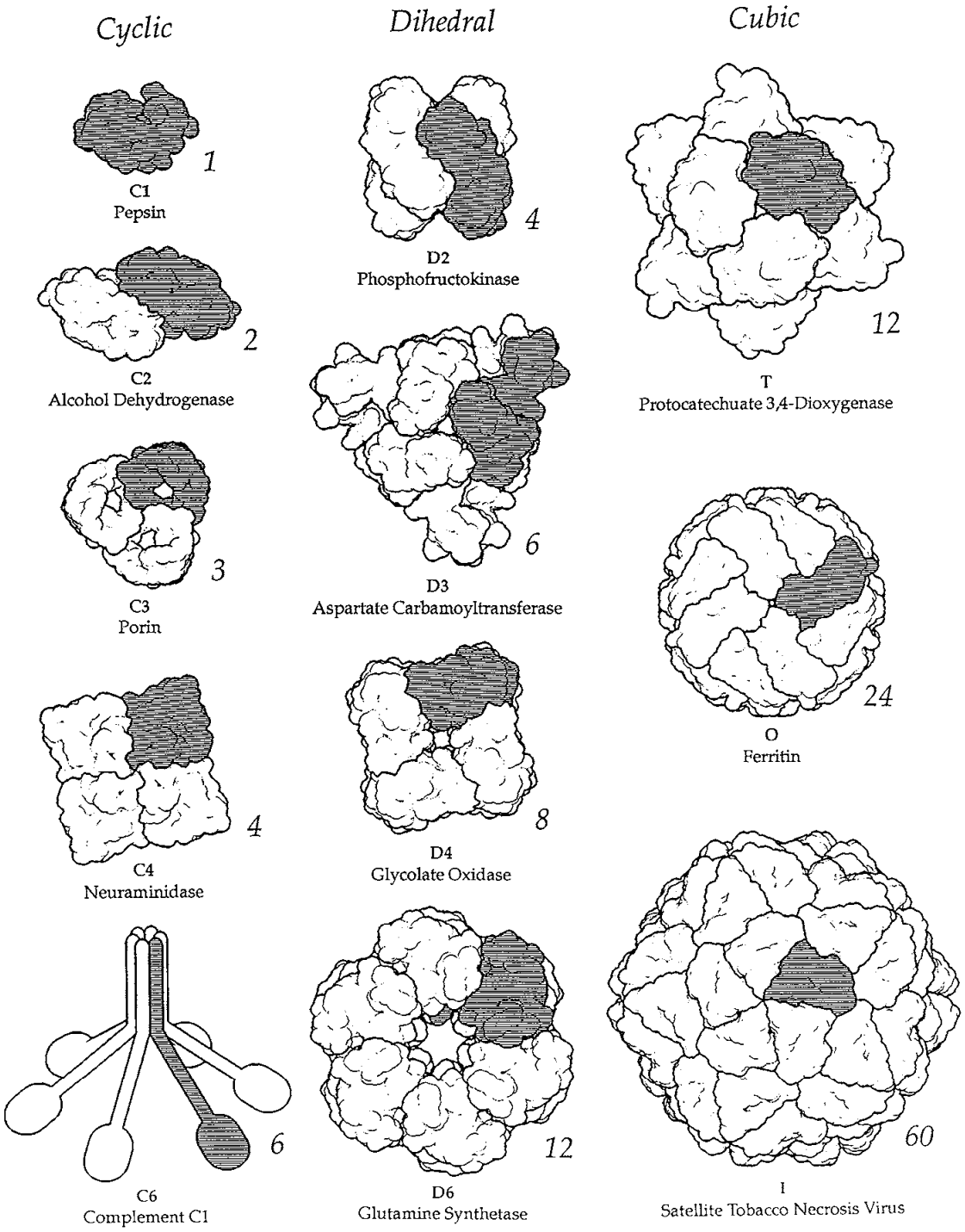


Figure 5 : Groupes de symétries cristallographiques.
Exemples de protéines avec le groupe de symétrie cristallographique qui leur correspond. Les symboles des groupes sont inclus en dessous de chaque structure (ex C1 et D2), et le nombre de sous unités identiques dans chaque groupe est inclus en dessous et à droite de la structure (ex 24 pour le groupe Octahédral). Une sous-unité est grisée dans chaque exemple. Notez que d'autres groupes non cristallographiques sont en accord avec la nature énantiomorphique des protéines, par exemple les symétries cycliques C5, C7 ou plus hautes, et les symétries dihédrales D5, D7 et au dessus. Les codes d'accession de la Protein Data Bank : pepsin, 5pep; alcohol dehydrogenase, 2ohx; porin, 2por; neuraminidase, 1ivb; phosphofructokinase, 1pfk; aspartate carbamoyltransferase, 1at1; glycolate oxidase, 1gox; glutamine synthetase, 2gls;

protocatechuate-3,4-dioxygenase, 3pcg; ferritin, 1hrs; satellite tobacco necrosis virus, 2stv. Source : Annu. Rev. Biophys. Biomol. Struct. 2000.29:105-153. (Goodsell and Olson, 2000)

La grande majorité des homo-oligomères est symétrique (Goodsell and Olson, 2000) (très peu d'exemples inverses existent, par exemple (Grimes et al., 1998 ; Kohlstaedt et al., 1992)). Quels en seraient les avantages ?

1. *La stabilité de l'association* : Blundell *et al.* (Blundell and Srinivasan, 1996) pensent que l'état de plus faible énergie est celui qui comporte une symétrie. Cornish-Bowden *et al.* (Cornish-Bowden and Koshland, 1971) ont fait plusieurs analyses thermodynamiques pour savoir si la symétrie présentait un avantage. Ils ont montré que les assemblages symétriques sont largement favorisés.
2. *Finition de l'assemblage* : l'agrégation peut avoir des conséquences délétères pour la protéine. La symétrie permet de créer des oligomères avec un nombre défini de copies, ce qui permet d'éviter l'agrégation.
3. *Efficacité de repliement* : Wolynes (Wolynes, 1996) suppose que les protéines symétriques ont moins de barrières cinétiques pour se replier. Il se base sur une analogie avec de simples groupes d'atomes.

I.A.2.c Fonction des complexes protéiques

Il existe un certain nombre de protéines qui possèdent plusieurs sites actifs (Goodsell and Olson, 2000). Ces protéines peuvent paraître désavantagées car certaines mutations pourraient invalider plusieurs sites actifs à la fois. Il y a certains avantages : la fonction de la protéine peut être modifiée par l'interaction entre plusieurs sites (coopérativité, inhibition) (Schultz and Schirmer, 1979).

Certaines fonctions vont également favoriser des protéines symétriques. En effet, la symétrie permet l'assemblage d'un nombre défini de monomères, ce qui peut permettre de mesurer des distances, d'entourer des cibles moléculaires ou de créer des contenants de taille fixe. Par exemple, la symétrie C2 crée une règle allostérique qui permet aux répresseurs de mesurer la longueur d'une répétition sur l'ADN. La symétrie C2 facilite également la régulation allostérique : un peu comme un compas, la fixation d'un inducteur ou d'un inhibiteur à un endroit charnière peut modifier la distance entre deux extrémités fonctionnelles.

Les symétries de type C3 ou plus peuvent former des pores ou des cavités. Kelman *et al*. (Kelman et al., 1995) ont remarqué que plusieurs des protéines qui encerclent l'ADN double brin ont des symétries de rotation à 60°, et sont composées soit de six unités distinctes, soit de six domaines similaires. Par exemple les facteurs d'élongation de l'ADN polymérase ont une symétrie approximative C6, avec six domaines arrangés en anneau autour de l'ADN. Chez *E. coli*, la polymérase III contient deux sous-unités β de trois domaines chacun (Kong et al., 1992) alors que le facteur d'élongation de l'ADN polymérase δ eucaryote contient trois sous-unités de deux domaines chacun et adopte la même forme (Krishna et al., 1994).

Les anneaux oligomériques sont également utilisés comme pores dans la bicouche lipidique (Goodsell and Olson, 2000). Par exemple la perforine forme des pores de taille variable et le pore nucléaire contient de nombreuses protéines et présente une symétrie de type C8. Les anneaux servent également à la création de moteurs de rotation comme ceux des flagelles d'*E. coli* et de *Salmonella typhimurium*.

Les protéines oligomériques sont utiles pour attraper et stocker des grandes molécules ou un grand nombre de molécules. Par exemple, ces protéines peuvent présenter une symétrie cyclique, et former une cavité au centre. D'autres protéines ont une symétrie dihédrale formée par l'assemblage dos à dos de 2 sous-unités ayant une symétrie cyclique. C'est le cas de la chaperonne bactérienne GroEL, qui a une symétrie D7, et forme une cavité pour guider les protéines immatures (Xu et al., 1997). Les symétries cubiques sont également utilisées pour construire des contenant encore plus grands. Par exemple, la ferritine utilise une symétrie octahédrale pour contenir les ions fers (Theil, 1987).

Il existe à l'inverse plusieurs fonctions biologiques qui ont besoin de protéines non symétriques. Par exemple les polymérases et les ribosomes ont des réactions asymétriques et directionnelles. Par contre, certaines sous-unités de la polymérase peuvent avoir des symétries locales. Une exception est l'ADN polymérase III bactérienne, qui est un gros complexe enzymatique et a une symétrie globale de type C2.

I.A.2.d Origine et évolution des protéines oligomériques

Les complexes protéiques seraient apparus tôt dans l'évolution : le génome de LUCA (Last Universal Commun Ancestor) aurait contenu des oligomères, comme par

exemple l'ARN polymérase (Makarova et al., 1999). Les protéines essentielles sont pour la majorité des oligomères (observé chez les levures), et les protéines essentielles sont souvent très conservées dans l'évolution donc il se pourrait que les oligomères soient apparus tôt dans l'évolution et soient conservés dans la majorité des espèces (Pereira-Leal and Teichmann, 2005).

Entre 7% et 20% des complexes protéiques ont des similarités avec d'autres complexes, et auraient donc évolué par duplication (Pereira-Leal et al., 2006; Pereira-Leal and Teichmann, 2005). Un grand nombre de complexes auraient évolué progressivement par duplications partielles. En effet, il est peu probable que tous les gènes qui codent pour un complexe soient dupliqués en même temps, excepté si ces gènes sont localisés les uns à côté des autres dans le génome, par exemple dans un opéron (Reams and Neidle, 2004), ou lors d'une duplication totale de génome (Skrabanek and Wolfe, 1998). Cependant, la duplication d'une partie seulement d'un complexe risque d'entraîner des différences de dosage qui peuvent avoir des effets délétères (Veitia, 2004).

Les protéines qui participent à la formation de plusieurs complexes sont très fréquentes et sont plus susceptibles d'être codées par des gènes essentiels. Les complexes dupliqués gardent globalement la même fonction, mais ont des spécificités différentes de liaison et de régulation, ce qui indique que les duplications seraient liées à une spécialisation fonctionnelle. Au moins 40% des interactions chez la levure *Saccharomyces cerevisiae* résultent de duplication avec conservation des interactions (Pereira-Leal and Teichmann, 2005). Les protéines évolueraient par duplication, suivie de divergence par rapport aux interactions de la protéine ancestrale.

Est-ce que les complexes proviennent de l'assemblage de parties ou de la division d'un tout ? Le fait d'avoir plusieurs gènes serait intéressant pour limiter la fonction de chaque gène et les effets délétères de certaines mutations. De plus, la fusion est environ quatre fois plus fréquente que la fission (Kummerfeld and Teichmann, 2005; Snel et al., 2000), ce qui est en faveur de la première hypothèse.

90% des complexes protéiques contiennent des contacts entre des protéines identiques dans le même complexe (Pereira-Leal et al., 2006). Les interactions des homo-oligomères sont principalement hydrophobes (Bahadur et al., 2003), ce qui facilite et optimise la création de ces structures. Dans les complexes stables, l'interface

entre les protéines est souvent large, avec une surface enfouie supérieure à 2500 $Å^2$ (Janin and Chothia, 1990). Cependant, de nombreux complexes ne sont pas stables et s'assemblent et se séparent très rapidement, ils impliquent des interfaces plus petites, de moins de 2000 $Å^2$ (Pereira-Leal et al., 2006).

Il semble y avoir une corrélation entre le nombre d'interactions protéiques et le fait que ces protéines soient indispensables : les protéines qui ont le plus d'interactions sont les plus indispensables (Jeong et al., 2001), mais cet effet est faible (Rocha, 2006).

I.A.3 Les domaines protéiques

I.A.3.a Définition

Le domaine est un niveau de description intermédiaire entre la structure secondaire et la structure tertiaire des protéines. Les protéines peuvent contenir un domaine ou plusieurs, selon leur taille, mais la majorité contient plusieurs domaines (Vogel et al., 2004a). Les domaines ont une taille entre 100 et 250 résidus environ (Chothia et al., 2003). Le domaine peut être défini de plusieurs façons.

Le domaine a tout d'abord été décrit de façon structurale (Wetlaufe, 1973) : c'est une région globulaire compacte également contiguë en séquence. Il est également défini comme une unité structurale distincte, compacte et stable qui se replie indépendamment des autres unités (Branden and Tooze, 1999).

En génomique comparative, un domaine est une unité qui a une origine évolutive commune avec les autres domaines de la même famille, et qui peut avoir été transférée et recombinée dans une autre protéine (Rossmann et al., 1974).

Ces définitions de domaines sont différentes, mais la plupart du temps, les familles de domaines qui en résultent sont les mêmes (Ekman et al., 2005). Les domaine peuvent avoir une fonction propre ou contribuer à la fonction d'une protéine multi-domaines en coopération avec les autres domaines (Vogel et al., 2004a).

L'identification des domaines a un intérêt du point de vue structural : s'il est possible de définir des unités qui se replient indépendamment, il pourrait être plus facile de prédire le repliement d'une protéine contenant plusieurs domaines dans son ensemble, et donc sa structure.

De plus, la prédiction de domaine à partir des séquences pourrait apporter des informations précieuses pour l'annotation des génomes, dans la mesure où la fonction, ou sous-fonction, conférée par le repliement du domaine est souvent conservée lorsque le domaine se lie à d'autres domaines (Han et al., 2006).

Les domaines sont donc très étudiés en tant qu'unité d'évolution à part entière, et un grand nombre d'études de duplications intragéniques ont été réalisées sur les domaines, elles seront présentées dans la 2ème partie de cette introduction.

I.A.3.b Les bases de données de domaines

Il existe plusieurs bases de données de domaines qui ont chacune leurs spécificités. Elles diffèrent concernant leur définition du domaine, le type de représentation formelle du domaine, le mode d'acquisition de nouvelles données (automatique, manuel, hybride). J'ai choisi de décrire cinq bases de données, sélectionnées parce qu'elles sont très utilisées, et certaines ont servi à des analyses dans ma thèse. Les deux premières sont des bases de données de séquences et les trois suivantes des bases de données de structures, même si les premières contiennent des données structurales et les dernières se basent souvent sur des alignements de séquences pour définir leurs classes.

- *Pfam (Finn et al., 2008)*

Les domaines de Pfam sont créés à partir d'alignements multiples et des profils HMM[7] sont calculés pour chaque domaine. Les domaines Pfam ont une longueur moyenne de 145 acides aminés (Bornberg-Bauer et al., 2005). Cette base de données est divisée en deux parties : Pfam-A, qui est validée manuellement et Pfam-B qui est générée automatiquement.

Dans Pfam-A, un alignement multiple est fait à partir d'un petit nombre de séquences représentatives contenant le domaine étudié, et il est vérifié manuellement. A partir de cet alignement, un profil HMM est calculé. Il permet de générer un alignement complet de toutes les séquences qui contiennent ce domaine. Les domaines Pfam-B sont générés automatiquement à partir d'alignements multiples de la base de données Prodom présentée ci-après ((Sonnhammer and Kahn, 1994)).en enlevant les séquences correspondant à Pfam-A.

[7] HMM (Hidden Markov Model), chaînes de Markov cachées.

La version 23.0 disponible en septembre 2008 contenait 10340 familles, couvrant plus de 70% des protéines de SWISS-PROT et TrEMBL.

- *ProDom (Sonnhammer and Kahn, 1994)*

ProDom est une base de données de domaines générée automatiquement et basée sur la comparaison globale de toutes les séquences protéiques disponibles de SWISS-PROT et TrEMBL. Dans un premier temps, les domaines homologues à des domaines connus sont identifiés par l'algorithme mkdom2 (Gouzy et al., 1999) qui est un algorithme récursif basé sur PSI-BLAST (Altschul et al., 1997). Le principe est qu'on considère que la séquence la plus courte correspond à un domaine et cette séquence est cherchée contre la base de données pour trouver des domaines homologues. Les séquences homologues sont ensuite retirées de la base de recherche, mais si ces séquences sont plus longues, la partie qui ne correspond pas au domaine est conservée dans la base de recherche. Le processus est itéré ensuite en utilisant à chaque fois la séquence la plus courte comme requête.

Certaines améliorations sont apportées pour augmenter l'efficacité de la méthode : les fragments de séquence sont supprimés de la base pour ne conserver que les protéines entières, et les séquences de faible complexité sont masquées pour éviter les faux positifs. Les domaines doivent avoir plus de vingt acides aminés. Les répétitions internes sont utilisées comme requêtes à la place de la séquence entière. Pour limiter les faux positifs, le seuil d'e-value de PSI-BLAST est de 10^{-6} et l'e-value est calculée par rapport à la taille initiale de la base de données.

Pour augmenter la rapidité de la procédure, une matrice de score position spécifique (PSMM) est construite à partir des domaines connus, puis lancée comme requête contre la base de données au début du processus. La plupart de ces domaines connus sont issus de SCOP.

Pfam et ProDom sont maintenant réunies avec PROSITE et PRINTS dans la base InterPro (2007). Ce projet rassemble ces bases de données sous le même format.

En septembre 2008, InterPro contenait 4941 domaines et 11 128 familles.

- *SCOP (Strucural Classification Of Proteins) (Murzin et al., 1995)*

La classification de SCOP est faite manuellement. Les structures sont d'abord découpées en domaines structuraux (régions avec un cœur hydrophobe et peu d'interactions avec le reste de la protéine). Ensuite elles sont classées selon 4 niveaux :

1. *Classe* : selon la composition en structures secondaires : il y a quatre classes principales : tout α, tout β, α/β (α et β mêlés), $\alpha+\beta$ (α et β) dans des régions séparées. Les sept autres classes sont beaucoup plus petites.
2. *Repliement* : similarités dans l'arrangement spatial et dans les connexions entre structures secondaires.
3. *Super-famille :* les structures et fonctions suggèrent une évolution commune. L'identité de séquence peut être faible.
4. *Famille :* au moins 30% d'identité de séquence ou alors des fonctions et structures très similaires.

Le domaine est considéré comme une unité d'évolution. Il peut être seul dans une protéine, ou associé à d'autres. Les domaines SCOP ont en général 100 à 250 résidus, avec une moyenne de 175 acides aminés, et sont plus longs que ceux de Pfam (Gerstein, 1997). En 2004, la base de données SUPERFAMILY basée sur la classification SCOP contenait des domaines homologues à au moins un domaine de 59% des protéines des génomes séquencés et 49% des résidus de ces protéines (Madera et al., 2004)

En septembre 2008, la base de données contenait 97 178 domaines, qui étaient répartis en 1086 repliements, 1777 super-familles et 3464 familles.

- *Astral (Chandonia et al., 2004)*

Astral provient de la base de données SCOP et la complète.

Les structures de la PDB se voient attribuer un score AERO SPACI qui représente la qualité de la structure. Il prend en compte la résolution de la protéine, le facteur R qui indique la qualité de la structure, des paramètres stéréochimiques de vérification et il pénalise les structures modifiées.

Astral est basé sur les champs SEQRES et ATOM des fichiers PDB, qui contiennent la liste des résidus. Les différentes chaînes sont concaténées en reprenant

l'ordre dans le gène. Les chaînes de moins de vingt résidus et celles qui contiennent plus de 20% de résidus inconnus sont enlevées.

Il existe trois classements pour créer des groupes non redondants. Dans chaque groupe, la structure qui a le meilleur score AEROSPACI est conservée comme représentant. La classification est basée sur les domaines :

1. Identité de séquence avec Blast dans les deux séquences, les séquences sont faites à partir des champs SEQRES et ATOM des fichiers PDB.
2. Identité de séquence avec Blast dans les deux séquences, mais en plus la e-value est calculée sur une banque de 100 000 000 résidus (taille de SWISS-PROT ou TrEMBL en 2000), ce qui permet de ne pas sur-estimer la significativité des paires dans une base de données petite.
3. Classification basée sur les domaines SCOP. Il y a cinq classes : « classe », « repliement », « superfamille », « famille », et « protéine et espèces ».

- *CATH (Class Architecture Topology Homology) (Orengo et al., 1997)*

Cette classification est faite par une méthode hybride, c'est-à-dire automatique et manuelle. Les protéines sont découpées en domaines selon un consensus de trois méthodes et si elles s'accordent, c'est le découpage de DETECTIVE (Swindells, 1995) qui est conservé. Dans les autres cas, le découpage est effectué manuellement.

CATH contient quatre niveaux principaux et trois niveaux supplémentaires :

1. *Classe :* les quatre classes sont : « mainly α », « mainly β », « mixed α–β », « few secondary structures ». L'assignation est automatique dans 90% des cas.
2. *Architecture :* similarité dans l'organisation générale des structures secondaires. Cette classification est manuelle.
3. *Topologie :* similarité dans le nombre, l'ordre et les connexions des structures secondaires.
4. *Superfamille homologue*: similarité de fonction, laissant supposer un ancêtre commun.

Les autres niveaux correspondent à des protéines ayant une identité de séquence >35%, >95% et >100%. L'alignement est effectué par l'algorithme de Needleman et Wunsch (Needleman and Wunsch, 1970).

La version 3.2.0 de septembre 2008 contenait 114 215 domaines.

I.A.3.c Le domaine comme unité d'évolution

Plusieurs travaux ont étudié l'évolution des domaines plutôt que l'évolution des protéines, en considérant le domaine comme une unité d'évolution. Ces études ont porté à la fois sur les domaines structuraux et les domaines protéiques.

- *L'évolution des domaines*

D'après Chothia *et al*. (Chothia et al., 2003), les principaux mécanismes qui permettent l'accroissement du répertoire protéique sont (1) la duplication de séquences qui codent pour un ou plusieurs domaines, (2) la divergence des séquences dupliquées par mutation, délétion, insertion pour modifier les structures qui pourront acquérir de nouvelles propriétés et être sélectionnées ; et dans certains cas (3) la recombinaison de gènes qui crée un nouvel arrangement de domaines. En effet, il est plus facile de créer un nouveau domaine ou une nouvelle protéine par duplication puis divergence que *de novo*.

Fong *et al*. (Fong et al., 2007) proposent que la majorité des protéines multidomaines évoluent à partir de fusion ou fission. Ils expliquent 87% des architectures de domaines protéiques par des événements de fusion, de fission ou l'introduction de nouveaux domaines. Les réarrangements seraient peu nombreux. La fusion de domaines est entre 4 et 5,6 fois plus fréquente que la fission (dans les domaines SCOP et Pfam) (Fong et al., 2007; Kummerfeld and Teichmann, 2005). Les protéines composées d'un seul domaine apparaissent souvent comme nouveau domaine et non comme fission d'une protéine multidomaine.

Les protéines peuvent aussi évoluer par insertion ou délétion de domaines. Bjorklund *et al*. (Bjorklund et al., 2005) ont étudié les domaines SCOP et Pfam et ont trouvé que les insertions / délétions sont plus fréquentes que les duplications internes, et que les échanges de domaines sont rares. Les insertions n'ont pas toujours lieu entre deux domaines et peuvent survenir à l'intérieur d'un domaine, dans ce cas, la probabilité de l'insertion est proportionnelle à la longueur du domaine dans lequel un

autre domaine est inséré (Aroul-Selvam et al., 2004). Les indels et les répétitions sont plus fréquentes aux extrémités N et C terminales des protéines (Bjorklund et al., 2005). Parfois, le domaine n'est pas réellement délété, mais il peut avoir évolué rapidement, n'est pas détectable par les méthodes classiques, et pourrait ne plus être fonctionnel (Bornberg-Bauer et al., 2005)

Après duplication, un domaine peut évoluer par divergence et éventuellement se combiner avec d'autres domaines pour former une protéine multidomaines. Dans certains cas, il peut acquérir une nouvelle fonction ou sa fonction peut être modifiée (Vogel et al., 2004a). Il est possible que certains événements de recombinaison s'effectuent au niveau des introns (Patthy, 1999). Une étude a montré que les exons étaient significativement corrélés aux domaines Pfam (Liu and Grigoriev, 2004). De Souza *et al*. (de Souza et al., 1996; de Souza et al., 1997) suggèrent que les régions qui bordent les introns pourraient correspondre aux régions « linker » qui relient deux domaines. Les domaines sont définis comme des segments qui, une fois repliés, tiennent dans une sphère de 28 Å de diamètre. Les auteurs trouvent un excès significatif de « linkers » parmi les régions bordant les introns des protéines étudiées.

- *Les domaines dans les différents règnes*

La plupart des protéines contiennent plusieurs domaines. Selon les calculs, sur des domaines protéiques et structuraux, le nombre de protéines multidomaines varie de 40% à 70% pour les procaryotes et de 65% à 80% pour les eucaryotes (Ekman et al., 2005 ; Liu and Rost, 2004 ; Teichmann et al., 1998). Les protéines multidomaines sont peu fréquentes chez les archées, cela pourrait être dû au fait que les grosses protéines sont peu stables dans les environnements hyperthermophiles (Koonin et al., 2002).

Chothia *et al*. (Chothia et al., 2003) ont étudié 429 familles de domaines SCOP présentes chez 14 génomes eucaryotes. Les protéine appartenant à ces familles correspondent à 80% des domaines chez les animaux, et 90% des domaines chez les champignons et les plantes. Ils ont remarqué que la plupart de ces protéines sont communes à tous les eucaryotes et remonteraient donc à l'origine des eucaryotes, voire à l'origine des différents domaines.

I.A.3.d La dynamique des domaines

- *Le nombre de domaines est limité – loi de puissance du peuplement des domaines*

Le nombre de séquences possibles paraît presque illimité, mais le nombre de domaines est fini et relativement restreint. Environ 90% des 600 enzymes des voies métaboliques de petites molécules d'*E. coli* sont construites à partir de 213 familles de domaines (Teichmann et al., 2001b). Le nombre de repliements possibles est estimé entre 650 et 10 000 (Orengo et al., 1994 ; Wang, 1998 ; Wolf et al., 2000). Ces repliements suivent une loi de puissance (Qian et al., 2001) : quelques familles de domaines sont très fréquentes, la majorité des protéines appartiennent à 1000 repliements différents ; mais la plupart des familles sont peu fréquentes. C'est pour ces repliements que l'estimation varie le plus. En 2004, il était estimé qu'entre 1/3 et 2/3 des résidus des génomes séquencés pouvaient être associés à l'une des 800 famille structurale SCOP ou CATH (Grant et al., 2004).

Pourquoi le peuplement de ces familles de domaines suit-il une loi de puissance ? D'après Bonrberg-Bauer *et al*. (Bornberg-Bauer et al., 2005), ce serait dû à une combinaison d'événements évolutifs neutres et de sélections pour les propriétés fonctionnelles de certaines familles. Certains domaines évolueraient par convergence, auraient donc une origine évolutive différente et n'auraient pas de similarité de séquence. Ils convergeraient vers le même repliement car certaines contraintes fonctionnelles favorisent certains arrangements de chaînes (Gerstein, 1997).

- *L'ordre des domaines est conservé*

L'ordre des domaines est très conservé : dans 90% des cas, deux domaines sont dans le même ordre dans une protéine (Apic et al., 2001a). Bashton et Chothia (Bashton and Chothia, 2002) ont observé que la plupart du temps, les domaines A et B sont trouvés dans l'ordre AB ou BA, mais seulement 2% des protéines existent à la fois dans l'ordre AB et dans l'ordre BA. Il y a moins de 1% des combinaisons de domaines possibles qui sont observées chez les protéines multidomaines (Apic et al., 2003 ; Bornberg-Bauer et al., 2005), cela indique une forte conservation de l'ordre des domaines au cours de l'évolution. La même observation est faite avec des triplets de

domaines (Vogel et al., 2004b), et indique l'existence de « supradomaines » qui sont des combinaisons de domaines très conservées.

Deux explications sont possibles : soit cet ordre est nécessaire pour la fonction de la protéine, soit cet arrangement a été créé une fois dans l'évolution puis dupliqué pour créer plusieurs protéines. Bornberg-Bauer *et al*. (Bornberg-Bauer et al., 2005) proposent que l'architecture des domaines est conservée parce que la duplication est plus fréquente que la recombinaison. En effet, la plupart des réarrangements de l'ordre des domaines testés in vitro montrent que l'ordre des domaines a peu ou pas d'effet sur la fonction de la protéine (Lindqvist and Schneider, 1997). Il existe également des permutations cycliques ou circulaires sur les domaines, par exemple chez la méthyltransférase d'ADN ou la lectine (Jeltsch, 1999 ; Young et al., 1982), ce qui est souvent utilisé comme argument pour dire que l'ordre des domaines influe peu sur la fonction (Jung and Lee, 2001). Toutes ces observations sont en faveur d'une seule origine pour la combinaison de domaines, suivie par des duplications.

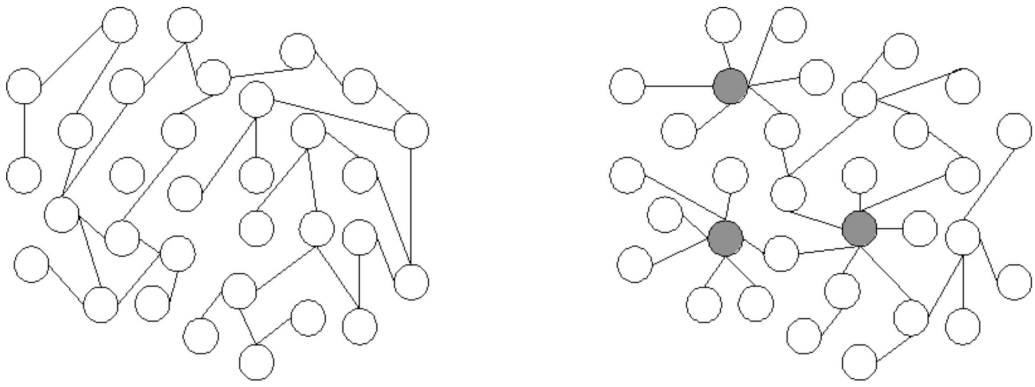
La sélection des combinaisons de domaines serait basée sur leurs fonctions (Vogel et al., 2004b). Les combinaisons de domaines et les superfamilles pourraient avoir joué un rôle important dans l'émergence des organismes plus complexes. Les protéines qui ont la même combinaison de domaines ont probablement un ancêtre commun et des fonctions communes.

- *Les combinaisons de domaines suivent une loi de puissance*

Les combinaisons de domaines suivent une loi de puissance : quelques familles de domaines sont combinées à beaucoup d'autres familles, mais la plupart des familles sont combinées seulement à quelques autres familles (Apic et al., 2001a ; Apic et al., 2001b ; Wuchty, 2001). Environ la moitié des familles ne se combine pas avec d'autres familles, et environ un tiers ne jouxte qu'une autre famille. Ces combinaisons peuvent également être représentées par un graphe « scale-free » (Figure 6). Les familles combinées à beaucoup d'autres familles sont également les plus abondantes dans les génomes (Vogel et al., 2005). Cela suggère une certaine stochasticité des recombinaisons de domaines. Cependant, toutes les combinaisons possibles de domaines ne sont pas observées, ce qui correspond également à l'observation des duplications de combinaison de domaines. Donc la plupart des familles ont peu de voisins. Quelques familles ont beaucoup de voisins et ce sont les familles les plus grandes dans les génomes.

Les domaines les plus fréquents et qui ont le plus de partenaires ont souvent des fonctions particulières. Par exemple les familles des hydrolases de nucléotide triphosphate peuvent agir comme kinases ou hydrolases, seules ou combinées à d'autres familles. C'est également le cas des domaines qui se lient à l'ADN ou à l'ARN, et qui, combinés à d'autres domaines, sont responsables d'une régulation spécifique.

Ye et Godzik (Ye and Godzik, 2004) ont trouvé des groupes de domaines très connectés, qui correspondent à des fonctions particulières comme la transduction du signal ou l'adhésion cellulaire. De plus, l'étude de 255 familles de protéines présentes dans les trois règnes, montre qu'elles correspondent à 80% des familles de protéines archées, 60% des familles de protéines d'eubactéries et 50% des familles de protéines eucaryotes (Apic et al., 2001b). 60% à 90% des combinaisons de domaines spécifiques d'un règne sont faites à partir de ces familles qui sont communes à tous les règnes, mais qui se combinent de façon différente. Cela montre que les nouvelles combinaisons de domaines, même avec des domaines anciens, contribuent au processus de divergence entre les génomes qui inclut la divergence de séquence, l'expansion et la contraction de familles de domaines.



(a) Réseau aléatoire (b) Réseau « scale-free »

Figure 6 : (a) Représentation d'un réseau aléatoire et (b) d'un réseau « scale-free ».
Dans le réseau « scale-free », quelques nœuds sont connectés à beaucoup d'autres (en gris) et la plupart des nœuds n'ont que quelques connections (en blanc). (Issu de http://commons.wikimedia.org/wiki/Image:Scale-free_network_sample.png)

- *La pierre de rosette des protéines*

L'évolution des domaines, par la méthode de la pierre de rosette pour les protéines, a été proposée par Marcotte *et al*. (Marcotte et al., 1999a). Le mécanisme est

le suivant : deux protéines fusionnent en un seul gène. L'interface entre les deux domaines évolue vers une plus haute stabilité, avec par exemple des interactions hydrophobes. Après un grand nombre de générations, une fission du gène a lieu. Les protéines continuent à s'assembler via l'interface qu'elles ont formée. Il est possible de trouver quelques une de ces protéines avant et après fusion dans des bases de données (Bornberg-Bauer et al., 2005). Un des avantages de la fusion de gènes pourrait être la co-régulation après la fission, ou la réduction de génome : il y a 15 gènes qui sont fusionnés chez *Mycoplasma genitalium* et qui sont présents en 2 gènes séparés chez *Mycoplasma pneumonia,* qui a un génome plus grand de 17% (Enright and Ouzounis, 2001).

- *Les domaines et les fonctions des protéines*

Le fait d'ajouter un domaine peut modifier la fonction de la protéine (Hegyi and Gerstein, 2001) : il a été estimé que deux protéines monodomaines contenant un domaine (de SCOP) de la même famille ont 67% de probabilité d'avoir la même fonction contre 35% pour des protéines qui possèdent deux domaines dont un en commun.

La complexité d'un organisme n'est pas strictement liée à son nombre de gènes, le nématode en a plus que la drosophile, et l'homme en a moins que la paramécie (Arnaiz et al., 2007), mais la complexité pourrait être liée à l'expansion de certaines familles particulières qui sont à la base des formes de vie les plus complexes (Chothia et al., 2003). Pour Levine *et al*. (Levine and Tjian, 2003), la complexité d'un organisme est liée à la complexité de son réseau de régulation et à son nombre de facteurs de transcription. En effet, dans les organismes complexes, le nombre moyen de facteurs de transcription pour un gène est plus élevé (van Nimwegen, 2003). De plus, dans beaucoup d'organismes, les domaines de liaison à l'ADN descendent d'un petit nombre de domaines qui ont été utilisés dans des combinaisons variées et dont la fréquence dépend de la lignée.

Certaines familles d'orthologues sont conservées même dans des génomes assez éloignés (Koonin et al., 2002). Cependant, il existe des expansions de familles de protéines qui proviennent de duplications récentes. Ces expansions sont particulièrement fréquentes chez les eucaryotes et correspondent souvent à des

adaptations spécifiques de ces types d'organismes (Jordan et al., 2001 ; Lespinet et al., 2002; Remm et al., 2001).

Une étude a essayé de savoir comment les protéines multidomaines s'adaptent aux voies métaboliques après duplication. Deux solutions sont envisagées : soit l'enzyme garde la même spécificité de substrat mais change d'action catalytique, soit l'enzyme garde sa fonction catalytique et change de substrat. L'analyse des réseaux métaboliques des petites molécules d'*E. coli* indique que presque toujours le site catalytique est conservé, et que c'est le substrat qui change (Teichmann et al., 2001a). Cela conduit à des réseaux métaboliques très hétérogènes, où les enzymes d'un même réseau ont peu ou pas d'origine évolutive commune. La comparaison des voies métaboliques de plusieurs espèces montre la présence de protéines qui ont la même fonction mais appartiennent à des familles différentes, c'est un déplacement non orthologue (Koonin et al., 1996). Chez certains organismes, une partie de la voie métabolique classique n'existe pas et est remplacée par une voie alternative (Dandekar et al., 1999).

I.B Le rôle des duplications dans l'évolution des protéines

Les principaux mécanismes qui permettent l'accroissement et la diversification du répertoire protéique sont les duplications de séquences, leur divergence par mutation, insertion et délétion et la recombinaison des gènes (Chothia et al., 2003). Les duplications jouent donc un rôle important et participent à la modification des fonctions existantes et à la création de nouvelles fonctions.

Je m'intéresserai dans un premier temps aux répétitions en général, qui ont lieu à plusieurs échelles, et qui ont des caractéristiques différentes liées à leur mécanisme de création. Dans une seconde partie, je traiterai les répétitions intragéniques qui sont celles qui font l'objet des études de cette thèse. Elles occupent une place particulière dans la mesure où elles vont modifier une protéine existante et pourront avoir des conséquences sur sa fonction.

I.B.1 Les répétitions

I.B.1.a Historique des répétitions

L'article de Taylor et Raes (Taylor and Raes, 2004) retrace l'histoire de l'étude des répétitions. Les premières observations en 1911 portent sur le nombre de chromosomes : certaines plantes ont leur nombre de chromosomes doublé et ces différences influent sur leur phénotype (Kuwada, 1911 ; Tischler, 1915). Une expérience de Stadler (Stadler, 1929) a montré que les espèces polyploïdes d'*Avena* (avoine) et de *Triticum* (orge) étaient moins sensibles à l'irradiation que les espèces non polyploïdes.

Plus tard, Muller *et al*. (Muller and Gershenson, 1935) ont produit des drosophiles avec une partie du chromosome X dupliqué et inséré dans le chromosome 2. Le mutant est viable, et ils proposent ensuite que ce type de duplications pourrait exister dans la nature et serait un moyen d'augmenter le nombre de gènes sans les conséquences négatives de l'aneuploïdie[8]. L'une des portions redondantes pourrait ensuite diverger. Serebrovsky (Serebrovsky, 1938) proposa en 1938 que la sélection pourrait être

[8] Gain ou perte d'un ou plusieurs chromosomes

relâchée après duplication, à cause de la redondance, et que cela permettrait aux gènes de se spécialiser.

Plusieurs chercheurs proposent que la duplication de génomes, ou en tout cas un accroissement important du génome, serait à l'origine d'événements de spéciation et permettrait une plus grande complexité des organismes (Bridges, 1935 ; Gulick, 1944 ; Stephens, 1951). Plus tard, Britten et Davidson (Britten and Davidson, 1969) constateront que le contenu en ADN n'est pas toujours proportionnel à la complexité des organismes.

Ces idées seront reprises par Ohno (Ohno, 1970) : il propose qu'après duplication, un des deux gènes paralogues est libre d'accumuler des mutations et d'évoluer. Ce gène deviendra souvent un pseudogène et disparaîtra mais dans certains cas, il pourra acquérir une nouvelle fonction. Ohno pense que la duplication de gènes serait le facteur le plus important de l'évolution et que sans les duplications de gènes, l'émergence des métazoaires, vertébrés et mammifères à partir des organismes unicellulaires aurait été impossible. Un événement de duplication totale du génome aurait donc facilité l'émergence des vertébrés.

Vers la fin des années 60, le gel d'électrophorèse a permis de tester plus facilement la duplication de gènes en comptant le nombre de copies de loci codant pour certains isozymes[9]. Des études montrent que les gènes PGI (phospho glucose isomérase) sont dupliqués et exprimés dans différents tissus, et que cette différence d'expression arrive peu après la duplication (Avise and Kitto, 1973). Avec le séquençage de l'ADN, les répétitions ont pu être détectées plus facilement dans l'ADN génomique à l'aide de primers de PCR dégénérés ou de sondes d'ADN. D'autres technologies liées au séquençage ont également permis de nombreuses avancées comme l'hybridation in situ, les puces à ADN basées sur les EST ou les ADNc. Enfin, les méthodes bioinformatiques permettent de détecter les répétitions à grande échelle dans les séquences et génomes séquencés avec des critères statistiques permettant leur caractérisation fine.

[9] Un isozyme est une enzyme présentant une séquence d'acides aminés différente d'une enzyme de référence mais catalysant la même réaction chimique.

I.B.1.b Mécanismes de création et caractéristiques des répétitions

Il existe plusieurs mécanismes qui créent des répétitions avec des caractéristiques différentes, comme leur longueur, leur espacement, ou leur positionnement sur le chromosome.

- *Le dérapage de l'ADN polymérase (Figure 7)*

Lors de la synthèse du néo-brin d'ADN, le néo-brin peut se désapparier et de réapparier de façon décalée. Selon le décalage produit, il peut entraîner un gain ou une perte d'une unité répétée. Les répétitions ainsi créées sont souvent de petite taille et présentes en tandem. Ce mécanisme existe chez tous les organismes.

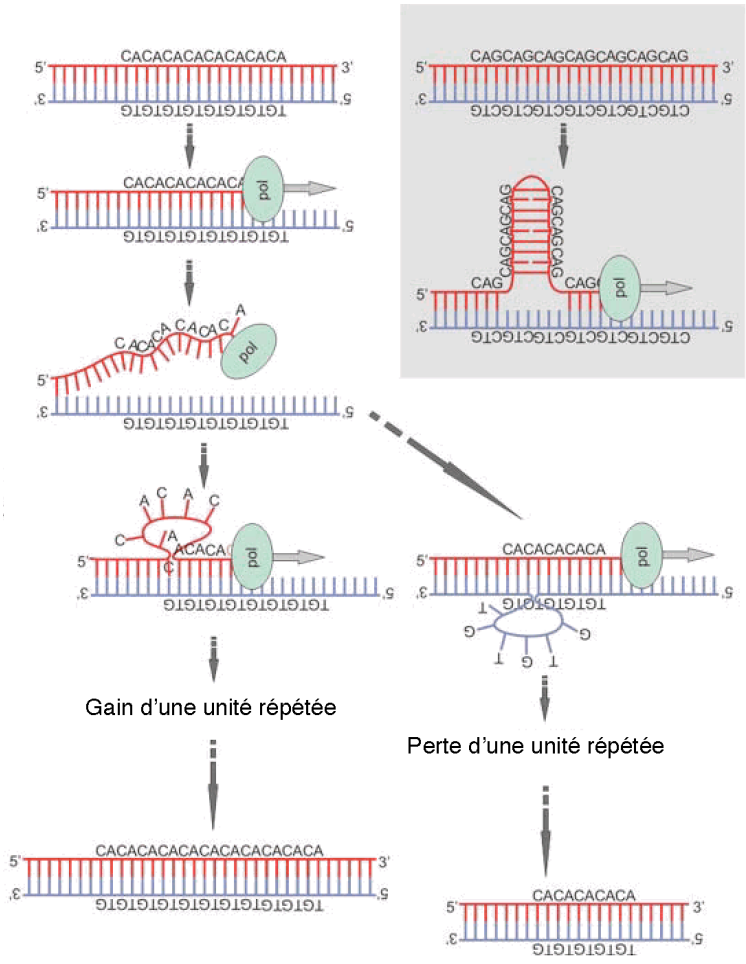


Figure 7 : Schéma de la création de répétitions par dérapage de l'ADN polymérase. (Traduit de http://biol.lf1.cuni.cz/ucebnice/en/repetitive_dna.htm).

- *La recombinaison homologue (Figure 8)*

Ce mécanisme est présent chez tous les organismes. Les événements de recombinaison arrivent grâce à une structure particulière appelée amplicon, il s'agit d'une séquence entourée de deux répétitions directes appelées pieds. Trois mécanismes existent pour expliquer leur duplication. Le premier est le crossing-over inégal entre chromosomes homologues ou entre chromatides sœurs lors de la méiose. Cela entraîne le gain de l'amplicon pour un chromosome (ou chromatide) et sa perte pour l'autre (Anderson and Roth, 1981). Ce mécanisme a été observé chez certains chromosomes et plasmides (Dianov et al., 1991) et conduit à la formation de répétitions en tandem de taille variable. La deuxième possibilité est l'excision de l'amplicon par recombinaison entre ses pieds, et l'insertion de l'amplicon dans le génome (sur le même chromosome ou sur un autre) par une autre recombinaison. De même que pour le mécanisme précédent, chaque répétition est associée à une délétion. La détection d'intermédiaires libres prédits par ce mécanisme suggère qu'il pourrait avoir lieu (Flores et al., 1993). Le 3ème mécanisme est la réplication circulaire de l'amplicon : deux événements de réplication entre les pieds de l'amplicon permettent la création d'un amplicon multicopies, ce qui évite la perte d'un amplicon comme dans les cas précédents (Young and Cullum, 1987). L'observation de structures prédites par ce mécanisme indique qu'il pourrait être utilisé (Petit et al., 1992).

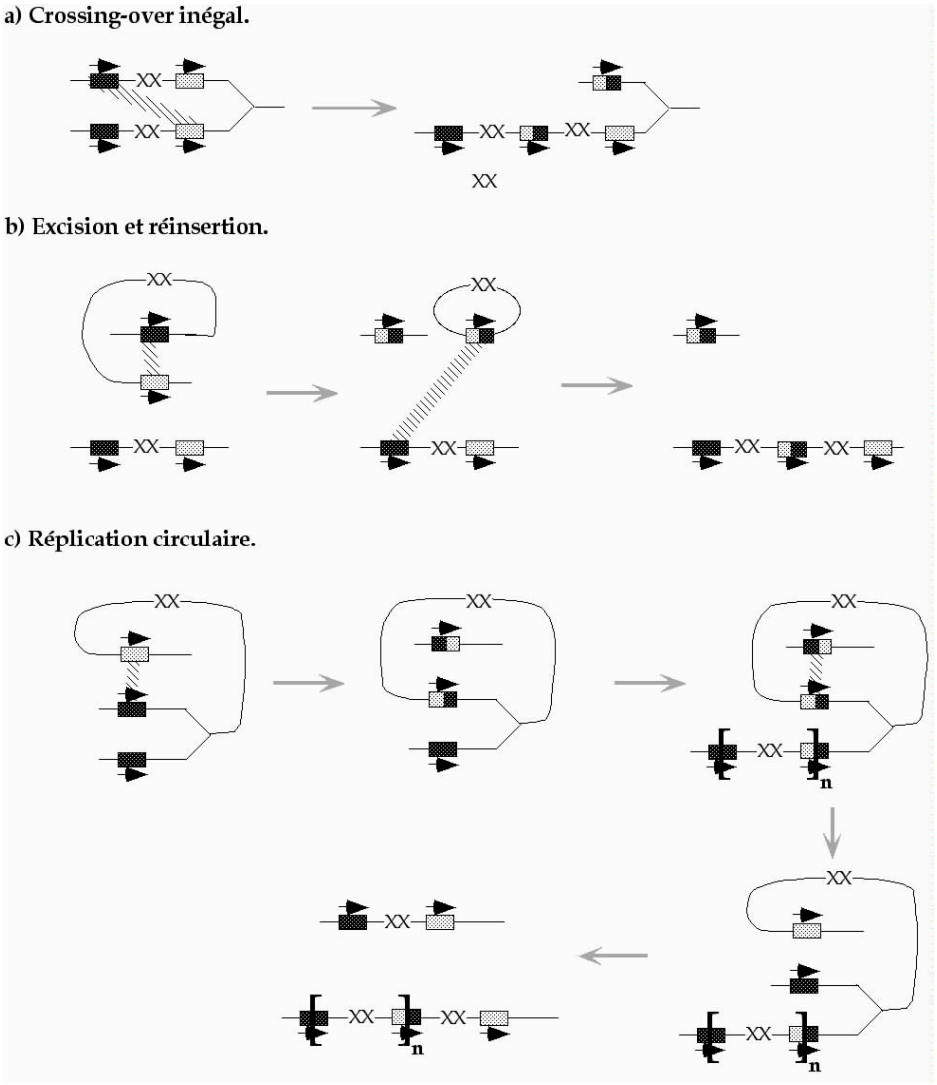


Figure 8 : Mécanismes pour l'amplification d'un amplicon.
a) Crossing-over inégal entre deux chromosomes. b) Excision et réinsertion dans le génome. La réinsertion peut avoir lieu à un locus différent de celui de l'excision. c) Amplification par réplication circulaire. Ce mécanisme permet de créer en une seule étape de nombreuses copies de l'amplicon (Tiré de (Achaz, 2002)).

- *Les répétitions en tandem*

Les deux premiers mécanismes de création de répétitions, dérapage de l'ADN polymérase et recombinaison homologue, peuvent conduire à la fois au gain ou à la perte d'une unité répétée en tandem. Par contre, les répétitions éloignées ne peuvent être délétées que par recombinaison homologue, ou par excision pour les éléments transposables (traités ci-après), et sont donc conservées plus longtemps dans les génomes. Achaz *et al*. (Achaz et al., 2000) proposent que la plupart des répétitions seraient créées en tandem et ensuite éloignées par des réarrangements chromosomiques. En effet, les répétitions en tandem sont beaucoup plus nombreuses que les répétitions éloignées et elles sont moins divergentes.

De plus, les répétitions inverses sont beaucoup moins fréquentes que les répétitions directes, en particulier pour les génomes les plus stables. L'évitement de répétitions est lié à la stabilité d'un génome : les génomes proches qui divergent en terme de stabilité divergent aussi en terme de répétitions inversées. La mesure du potentiel de recombinaison des répétitions inversées (Rocha, 2003) et de leur impact éventuel sur la structure du chromosome montre que ces répétitions sont préférentiellement situées à des endroits dans le chromosome où les recombinaisons homologues produiraient un échange de brin plus petit qu'attendu par hasard.

- *Les éléments transposables*

Ils peuvent se déplacer dans le génome ou se dupliquer en produisant des répétitions espacées. Il y a deux sortes d'éléments transposables : ceux qui passent par un intermédiaire ARN et ceux qui passent par un intermédiaire ADN. Les premiers sont appelés rétrotransposons ou éléments de classe I. Ils se dupliquent par un mécanisme de transposition réplicative : après transcription du gène, celui-ci est rétro-transcrit en ADN puis inséré dans le génome (Figure 9) (Simon et al., 2008). Cela crée des répétitions espacées suivant la taille du rétrotransposon. Il existe deux classes de rétrotransposons : avec ou sans LTR (Long Terminal Repeat), les premiers sont encadrés de longues répétitions inversées. Celles-ci permettent au rétrotransposon, après rétrotranscription, de s'insérer dans le génome par recombinaison dans une séquence du génome contenant des LTR. Les rétrotransposons sans LTR s'insèrent dans une séquence contenant une cassure double brin.

La 2ème catégorie est celle des transposons à ADN ou éléments de classe II, qui passent par un intermédiaire ADN. Ils se déplacent par un mécanisme de « couper-coller » : l'élément est excisé du génome grâce à une transposase et inséré à un autre endroit (Curcio and Derbyshire, 2003). Cela forme une coupure double brin de l'ADN qui sera réparée par les mécanismes de réparation présentés ci-après, et dans certains cas, la cassure sera réparée en copiant une séquence similaire, et cela créera une répétition. Certains de ces éléments peuvent également se dupliquer par transposition réplicative (comme pour les rétrotransposons). Ces éléments peuvent également se dupliquer si la transposition a lieu au moment de la réplication, et que le site donneur (d'où le transposon a été excisé) a été répliqué mais pas le site cible. Les séquences d'insertion sont les transposons à ADN les plus simples, et contiennent un gène codant

pour la transposase. Les transposons composites sont composés d'un ou plusieurs gènes encadrés par deux séquences d'insertion.

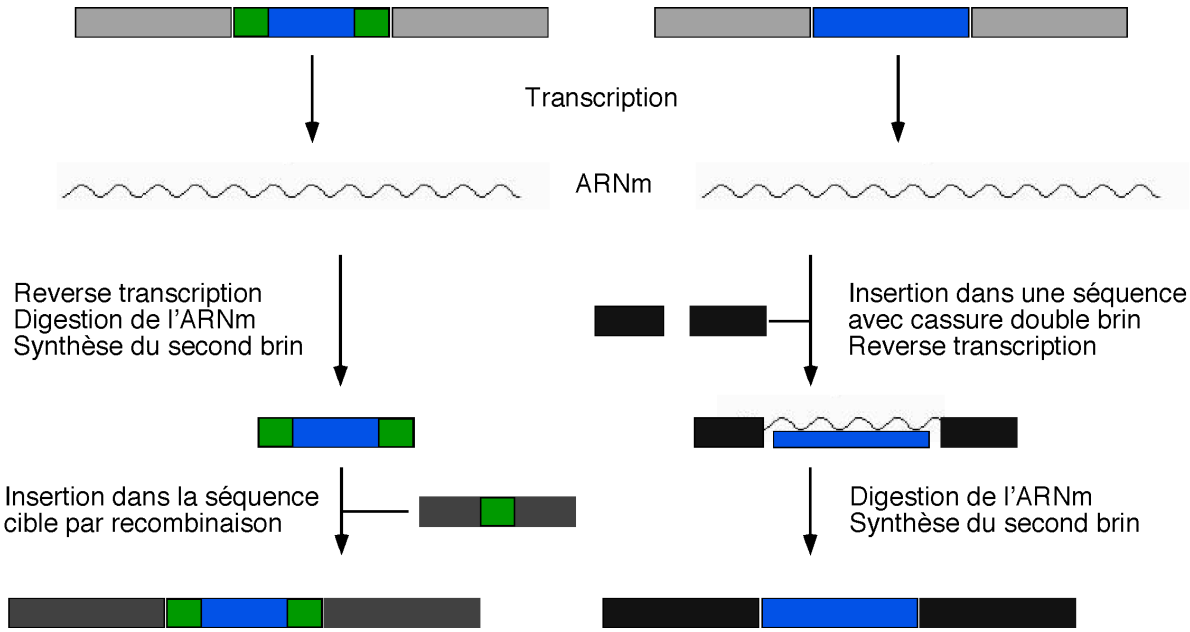


Figure 9 : Schéma simplifié de la rétrotransposition.
(A gauche : rétrotransposons à LTR, à droite rétrotransposons sans LTR, les LTR sont représentés en vert).

- *La duplication de chromosome(s)*

Suite à une mauvaise ségrégation des chromosomes pendant la méiose, un organisme peut se retrouver avec un ou plusieurs chromosome(s) excédentaire(s). La duplication de génome complet a deux origines possibles : soit l'autotétraploïdie (duplication de chaque chromosome) soit l'allotetraploïdie (fusion de deux génomes diploïdes). Le dernier cas a l'avantage d'apporter d'avantage de diversité génétique.

Lorsque l'organisme possède un chromosome de trop, cela peut entraîner des maladies, comme dans le cas de la trisomie 21. L'étude de génomes séquencés comme la paramécie (Aury et al., 2006) ou le tétraodon (Brunet et al., 2006) ont montré l'existence de plusieurs duplications complètes des génomes dans l'évolution de ces espèces, qui auraient contribué à l'expansion des familles de gènes et à l'évolution des génomes chez une grande diversité d'espèces. Une des plus importantes conséquences de la duplication de génome pourrait être la spéciation, suite à un isolement post-reproductif (Lynch and Conery, 2000).

- *La réparation de l'ADN double brin (Figure 10)*

La cellule a trois possibilités pour réparer une cassure double brin dans l'ADN. Si les extrémités double brin correspondent à une séquence qui existe dans le génome, par exemple sur le chromosome homologue, elle pourra être copiée pour réparer l'ADN par recombinaison homologue, ce qui créera une répétition (Paques and Haber, 1999). De façon alternative, la cellule peut utiliser un des deux mécanismes suivants. Si l'une des extrémités de la cassure double brin correspond à une séquence existant dans le génome, les mécanismes de réparation pourront copier en partie cette séquence, ce qui initie une réplication d'un fragment du chromosome et crée donc une répétition (réplication induite par une cassure) (Kraus et al., 2001). Dans le dernier cas, aucune extrémité ne correspond à une séquence connue, les mécanismes peuvent coller les deux extrémités double brin soit entre elles, soit en insérant une séquence d'ADN double brin réparation par jonction des extrémités non homologues). Cela peut aboutir dans certains cas à l'insertion ou à la perte d'une séquence ; dans le cas d'une insertion, cela peut créer une répétition, si la séquence existe ailleurs dans le génome, par exemple dans le cas d'un rétrotransposon sans LTR (Critchlow and Jackson, 1998). Dans tous les cas, les répétitions créées sont espacées et de tailles variables.

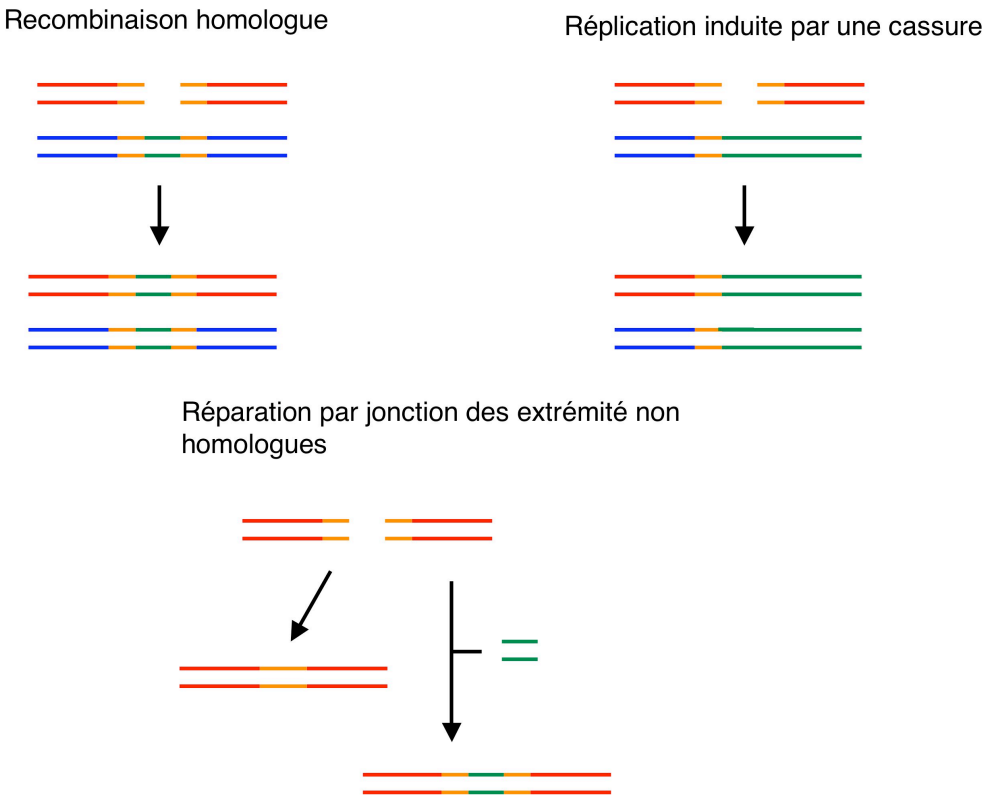


Figure 10 : Mécanismes de réparation des cassures double brin.

- *Les répétitions satellites*

Ces petites répétitions ont une taille comprise entre 1 et 100 pb de long, ce sont des répétitions multicopies en tandem. Elles sont instables et évoluent rapidement (Thomas, 2005). Il existe trois classes de petites répétitions.

Les microsatellites ou SSR (Simple Sequence Repeats) sont des répétitions en tandem de motifs de 1 à 9 pb de long. Elles peuvent être répétées des dizaines voire des centaines de fois. Elles sont plus fréquentes et plus fréquentes chez les eucaryotes. Elles sont formées par dérapage de l'ADN polymérase ou crossing over inégal. Elles sont très instables et perdent ou gagnent facilement des unités répétées. La variation du nombre de copies de ces répétitions serait associées à la diversité des races de chien (Fondon and Garner, 2004), et pourrait jouer un rôle dans l'évolution rapide de variants morphologiques. Les répétitions instables seraient à l'origine de plus de 40 maladies neurologiques, neurodégénératives et neuromusculaires (Pearson et al., 2005). Ce type de répétitions est aussi utilisé par certaines bactéries, par exemple le pathogène

Haemophilus influenzae, pour échapper à la réponse immunitaire de l'hôte : les gènes codant pour des protéines de surfaces ont un nombre variable d'éléments répétés et donc se diversifient très vite (Moxon et al., 1994).

Les minisatellites ou VNTR (Variable Number of Tandem Repeat) sont des répétitions en tandem de 10 à 100 pb de long. Elles sont souvent formées par recombinaison inégale entre chromosomes ou à l'intérieur du même chromosome. Elles seraient souvent présentes à côté des points chauds de recombinaison (hot spot, où les coupures double brin sont fréquentes) et pourraient également être créées par les mécanismes de réparation des cassures double brin de l'ADN.

Les répétitions centromériques ou satellites : ce sont des répétitions d'au moins 100 pb présentes au niveau du centromère, probablement créées par des crossing-over inégaux.

I.B.1.c L'évolution des répétitions après duplication

Les répétitions ont des destins différents en fonction de leur type. Les répétitions en tandem, dont les répétitions satellites, peuvent gagner ou perdre des unités répétées au fur et à mesure des générations. Les répétitions qui sont créées à l'intérieur d'un gène seront plus longuement étudiées par la suite. Je développerai maintenant les cas de duplications d'un ou plusieurs gènes, qui ont été particulièrement étudiées dans la mesure où elles ne sont pas délétères pour la protéine et peuvent conduire à des innovations fonctionnelles.

Comme vu précédemment, Ohno (Ohno, 1970) a proposé qu'après duplication, un des deux gènes paralogues est libre d'évoluer et d'accumuler des mutations. Il deviendra souvent un pseudogène car les mutations délétères sont plus fréquentes que les mutations avantageuses (Wagner, 1998) et la plupart des dupliquas deviennent des pseudogènes en quelques millions d'années (Lynch and Conery, 2000). Mais dans certains cas, le gène dupliqué peut acquérir une nouvelle fonction (Kimura and Ota, 1974) ou se sous-fonctionnaliser (les deux gènes partagent la fonction ancestrale) (Force et al., 1999). Ce cas est illustré par l'exemple du gène « engrailed » du poisson zèbre qui existe en deux exemplaires : l'un est exprimé dans une partie du cerveau et dans les neurones spinaux, et l'autre dans un bourgeonnement pectoral. Le génome du poulet ne contient qu'une copie de ce gène qui est exprimée à ces deux endroits.

D'autres exemples ont été mis en évidence par la suite. Chez la drosophile, il existe deux gènes qui codent pour deux types de soies qui proviendraient d'un gène ancestral qui coderait pour toutes ces soies. Après duplication, une des copies aurait eu une mutation qui l'empêche de contrôler un type particulier de soies, et l'autre copie aurait compensé cette perte (Modolell and Campuzano, 1998). Un autre exemple trouvé chez le poisson (Dermitzakis and Clark, 2001) montre deux familles de gènes dont les équivalents humains codent pour des protéines avec un épissage alternatif différent. Certains gènes chez le poisson contiennent une mutation qui empêche l'épissage alternatif. Il existe également des exemples d'acquisition de nouvelles fonctions. C'est le cas de la RNase des singes colombiens mangeurs de feuilles (les autres singes mangent des fruits et des insectes). Ils font fermenter les feuilles dans une poche intestinale antérieure par des bactéries symbiotiques. Leur RNase I a été dupliquée (Zhang et al., 2002). Une copie a gardé la fonction ancestrale de couper l'ARN double brin, et l'autre copie a évolué et lui permet de digérer l'ARN bactérien dans sa poche acide. Cette acquisition a été faite par évolution darwinienne positive, car le nombre de mutations non synonymes par site non synonyme est plus élevé que le nombre de mutations synonymes par site synonyme.

Cependant, la théorie de la néo-fonctionnalisation n'explique pas comment les gènes dupliqués sont conservés dans les génomes suffisamment longtemps pour acquérir une nouvelle fonction, car il est probable qu'ils accumulent des délétions et deviennent des pseudogènes rapidement. De plus, après duplication, les deux copies sont sous faible pression de sélection et évoluent à la même vitesse (Kondrashov et al., 2002). La présence d'un gène en deux copies peut être défavorable à cause du coût métabolique ou d'une altération du dosage (Papp et al., 2003). D'autre part, la théorie de la sous-fonctionnalisation peut expliquer la conservation des deux copies mais suppose que le gène ancestral ait au moins deux fonctions.

La conservation des deux copies après duplication peut être avantageuse pour le dosage des gènes : de nombreux gènes essentiels sont conservés en deux copies et il peut être intéressant d'augmenter leur dosage. Plusieurs exemples ont été observés : l'opéron lactose d'*E. coli* ou la ribitol dehydrogenase, qui permettent de mieux s'adapter au milieu nutritif ou le facteur R de résistance aux antibiotiques, pour lequel le nombre de copies augmente en présence d'antibiotique, et diminue quand l'antibiotique est enlevé du milieu (Anderson and Roth, 1977). Il y a également des exemples

d'augmentation de gènes de protéines membranaires ou protéines sécrétées. Chez les bactéries, la majorité des gènes récemment dupliqués codent pour des molécules de surface qui, chez les pathogènes, sont impliquées dans l'interaction avec les cellules de l'hôte. Chez la levure, il existe des gènes dupliqués ayant une fonction de transporteurs membranaires, liée à la réponse au stress. Dans les cellules eucaryotes, il s'agit de récepteurs et de protéines sécrétées (Kondrashov et al., 2002).

Le fait que les gènes essentiels soient conservés à cause d'un plus fort dosage pourrait expliquer également pourquoi, après avoir subi un relâchement de pression de sélection juste après la duplication, ils évoluent en moyenne moins vite que les autres gènes (Davis and Petrov, 2004 ; Jordan et al., 2004 ; Kondrashov et al., 2002). Les protéines codées par les gènes essentiels ont souvent de fortes contraintes fonctionnelles.

Les gènes dupliqués issus de duplication totale de génome ou de duplications à plus petite échelle n'ont pas le même destin. La création de gènes dupliqués est en moyenne de 0,01 par gène et par million d'années chez les eucaryotes (Lynch and Conery, 2000). La demi-vie moyenne des dupliquas est de 4 millions d'années. Cependant, si ce taux de disparition est fort, des observations chez les espèces polyploïdes indiquent une forte conservation des gènes dupliqués après duplication de génomes complets (Ferris and Whitt, 1979) et la demi-vie des dupliquas après un événement de duplication de génome serait de 33 millions d'années chez *S. cerevisiae* (Hakes et al., 2007). Cela pourrait s'expliquer si le dosage des gènes joue un rôle important dans la conservation des dupliquas : le dosage des gènes est conservé en cas de duplication totale du génome, mais pas forcément en cas de duplication à petite échelle (Lynch and Force, 2000). Si le dosage n'est pas conservé, cela peut entraîner la formation d'oligomères différents qui peuvent être délétères (Veitia, 2004). Les gènes issus de duplications totales de génome ont tendance à garder une fonction plus similaire que ceux issus d'une duplication à petite échelle, sont moins indispensables, et font plus souvent partie de complexes. Au contraire, les gènes issus de duplications à petite échelle sont souvent essentiels, et même s'ils sont moins souvent conservés, lorsqu'ils le sont, ils sont plus susceptibles d'acquérir une nouvelle fonction (Hakes et al., 2007).

Un autre modèle explique la conservation des gènes dupliqués dans le génome avant acquisition d'une nouvelle fonction : il s'agit du modèle d'innovation,

amplification et divergence (Francino, 2005; Hendrickson et al., 2002). Il suppose que la protéine originale ait une fonction principale, et au moins une fonction secondaire, cette dernière n'étant pas forcément nécessaire à l'organisme. Après un changement de niche environnementale, cette fonction secondaire devient avantageuse, mais est présente en faible quantité, et dans ce cas une ou plusieurs duplication(s) du gène permet(tent) d'accroître l'expression de cette protéine. Ensuite, les copies du gène vont évoluer, jusqu'à ce que la fonction de l'une d'entre elles accumule des mutations avantageuses pour la fonction secondaire et qu'elle puisse remplacer toutes les copies du gène. Ces copies n'étant plus sous pression de sélection, elles pourront devenir des pseudogènes et disparaître par délétion. Le principe de ce modèle est que la présence de plusieurs exemplaires du gène, conservés à cause du dosage de gène sur la nouvelle fonction avantageuse, augmente la probabilité qu'au moins une des copies acquière une mutation qui améliore la fonction secondaire. Il est également possible que des recombinaisons entre les copies permettent d'atteindre la fonction voulue. Cette théorie est soutenue par plusieurs observations. Des bactéries *E. coli* contenant un gène *lacZ* avec une mutation qui diminue son activité, ont été placées sur un milieu minimum contenant du lactose. Après plusieurs générations, les génomes contenaient plusieurs copies de ce gène pour compenser sa faible activité, et il a été observé que certaines mutations ont permis de restaurer l'activité initiale de ce gène. Ceci est aussi en accord avec le fait que le peuplement de familles de gènes suivent une loi de puissance : certaines familles sont surreprésentées (Qian et al., 2001), il est donc possible que les familles présentes en grand nombre, dont certaines ont des fonctions spécifiques des métazoaires, aient évolué de cette façon. De plus, la plupart des paralogues conservés dans les génomes seraient sous pression positive, et ne connaîtraient pas de période avec une évolution neutre (Wagner, 2002). Tous ces arguments sont en accord avec la théorie « innovation, amplification, divergence ».

I.B.2 Les duplications intragéniques : création, évolution et conséquences

I.B.2.a Caractéristiques et répartition dans les différents règnes

La plupart des répétitions sont situées dans des régions non codantes, mais certaines d'entre elles sont à l'intérieur d'un gène (Marcotte et al., 1999b). Ces répétitions ont une variété considérable de taille, de la répétition d'un acide aminé, ou la

répétition en tandem de quelques acides aminés, à la répétition de domaines homologues (plus de 100 acides aminés, comme les domaines des anticorps) (Heringa, 1998). Si ces répétitions entraînent un changement du cadre de lecture, les conséquences pour la protéine peuvent être importantes dans la mesure où cela va modifier un certain nombre d'acides aminés ainsi que le codon stop, et la protéine risque de perdre sa fonction. Si le cadre de lecture est conservé, les conséquences seront moins délétères, mais la fonction de la protéine peut quand même être modifiée, ce qui peut entraîner des disfonctionnements comme des maladies. Seules les répétitions qui ne sont pas trop délétères pour les génomes seront conservées.

Marcotte *et al*. (Marcotte et al., 1999b) ont montré que plus de 14% des protéines contiennent des répétitions. Lavorgna *et al*. (Lavorgna et al., 2001) ont calculé qu'en moyenne 6% des protéines d'archées contiennent des répétitions, contre 5% pour les bactéries. Ces répétitions intragéniques sont trois fois plus nombreuses chez les eucaryotes que les chez les procaryotes (Marcotte et al., 1999b), et les protéines d'organismes unicellulaires (*S. cerevisiae*) contiennent moins de répétitions que les protéines d'organismes pluricellulaires (*C. elegans*), qui contiennent moins de répétitions que d'autres protéomes plus complexes. Le protéome humain est celui qui contient le moins de répétitions intragéniques. (Lavorgna et al., 2001). Parmi les bactéries, *Aquifex aeolicus* et *Thermotoga maritima* ont un taux de répétitions dans les protéines relativement important, cela pourrait être expliqué par un grand nombre de transferts avec les archées (Aravind et al., 1998 ; Nelson et al., 1999). A l'époque, le protéome contenant le moins de répétitions est celui de *Rickettsia prowazekii*, qui est une α-protéobactérie pathogène intracellulaire obligatoire. Elle a subi une réduction évolutive, qui a diminué sa complexité (Lavorgna et al., 2001). Le recouvrement des répétitions présentes chez les procaryotes et les eucaryotes est faible, 4%, et correspond en partie à des cassettes de liaison à l'ATP. Cela suggère que ces répétitions seraient apparues après la divergence entre procaryotes et eucaryotes ou qu'elles auraient beaucoup divergé depuis (Marcotte et al., 1999b).

L'augmentation des répétitions intragéniques pourrait être la cause ou la conséquence de l'augmentation de la taille des protéines durant l'évolution : les protéines procaryotes sont plus petites que les protéines eucaryotes (Lavorgna et al., 2001). Cependant, les protéines archées sont un peu plus petites que les protéines bactériennes, mais contiennent d'avantage de répétitions. A taille égale, il y a toujours

plus de répétitions dans les eucaryotes, que dans les archées, et que dans les bactéries. Les protéines de plus de 500 acides aminés contiennent une part dupliquée proportionnelle à leur longueur. Les répétitions intragéniques ont une composition particulière : dans les répétitions de haute complexité (par rapport aux répétitions qui contiennent un acide aminé répété, ou des acides aminés très similaires), les acides aminés polaires et/ou petits sont sur-représentés et les acides aminés gros et/ou non polaires sont sous-représentés (Marcotte et al., 1999b). De plus, la fréquence des répétitions est liée à la longueur de la séquence et la probabilité de générer des répétitions d'une certaine longueur décroît exponentiellement avec la longueur de la séquence, donc le mécanisme de production des répétitions serait préférentiellement la recombinaison (Marcotte et al., 1999b).

Les protéines qui contiennent des répétitions évolueraient plus vite que les autres protéines (Moxon et al., 1994). Par exemple, le nombre de répétitions des antigènes de surface peut changer rapidement, ce qui modifie le repliement de la protéine. Les génomes multicellulaires compenseraient leur faible taux de génération par cette source supplémentaire de variabilité présente dans les protéines répétées. Les protéines qui contiennent des répétitions en tandem sont souvent des protéines à la surface des cellules chez *S. cerevisiae* (Verstrepen et al., 2005) et chez les bactéries (Moxon et al., 1994). Le fait de contenir plusieurs répétitions peut changer les propriétés de la protéine : par exemple FLO1 adhère mieux quand elle contient d'avantage de répétitions.

Les protéines qui contiennent le plus de répétitions ont des fonctions spécifiques des eucaryotes, par exemple des protéines du tissu conjonctif, les protéines du cytosquelette, des protéines ribonucléiques, des protéines du muscle, des synapses ou du cerveau, ou de l'adhésion cellulaire (Marcotte et al., 1999b). Beaucoup de protéines transmembranaires ont évolué par duplication intragénique (Shimizu et al., 2004).

Les duplications ont été particulièrement étudiées chez les protéines très répétées ou dans le cadre des duplications de domaines.

I.B.2.b Les protéines très répétées

Il est estimé que ces répétitions sont présentes chez 0,79% des archées, 1,05% des bactéries et 5,31% des métazoaires (annotation « repeat » dans SWISS-PROT) (Andrade et al., 2001). Environ 20% des protéines humaines contiennent de multiples

unités répétées de 30 à 40 acides aminés (Ferreiro and Komives, 2007). Les protéines très répétées sont présentes dans plusieurs lignées. Ces protéines possèdent une structure secondaire régulière et forment des ensembles multi-répétés en structure 3D de différentes tailles et fonctions (Andrade et al., 2001). Les répétitions en tandem apparaissent souvent dans des arrangements réguliers, soit alignés les uns à côté des autres, soit dans des superhélices. Pour ces structures, il n'y a pas de limite théorique du nombre de répétitions.

Ces protéines pourraient avoir des avantages différents de ceux conférés aux domaines (Andrade et al., 2001). En général, ces protéines offrent des perspectives évolutives grâce à l'augmentation de la surface de contact. Elles sont relativement libres d'évoluer et leur taille peut augmenter sans affecter leur stabilité (c'est plus facile que par création *de novo*). Ces protéines ont une taille assez importante comparée aux domaines. Les contraintes sur la conservation de la séquence sont relativement faibles à cause des fonctions de liaison résultant des multiples répétitions. Leur succès pourrait être lié au fait qu'elles peuvent acquérir différentes fonctions, et se lier à des ligands différents en fonction de leur surface accessible au solvant.

Ces structures présentent une large surface accessible au solvant qui leur permet de se lier à de larges substrats comme les protéines et les acides nucléiques. Cette propriété peut augmenter le répertoire des fonctions cellulaires comme, par exemple, concernant le transport des protéines, l'assemblage de complexes protéiques ou la régulation des protéines. Par contre, les répétitions dans une superhélice entraînent une structure de type tonneau, avec une faible surface de contact disponible pour les interactions avec les ligands. Elles sont plus stables et compactes, et peuvent lier des petits ligands (Andrade et al., 2001). Les protéines non structurées qui ont des répétitions interagissent fréquemment avec des structures faites d'éléments répétés, par exemple l'histone H1 avec l'ADN (Wolffe and Guschin, 2000), les petites protéines riches en proline avec l'involucrine, ou d'autres répétitions avec l'enveloppe cellulaire (Steinert et al., 1999). Les protéines répétées non structurées ont peu de résidus non polaires et ne peuvent pas former de cœur hydrophobe pour se stabiliser. Les protéines qui ont des répétitions contenant plus de 20% de proline ont une forte probabilité de ne pas être structurées, sauf le collagène. C'est aussi le cas si la protéine répétée a moins de 30% de résidus polaires (Kajava, 2001).

Après duplication, les séquences peuvent évoluer rapidement. Par exemple, les répétitions HEAT chez les invertébrés et les mammifères ont en moyenne 13% d'identité de séquence (Andrade et al., 2001). Cela indique que les contraintes sur chaque répétition sont assez faibles, par rapport aux contraintes imposées à l'ensemble. Le nombre d'éléments répétés peut varier entre séquences homologues (Iturbe-Ormaetxe et al., 2005; Saupe et al., 1995), indiquant que le gain ou la perte d'une unité répétée survient fréquemment au court de l'évolution.

La plupart des protéines très répétées ne se replient pas indépendamment mais sont stabilisées en se repliant entre elles, c'est le cas des ankyrines (Ferreiro and Komives, 2007). Des analyses de cette protéine ont montré qu'en augmentant le nombre de répétitions, la stabilité de la protéine augmente, et si le nombre de répétitions est diminué, la protéine est déstabilisée (Tripp and Barrick, 2004). Si les délétions ont lieu à l'intérieur de la protéine et non à l'extrémité, la protéine est davantage déstabilisée, en effet, ces délétions produisent plus de variation structurale. Cela pourrait indiquer qu'il est plus avantageux pour les protéines que le nombre de répétitions augmente ou diminue aux extrémités. Les amplifications, ainsi que les insertions ou délétions, seraient en général moins délétères pour les protéines très répétées que pour les protéines globulaires.

Si les protéines très répétées ont été créées par duplication, il faut qu'il y ait un ancêtre commun avec une seule copie. Cependant cet ancêtre ne serait pas stable structurellement. Une explication possible est que cet ancêtre formait un homo-oligomère (Ponting and Russell, 2000). Cependant, il existe peu d'exemples, s'il y en a, de complexes formés d'une protéine répétée et d'un monomère. La protéine répétée serait beaucoup plus stable que l'oligomère, ce qui a peut-être conduit à la disparition de ce dernier.

Il y a de nombreuses familles de protéines très répétées (identifiables à partir des structures secondaires) et je vais détailler les six parmi les plus étudiées (Figure 11).

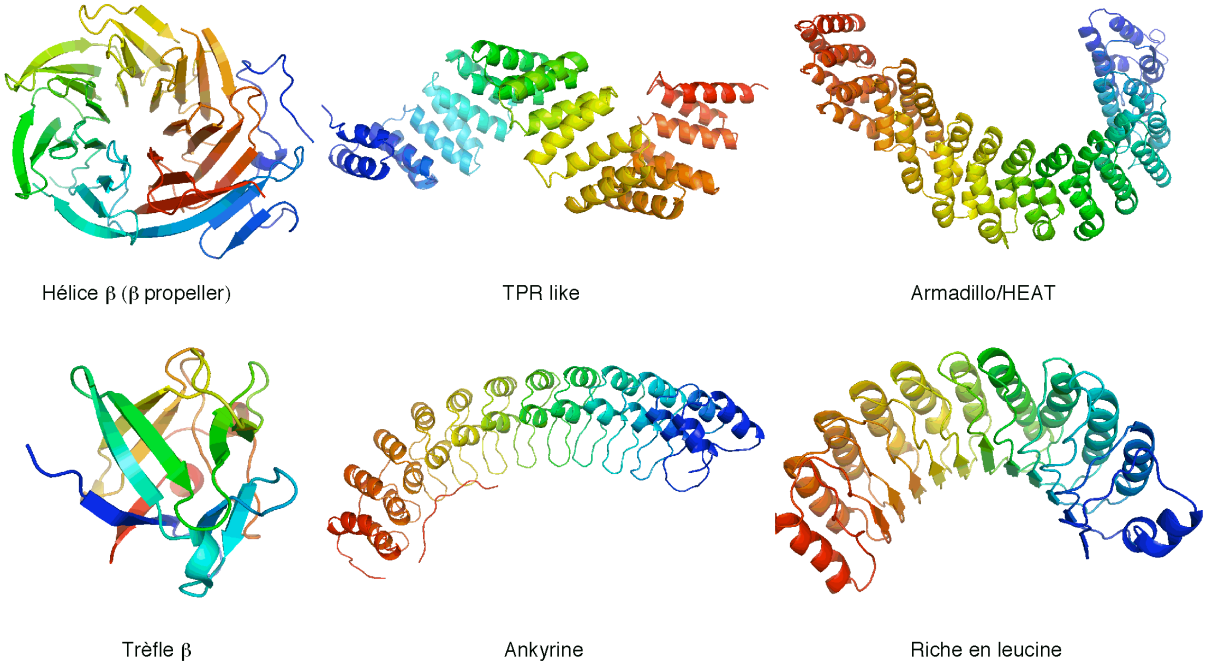


Figure 11 : Protéines faisant partie de six familles très répétées.
Hélice β (β propeller) (1erjA), trèfle β (1ijtA), TRP like (1w3bA), ankyrine (1n11A), armadillo/HEAT (1b3uA), riche en leucine (2ca6A).

- Les hélices β *(β propellers)*

Les répétitions WD40 (Neer et al., 1994) sont les répétitions les plus courantes dans les protéines, et sont de type tout-β. Elles font environ 40 acides aminés de long et contiennent des acides aminés W et D conservés. Chaque répétition est composée de quatre brins β antiparallèles arrangés en hélice autour d'un axe central, comme chez la galactose oxydase ou la neuraminidase. Plusieurs familles de domaines adoptent ce type de structure avec jusqu'à huit pales d'hélice. Ces protéines peuvent lier d'autres protéines ou ligands le long de leur axe, à l'intérieur des feuillets β. Ils pourraient sélectionner leur substrat en fonction de leur taille.

- *Le trèfle β*

C'est une autre structure de type tout-β, dont font partie l'interleukine 1s ou les facteurs de croissance des fibroblastes (Murzin et al., 1992 ; Ponting and Russell, 2000). Elle est constituée de six brins β dont trois forment un tonneau, et les trois autres forment une cape triangulaire autour. Les séquences ont peu de similarités, mais des analyses récentes des quatre trèfles β de protéines de liaison à l'actine montrent une triplication interne de ce trèfle qui conserve des similarités de séquence.

- *TPR like*

Les répétitions tétratricopeptides contiennent environ 34 acides aminés arrangés en deux hélices α, qui sont arrangées à la façon d'un bouton dans un trou (Sikorski et al., 1990). Ces répétitions ont une forte similarité de séquence et proviennent donc d'événements de duplication. Il s'agirait d'une répétition ancienne car elle existe chez les eucaryotes, les bactéries et les archées. Plusieurs répétitions TPR forment une superhélice droite avec une large surface disponible pour se lier à un ligand. Les protéines de ce type ont des fonctions variées comme la chaîne légère de la kinésine, la chaîne lourde de la clathrine, des phosphatases bactériennes, les protéines sécrétées SNAP, etc.

- *Ankyrine*

Chaque répétition contient environ 33 résidus et forme une structure en L constituée de deux hélices α antiparallèles suivies d'une épingle à cheveux β. Les épingles à cheveux β des répétitions successives sont maintenues entre elles par la formation de brins β antiparallèles (Lux et al., 1990). La fonction de ces protéines est de se lier à d'autres protéines, par exemple la p53 (un suppresseur de tumeur), la p65 (un régulateur transcriptionnel), la CDK6 (protéine kinase de la division cellulaire). A chaque fois, l'interaction met en jeu le sillon créé par les brins antiparallèles. Les protéines de cette famille incluent des facteurs de transcription, des régulateurs du développement, des protéines du cytosquelette et des toxines.

- *Armadillo / HEAT*

Les répétitions Armadillo sont composées de trois hélices α (Peifer et al., 1994). L'une d'entre elles est courte et perpendiculaire aux deux autres. Les répétitions HEAT sont composées de deux hélices antiparallèles, dont l'une forme un coude, ce qui fait qu'elle est l'équivalent de deux hélices d'une répétition Armadillo. Les hélices C-terminales de chacune de ces deux répétitions peuvent se superposer. L'assemblage de ces répétitions forme un solénoïde, qui contient un sillon formé par la dernière hélice de la répétition, et les interactions se font généralement également au niveau de ce sillon. Ces protéines sont par exemple dans la plakoglobine (protéine plaque de jonction), la β catenine, les importines β1 et β1.

- *Les répétitions riches en leucine*

Elles sont courtes, environ 20 acides aminés de long (Kobe and Deisenhofer, 1994). Elles contiennent une hélice α et un brin β orientés de façon antiparallèle. L'assemblage de ces répétitions forme une arche. Ces protéines sont impliquées dans la transduction du signal, les récepteurs membranaires, la réparation de l'ADN, l'adhésion cellulaire et les protéines de la matrice extracellulaire.

I.B.2.c Les duplications de domaines

La duplication est une des principales sources de création de domaines (Vogel et al., 2004a). Beaucoup d'études ont été réalisées sur les duplications de domaines.

- *Nombre et répartition dans les différents règnes*

Plus un génome contient de protéines, plus ses domaines sont dupliqués : 58% des domaines de *Mycoplasma* (Teichmann et al., 1998) et 98% des domaines de l'homme (Muller et al., 2002) sont dupliqués. Un peu moins de 10% des séquences contiennent des domaines en tandem (Apic et al., 2001b). Il y a une corrélation entre la complexité des fonctions contrôlées par le protéome d'un organisme et son degré de répétitions internes (Lavorgna et al., 2001), à la fois entre règnes - les protéines archées ont plus de répétitions internes que les bactéries - et à l'intérieur d'un règne - les protéines humaines ont plus de répétitions internes que leurs orthologues chez la drosophile.

De plus, les domaines en tandem sont plus longs chez les organismes multicellulaires que dans les organismes unicellulaires, (Apic et al., 2001b ; Ekman et al., 2005). Les familles impliquées dans les répétitions les plus longues sont spécifiques des métazoaires et ces protéines peuvent contenir 30 à 50 domaines (Apic et al., 2001b). Les répétitions de plus de dix domaines sont fréquentes chez les eucaryotes (Bjorklund et al., 2006).

Les familles, dont les répétitions de domaines sont les plus longues, sont spécifiques des métazoaires (Apic et al., 2001b). Ces familles contiennent des domaines extracellulaires impliqués dans l'adhésion cellulaire et les signaux, ou des familles intracellulaires de signalisation et de régulation. Plusieurs de ces familles spécifiques des métazoaires impliquées dans des longues répétitions sont flexibles en terme de

nombre de répétitions. Par exemple l'immunoglobuline existe en treize tailles différentes, de deux à cinquante-deux domaines (Teichmann and Chothia, 2000).

- *Mécanismes de création des répétitions de domaines*

Les répétitions de domaines sont supposées avoir été créées par recombinaison inégale ou par dérapage de l'ADN polymérase. Bkorklund *et al*. (Bjorklund et al., 2006) ont remarqué que les duplications de plusieurs domaines sont plus fréquentes que les duplications d'un seul domaine.

Street *et al*. (Street et al., 2006) ont observé que certaines duplications pourraient être favorisées par les introns. En effet, de nombreux motifs répétés ont une forte conservation de la position des introns et de la phase, et sont composés d'exons qui codent une ou deux répétitions complètes. Cela suggère la formation de gènes répétés par duplication locale. Il existe également des similarités de motifs d'acides aminés entre les répétitions proches d'un gène. Les introns pourraient faciliter les duplications intragéniques, et ainsi contribuer à l'abondance des répétitions dans les protéines eucaryotes.

- *Problèmes d'agrégation*

Les protéines qui ont plusieurs domaines ont évolué de façon à limiter les mauvais repliements. En effet, les domaines répétés en tandem peuvent poser des problèmes de mauvais repliement ou d'agrégation (Han et al., 2007). D'après Apic *et al*. (Apic et al., 2001b), moins de 10% des séquences génomiques contiennent des domaines en tandem (2 copies ou plus de la même famille), et seulement 10 à 20% des familles de domaines existent en tandem et donc ont sûrement évolué par duplication.

L'agrégation de domaines d'immunoglobuline décroît avec l'identité de séquence (Wright et al., 2005) : les domaines de 70% d'identité de séquence s'agrègent facilement alors que les domaines de 30 à 40% d'identité de séquence ne s'agrègent pas de façon détectable. L'analyse bioinformatique des domaines homologues consécutifs dans les protéines multidomaines montre que ces domaines ont généralement une identité de séquence inférieure à 40%, ils ont donc moins de risques de s'agréger. Les immunoglobulines et les fibronectines de type 3 ont fréquemment des domaines dupliqués en tandems. Pour les domaines de la famille Ig-like, seuls 10% des domaines

adjacents ont plus de 40% d'identité, et l'identité de séquence est significativement inférieure entre domaines adjacents qu'entre domaines non adjacents, ce qui indiquerait que les domaines adjacents sont sous plus grande pression de sélection pour diversifier les séquences. De plus, les protéines dont la similarité de séquence est forte peuvent se replier à l'aide de protéines chaperonnes. La divergence des séquences des domaines en tandem pourrait être expliquée par le fait que certaines protéines répétées en tandem ont évolué par duplication de plusieurs domaines à la fois (Bjorklund et al., 2006).

Certains domaines ne peuvent pas être exprimés s'ils sont en multiples copies en tandem, par exemple des protéines ayant une faible stabilité thermodynamique (Forman et al., 2005) ou des domaines qui ne sont jamais présents dans des protéines multidomaines. Cela suggère que les domaines présents dans les protéines multidomaines ont des signaux spécifiques dans leurs séquences pour éviter les mauvais repliements : ce sont des résidus « gatekeeper » qui sont très conservés et qui peuvent être des résidus chargés, des prolines ou glycines, ou des ponts disulfures (Han et al., 2007; Parrini et al., 2005 ; Steward et al., 2002).

II Objectifs

Les duplications internes peuvent être créées de plusieurs façons, principalement par dérapage de l'ADN polymérase ou par recombinaison inégale. Ces modifications, qui se produisent au niveau des gènes, sont répercutées sur la séquence d'acides aminés, et de ce fait perturbent la structure de la protéine. Les répétitions présentes dans les gènes ont souvent des conséquences importantes sur la fonction des protéines. Dans la plupart des cas, elles entraînent la synthèse de protéines non fonctionnelles, mais elles peuvent aussi conduire à la production de protéines avec une fonction partiellement ou totalement modifiée. Les duplications les moins délétères peuvent être conservées par l'organisme. Le nombre de répétitions à l'intérieur des protéines est assez important, de l'ordre de 14% selon Marcotte *et al*. (Marcotte et al., 1999b).

Il serait intéressant de mieux comprendre quelles sont les répétitions internes qui sont conservées par l'organisme et pourquoi elles le sont, quelles sont les protéines qui les contiennent, quelles sont les conséquences de ces répétitions sur ces protéines et comment les répétitions évoluent. Les structures des protéines évoluent moins vite que leurs séquences et elles gardent donc plus longtemps les traces des événements de duplication. Les répétitions visibles au niveau des séquences sont moins anciennes et apportent des informations sur leur création. Il paraît donc indispensable d'étudier les répétitions trouvées aux trois niveaux en parallèle, séquences nucléiques et protéiques, et structures tridimensionnelles, pour mieux comprendre leur évolution.

Les séquences et les structures connues sont en nombre important et permettent de faire des analyses à grande échelle. Il y a actuellement 54 génomes d'archées, 708 génomes de bactéries et 78 génomes d'eucaryotes disponibles. La base de données EMBL contient environ $6{,}5.10^6$ entrées. La résolution des structures tridimensionnelles des protéines prend plus de temps, et il y a donc beaucoup moins de structures connues que de séquences. Actuellement, la PDB compte 53 521 structures. Néanmoins il sera expliqué plus loin que ces repliements suffiraient à décrire plus de la moitié des protéines existantes.

L'analyse de cette masse de données nécessite des méthodes informatiques. La première étape est d'identifier les répétitions. De nombreux programmes ont été développés pour trouver des similarités locales ou globales dans les séquences ou les

structures, et certains seront détaillés dans le chapitre suivant, mais peu sont destinés à la recherche de similarités internes. Avant ce travail, aucun programme ne permettait de chercher des duplications internes à ces trois niveaux à la fois.

Nous avons donc conçu un nouveau programme, appelé Swelfe, qui permet de trouver ces répétitions. Il est innovant dans la mesure où il permet de chercher des répétitions à la fois dans les séquences d'ADN, d'acides aminés et les structures 3D des protéines. Il permet donc à la fois de trouver des répétitions anciennes qui ne seraient visibles que dans les structures et pourra donner des indications sur l'ancienneté de la duplication en fonction des similarités observées ou non au niveau des séquences correspondantes. Ce programme fait l'objet du prochain chapitre. Les chapitres suivants seront consacrés à l'analyse des répétitions trouvées par Swelfe.

III Swelfe : un outil pour détecter les répétitions dans les séquences et les structures des protéines

Il n'existait pas d'outil spécifique pour chercher les répétitions qui sont contenues dans les séquences d'ADN, les séquences d'acides aminés et les structures tridimensionnelles des protéines. Nous avons donc créé Swelfe. Je présenterai d'abord rapidement quelques algorithmes existant pour chercher des répétitions dans les séquences et les structures 3D, puis je détaillerai l'algorithme utilisé dans Swelfe, ainsi que les paramètres du programme et les vérifications effectuées. Je décrirai ensuite le site web qui a été créé pour mettre ce programme à la disposition de la communauté.

III.A Les algorithmes de comparaison de séquences et de structures

Les algorithmes de comparaison de séquences et de structures sont très souvent utilisés en bioinformatique. En effet, les mutations, ajout ou suppression de nucléotides, duplications, réarrangements qui vont modifier les séquences. Deux séquences présentes par exemple chez deux espèces et issues d'un événement de spéciation, ou chez la même espèce et issues d'un événement de duplication, présentent des différences, qui sont d'autant plus importantes que la séquence ancestrale commune est éloignée. Elles ont donc une origine évolutive commune : ce sont des gènes/protéines homologues, appelées orthologues dans le cas d'une spéciation et paralogues dans le cas d'une duplication au sein de la même espèce. Pour savoir si deux séquences d'ADN ou d'acides aminés se ressemblent, plusieurs algorithmes d'alignement de séquences existent.

La comparaison de séquences peut apporter des informations sur l'évolution des gènes et des protéines. De plus, la plupart du temps, deux séquences proches vont coder la même structure protéique 3D, du fait de la redondance du code génétique et de la similarité de certains acides aminés. Deux structures très proches ont souvent la même fonction. Donc deux gènes proches en séquence ont probablement la même fonction. Les algorithmes de comparaison de séquence sont donc aussi très utilisés pour l'annotation de gènes dans les génomes.

Deux structures proches ont très souvent la même fonction, de ce fait, plusieurs algorithmes de comparaison de structures ont été développés. Les structures évoluant moins vite que les séquences, il est possible de trouver des similarités en structure qui ont été effacées en séquence. Les applications seraient de regrouper les structures en familles pour pouvoir prédire les structures à partir des séquences, et mieux comprendre leur évolution.

De nombreux algorithmes ont été développés et adaptés afin de prendre en compte les spécificités des comparaisons de séquences et de structures. Je vais détailler quelques programmes parmi les plus utilisés.

III.A.1 Comparaison de séquences

Plusieurs algorithmes ont été développés pour comparer des séquences. Les algorithmes de référence sont celui de Needleman et Wunsch (Needleman and Wunsch, 1970) pour chercher la similarité globale entre deux séquences et celui de Smith et Waterman (Smith and Waterman, 1981) pour chercher des similarités locales entre deux séquences (détails ci-dessous). Des heuristiques ont ensuite été développées pour chercher rapidement des similarités entre une séquence et une banque, c'est le cas de Blast (Altschul et al., 1990) et Fasta (Pearson, 1990).

Certains programmes se sont spécialisés dans les alignements multiples, comme Clustal (Larkin et al., 2007), Multalign (Corpet, 1988) ou Muscle (Edgar, 2004).

D'autres programmes cherchent les répétitions dans des génomes complets, afin de mieux comprendre les événements de duplication à l'échelle du génome. C'est ce que font par exemple Repseek (Achaz et al., 2007), Reputer (Kurtz and Schleiermacher, 1999), Tandem Repeat Finder (Benson, 1999) ou MUMmer (Delcher et al., 2003).

Cependant, nous nous intéresserons ici seulement aux alignements entre deux séquences, et ceux-ci peuvent être adaptés à la recherche de répétitions internes.

III.A.1.a L'algorithme de Smith et Waterman

L'algorithme de Smith et Waterman (Smith and Waterman, 1981) est un algorithme qui utilise la programmation dynamique pour trouver le sous-alignement local optimal entre deux séquences. Il est inspiré de l'algorithme de Needleman et Wunsh (Needleman and Wunsch, 1970) qui calcule l'alignement global optimal entre

deux séquences. Le principe de la programmation dynamique est de résoudre un problème en trouvant les solutions optimales de ses sous-problèmes.

Le but de l'algorithme de Smith et Waterman est de trouver la paire de segments, provenant de deux séquences, telle qu'il n'y ait aucune autre paire de segments ayant plus de similarités entre elles. Pour trouver la paire de segments de plus haut score, il faut chercher l'alignement qui minimise le nombre de substitutions, insertions et délétions.

Dans ce chapitre, les notations suivantes seront utilisées pour tous les algorithmes :

A et B sont les séquences à aligner, de tailles respectives m et n : A= a_1 a_2 ... a_m et B = b_1 b_2 …b_n.

H est la matrice d'alignement, E et F sont les matrices d'extension de gap.

q est la pénalité d'ouverture de gap et r est la pénalité d'extension de gap, et le poids de gap total est : $W_k = q+kr$; k étant la taille du gap.

s(a,b) est le poids de la substitution de a par b.

Pour calculer le score de la paire de segments la plus similaire, une matrice H de taille m x n sera remplie de telle façon que chaque case $H_{i,j}$ contienne le score maximum de deux segments se finissant respectivement à a_i et b_j. Les matrices sont remplies à partir des formules suivantes, et les chemins empruntés sont mémorisés au fur et à mesure :

$$H_{(i,j)} = \begin{cases} \max\begin{Bmatrix} 0 \\ H_{(i-1,j-1)} + s_{(a,b)} \\ E_{(i,j)} \\ F_{(i,j)} \end{Bmatrix} & si\ i>0\ et\ j>0 \\ 0 & si\ i=0\ ou\ j=0 \end{cases}$$

$$E_{(i,j)} = \begin{cases} \max\begin{Bmatrix} E_{(i-1,j)} - r \\ H_{(i-1,j)} - q - r \end{Bmatrix} & si\ i>0\ et\ j>0 \\ -q & si\ i=0\ et\ j>0 \end{cases}$$

$$F_{(i,j)} = \begin{cases} \max\begin{Bmatrix} F_{(i,j-1)} - r \\ H_{(i-1,j)} - q - r \end{Bmatrix} & si\ i>0\ et\ j>0 \\ -q & si\ i>0\ et\ j=0 \end{cases}$$

Équation 2

Cette formule remet le score à 0 lorsqu'il devient négatif, ce qui réamorce l'alignement et permet de chercher une similarité locale. Concernant les matrices de gap, des vecteurs suffisent pour retenir l'information des extensions de gaps, cependant des matrices seront nécessaires pour les algorithmes suivant et sont présentées ici.

La case du début de l'alignement optimal est notée $H_{I,J}$ et la case de la fin de cet alignement $H_{I',J'}$.

Le meilleur alignement se finit à la case de $H_{I',J'}$ qui est la case où le score est le plus élevé. Ensuite, il faut remonter la matrice par le chemin conduisant à ce score maximal jusqu'à ce que le score soit inférieur ou égal à 0 : la case précédant le 0 correspond au début de l'alignement. Cet algorithme permet donc de trouver à la fois les sous séquences qui ont la plus grande similarité, mais aussi leur alignement.

Pour chercher les segments suivants les plus similaires, il faut trouver le 2nd meilleur score de la matrice H non associé au 1er alignement, puis appliquer la procédure qui consiste à remonter la matrice pour trouver le début de l'alignement. Il est nécessaire de recalculer la région de la matrice qui contient le 1er alignement pour s'affranchir de l'influence de cet alignement sur la matrice. Ce calcul est proposé par Waterman et Eggert (Waterman and Eggert, 1987) et sera décrit dans la section suivante.

Pour chercher les répétitions internes, il faut supprimer la diagonale (qui correspond à l'alignement de la séquence contre elle même) et ne calculer que la moitié de la matrice d'alignement (qui est symétrique) (Figure 12).

		A	D	Q	R	T	A	L	M	Q	K	T	A
	0	0	0	0	0	0	0	0	0	0	0	0	0
A	0	0	0	0	0	0	4	0	0	0	0	0	4
D	0	0	0	0	0	0	0	0	0	0	0	0	0
Q	0	0	0	0	1	0	0	0	0	5	1	0	0
R	0	0	0	0	0	0	0	0	0	1	7	0	0
T	0	0	0	0	0	0	0	0	0	0	0	12	4
A	0	0	0	0	0	0	0	0	0	0	0	4	16
L	0	0	0	0	0	0	0	0	2	0	0	0	8
M	0	0	0	0	0	0	0	0	0	2	0	0	0
Q	0	0	0	0	0	0	0	0	0	0	3	0	0
K	0	0	0	0	0	0	0	0	0	0	0	2	0
T	0	0	0	0	0	0	0	0	0	0	0	0	2
A	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 12 : Schéma de l'algorithme de Smith et Waterman adapté aux répétitions internes. La case de plus haut score est en rouge, l'alignement est en orange.

III.A.1.b L'algorithme de Waterman et Eggert

L'algorithme de Waterman et Eggert (Waterman and Eggert, 1987) permet le calcul des meilleurs alignements successifs indépendants sans recalculer toute la matrice H. Pour calculer le 2ème meilleur alignement et les suivants, il faut s'affranchir des cases de la matrice qui sont liées aux chemins des alignements précédents par des substitutions ou des gaps.

Pour cela, il suffit de recalculer la partie de la matrice qui a changé depuis le précédent alignement. Si le 1er alignement est entre les fragments $a_I \ldots a_{I'}$ et $b_J \ldots b_{J'}$, il suffit de recalculer au maximum le carré compris entre $H_{I,J}$ et $H_{m,n}$ (coté inférieur droit de la matrice). Les cases situées à gauche et au dessus du point de départ du dernier alignement n'auront pas changé. La nouvelle matrice est notée H*. Le calcul part donc du point de départ du dernier alignement $H_{I,J}$. Le score est recalculé :

$$H^*_{i,j} = \max\begin{Bmatrix} 0 \\ E_{i,j} \\ F_{i,j} \end{Bmatrix}$$

Équation 3

La similarité ou substitution correspondant au précédent alignement n'est donc pas autorisée. Il est ensuite possible de recalculer les valeurs de la ligne (i, j), pour tout I<i<n et pour tout J<j<m, jusqu'à ce que la valeur de $H^*_{i,j}$ soit identique à $H_{i,j}$. Si

$H^*_{i,j} = H_{i,j}$, il est évident que toutes les valeurs suivantes de la ligne seront ensuite identiques. Il est possible de continuer ainsi pour les autres colonnes. Pour chaque colonne, il faut aller au moins aussi loin que pour la colonne précédente. En prenant en compte les extensions de gap, il faut recalculer la matrice H* jusqu'à ce que les 3 conditions suivantes soient satisfaites :

$$\begin{cases} H^*_{i,j} = H_{i,j} \\ E^*_{i,j} = E_{i,j} \\ F^*_{i,j} = F_{i,j} \end{cases}$$

Équation 4

III.A.2 Comparaison de structures

Plusieurs méthodes sont utilisées pour comparer les structures. Je présenterai ici cinq méthodes qui donnent de bons résultats (Novotny et al., 2004) et sont parmi les plus utilisées.

III.A.2.a DALI

La méthode DALI a été proposée par Holm et Sander (Holm and Sander, 1993). Chacune des deux structures est représentée par une matrice de distances internes contenant toutes les distances entre tous les couples de Cα de chaque structure. Les matrices sont ensuite divisées en sous matrices chevauchantes de taille six. Seules les 40 000 meilleures paires de sous-matrices sont conservées, puis les résidus correspondant des deux protéines sont alignés. Les sous-matrices de taille six sélectionnées sont divisées en trois sous-matrices de taille quatre chevauchantes, afin de trouver la meilleure série de matrices correspondantes entre les deux protéines. Ensuite les sous-matrices des deux structures sont comparées pour trouver des sous-matrices similaires avec la méthode de Monte-Carlo. Comme il serait trop long de calculer toutes les combinaisons de sous-matrices, cette méthode permet d'explorer aléatoirement l'espace des solutions. Le but est de trouver la meilleure série de sous-matrices de taille quatre similaire entre les deux protéines, et qui maximise un score de similarité structurale défini en terme d'équivalence des distances intra-moléculaires. Le résultat est donc une correspondance entre les résidus de deux tétramatrices et non un alignement. L'algorithme est initié avec plusieurs points de départ, c'est à dire avec plusieurs correspondances entre les résidus. Finalement, la correspondance optimale entre les

deux structures est affinée en enlevant 30% des correspondances et en réitérant l'algorithme à partir de ces nouveaux points de départ.

Le programme est mis à disposition sur le site http://ekhidna.biocenter.helsinki.fi/dali_server/ et peut être téléchargé. Il a servi pour la classification FSSP (Families of Strucurally Similar Proteins). Il est parmi ceux qui donnent les meilleurs résultats (Novotny et al., 2004).

III.A.2.b CE

La méthode CE a été publiée par Shindyalov et Bourne (Shindyalov and Bourne, 1998). Elle se déroule en deux étapes : la recherche de fragments similaires de 8 résidus, puis l'assemblage de ces fragments.

Lors de la 1ère étape, les fragments similaires de huit résidus contigus sont recherchés. Deux fragments sont considérés comme similaires si la moyenne des différences des distances internes est inférieure à 3Å : ils sont appelés AFP (Aligned Fragment Pair). Dans un deuxième temps, ces fragments similaires sont assemblés si plusieurs conditions sont respectées : le nouveau fragment ne doit être chevauchant avec aucun des fragments déjà présents, il doit être contigu avec un AFP déjà présent sur au moins une protéine, les gaps éventuels doivent être de taille inférieure à 30 résidus, la moyenne des distances internes entre toutes les AFP réunies, y compris la nouvelle, doit être inférieure à 4Å.

Parmi les 20 meilleurs alignements obtenus, seul celui qui a le meilleur RMSD est conservé. Il est ensuite affiné entre autre par une procédure de superposition-alignement de type Needleman et Wunsch (Needleman and Wunsch, 1970). Un Z-score est également calculé. Il est disponible sur le site http://cl.sdsc.edu/ce.html et peut être utilisé en ligne ou téléchargé.

III.A.2.c VAST

La méthode VAST (Gibrat et al., 1996) est basée sur la théorie des graphes. Les structures secondaires des protéines sont représentées par un nœud sur le graphe. Les nœuds sont reliés entre eux si les structures secondaires des deux protéines se ressemblent suffisamment. Il faut ensuite rechercher le sous-graphe le plus grand, tel que chaque nœud du sous-graphe est connecté à un autre nœud du sous-graphe et qu'il ne soit pas inclus dans un autre sous-graphe qui ait la même propriété. Cette méthode

permet aussi de calculer la significativité des alignements trouvés. La p-value calculée est la probabilité que ce score soit obtenu par hasard en dessinant aléatoirement les paires de structures secondaires à partir de la banque multiplié par le nombre d'alignements sub-structuraux alternatifs possibles pour une paire de structures donnée. VAST peut être téléchargé à http://mig.jouy.inra.fr/logiciels/vast.

III.A.2.d MATRAS

Matras (Kawabata and Nishikawa, 2000) est basé sur le principe de matrices de substitution analogues à celle de Dayhoff (Dayhoff et al., 1978). Ces matrices sont calculées à l'aide d'un modèle de transition markovien. Le score de substitution adopté est le suivant :

$$S_{i,j} = \log \frac{P(i \to j)}{p(j)}$$

Équation 5

$P(i \to j)$ est la probabilité que l'état i change vers l'état j au cours de l'évolution, $p(j)$ est la probabilité que l'état j apparaisse par hasard. Une matrice de probabilités de transition est calculée à partir d'alignements de structures homologues, sélectionnées en fonction de leur similarité de séquence. Il y a trois types de scores : un score SSE sur les changements de structures secondaires, un score environnement sur l'état des structures secondaires (enfoui ou exposé au solvant) (SSE), et un score de distances basé sur les distances internes entre les résidus.

Ensuite, un premier alignement est effectué sur les scores SSE et d'environnement, puis à partir de cet alignement, plusieurs autres alignements sont effectués par programmation dynamique à partir des scores de distance, jusqu'à convergence.

MATRAS est disponible à l'adresse http://biunit.aist-nara.ac.jp/matras/. Il permet de faire à la fois des comparaisons entre deux structures, des comparaisons multiples de structures, et des recherches de répétitions internes.

III.A.2.e YAKUSA

Yakusa (Carpentier et al., 2005) cherche les similarités entre une structure et une banque. Les structures sont codées en angles α entre –180° et +180° (angle dièdre entre 4 Cα successifs, cf. III.B.2), ce qui permet de les représenter par une suite de symboles.

Les angles α de la structure requête sont rangés dans un automate par groupe de k angles successifs chevauchants, avec leur position. L'automate contient aussi les motifs similaires avec des angles α proches. Ensuite, pour chaque structure de la banque, l'automate est parcouru à la recherche de motifs communs aux deux structures, les graines. Ensuite les graines sont sélectionnées et étendues en segments structuraux les plus longs possibles : les SHSP (Structural High Scoring Pairs). Les SHSP compatibles sont ensuite sélectionnés et un score est calculé.

Il est utilisable en ligne ou téléchargeable à l'adresse suivante : http://bioserv.rpbs.jussieu.fr/Yakusa/index.html. Il donne de bons résultats et est plus rapide : moins d'une minute pour comparer une structure contre une banque, contre au moins 5 minutes pour DALI (Novotny et al., 2004).

III.B L'algorithme utilisé dans Swelfe

III.B.1 L'algorithme SIM

L'algorithme SIM de comparaison de séquences, proposé par Huang et Miller (Huang and Miller, 1991; Huang et al., 1990), est utilisé lorsque la taille des séquences est plus limitante que le temps de calcul.

L'algorithme classique de Smith et Waterman (Smith and Waterman, 1981) mémorise toute la matrice d'alignement (de taille m x n) pour retrouver le chemin suivi par le meilleur alignement. La mémoire utilisée pour mémoriser les extensions de gaps est importante si l'algorithme de Waterman et Eggert (Waterman and Eggert, 1987) est utilisé. Lorsque les séquences à comparer sont très longues, c'est souvent la taille des séquences et donc la mémoire, plus que le temps de calcul, qui est le facteur limitant. L'algorithme SIM permet de résoudre ce problème. Pour notre jeu de données, seule une séquence d'ADN très longue, celle de la titine (plus de 10 000 nt), posait des problèmes de mémoire.

L'algorithme se déroule en trois étapes résumées sur la Figure 13 qui seront détaillées ensuite :

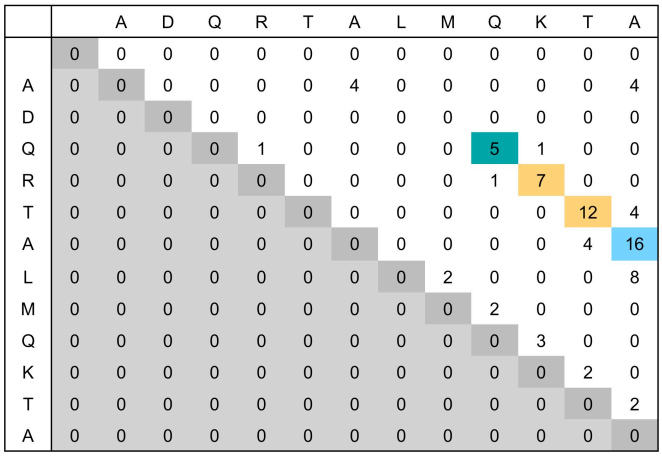
- **Trouver la case de plus haut score dans la matrice d'alignement.** Au fur et à mesure du calcul de la matrice d'alignement, une ligne et une case sont mémorisées à chaque étape. Le meilleur score et sa position

sont gardés en mémoire. À la fin du calcul de la matrice, le meilleur score retenu correspond à la fin du meilleur alignement.

- **Trouver le point de départ de l'alignement correspondant au plus haut score.** Pour cela, une partie de la matrice est recalculée dans le sens inverse à partir de la case de plus haut score.
- **Refaire l'alignement sur la portion de la matrice entre le point de départ et le point d'arrivée pour connaître le meilleur chemin.** L'algorithme de Myers et Miller (Myers and Miller, 1988) est utilisé.

		A	D	Q	R	T	A	L	M	Q	K	T	A
	0	0	0	0	0	0	0	0	0	0	0	0	0
A	0	0	0	0	0	0	4	0	0	0	0	0	4
D	0	0	0	0	0	0	0	0	0	0	0	0	0
Q	0	0	0	0	1	0	0	0	0				
R	0	0	0	0	0								
T	0	0	0	0	0	0							
A	0	0	0	0	0	0	0						
L	0	0	0	0	0	0	0	0					
M	0	0	0	0	0	0	0	0	0				
Q	0	0	0	0	0	0	0	0	0	0			
K	0	0	0	0	0	0	0	0	0	0	0		
T	0	0	0	0	0	0	0	0	0	0	0	0	
A	0	0	0	0	0	0	0	0	0	0	0	0	0

		A	D	Q	R	T	A	L	M	Q	K	T	A
	0	0	0	0	0	0	0	0	0	0	0	0	0
A	0	0	0	0	0	0	4	0	0	0	0	0	4
D	0	0	0	0	0	0	0	0	0	0	0	0	0
Q	0	0	0	0	1	0	0	0	0	5	1	0	0
R	0	0	0	0	0	0	0	0	0	1	7	0	0
T	0	0	0	0	0	0	0	0	0	0	0	12	4
A	0	0	0	0	0	0	0	0	0	0	0	4	16
L	0	0	0	0	0	0	0	0	2	0	0	0	8
M	0	0	0	0	0	0	0	0	0	2	0	0	0
Q	0	0	0	0	0	0	0	0	0	0	3	0	0
K	0	0	0	0	0	0	0	0	0	0	0	2	0
T	0	0	0	0	0	0	0	0	0	0	0	0	2
A	0	0	0	0	0	0	0	0	0	0	0	0	0



		A	D	Q	R	T	A	L	M	Q	K	T	A
	0	0	0	0	0	0	0	0	0	0	0	0	0
A	0	0	0	0	0	0	4	0	0	0	0	0	4
D	0	0	0	0	0	0	0	0	0	0	0	0	0
Q	0	0	0	0	1	0	0	0	0	5	1	0	0
R	0	0	0	0	0	0	0	0	0	1	7	0	0
T	0	0	0	0	0	0	0	0	0	0	0	12	4
A	0	0	0	0	0	0	0	0	0	0	0	4	16
L	0	0	0	0	0	0	0	0	2	0	0	0	8
M	0	0	0	0	0	0	0	0	0	2	0	0	0
Q	0	0	0	0	0	0	0	0	0	0	3	0	0
K	0	0	0	0	0	0	0	0	0	0	0	2	0
T	0	0	0	0	0	0	0	0	0	0	0	0	2
A	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 13 : Schéma de l'algorithme SIM.

Le détail de chaque étape est expliqué ci-dessous. Les notations utilisées sont les même que pour l'algorithme de Smith et Waterman (équation 2).

$$H_{(i,j)} = \begin{cases} \max\left\{ \begin{matrix} 0 \\ H_{(i-1,j-1)} + s_{(a,b)} \\ E_{(i,j)} \\ F_{(i,j)} \end{matrix} \right. & si\ i > 0\ et\ j > 0 \\ 0 & si\ i = 0\ ou\ j = 0 \end{cases}$$

$$E_{(i,j)} = \begin{cases} \max\left\{ \begin{matrix} E_{(i-1,j)} - r \\ H_{(i-1,j)} - q - r \end{matrix} \right. & si\ i > 0\ et\ j > 0 \\ -q & si\ i = 0\ et\ j > 0 \end{cases}$$

$$F_{(i,j)} = \begin{cases} \max\left\{ \begin{matrix} F_{(i,j-1)} - r \\ H_{(i-1,j)} - q - r \end{matrix} \right. & si\ i > 0\ et\ j > 0 \\ -q & si\ i > 0\ et\ j = 0 \end{cases}$$

Équation 6

Le meilleur alignement local finit à (I', J') tel que :

$$H_{(I',J')} = \max\left\{ H_{(i,j)} : 1 \le i \le m\ et\ 1 \le j \le n \right\}$$

Équation 7

- *Étape 1*

L'objectif est de trouver la case de meilleur score, il suffit donc de mémoriser les lignes récemment calculées de H et E (Figure 14). Si les lignes i-1 de H et E sont mémorisées dans les vecteurs HH et EE, il suffit d'écraser ces valeurs au fur et à mesure du calcul de la ligne i. Il y a aussi besoin de trois scalaires : f, h, p définis ci-après. f

contient le poids d'extension de gap, il n'y a pas besoin de vecteur car les valeurs de F ne dépendent que de la ligne i et non de la ligne i-1. h contient le nouveau score qui a été calculé. p contient la case $H_{(i-1, j-1)}$ qui est nécessaire pour calculer le score suivant mais qui va être effacée dans HH.

A chaque étape (i, j>0), on a :

$$HH_{(k)} = \begin{cases} H_{(i,k)} \\ H_{(i-1,k)} \end{cases} \begin{matrix} si\ k < j \\ si\ k \geq j \end{matrix}$$

$$EE_{(k)} = \begin{cases} E_{(i,k)} \\ E_{(i-1,k)} \end{cases} \begin{matrix} si\ k < j \\ si\ k \geq j \end{matrix}$$

$$f = F_{(i,j-1)}$$

$$h = H_{(i,j-1)}$$

$$p = H_{(i-1,j-1)}$$

Équation 8

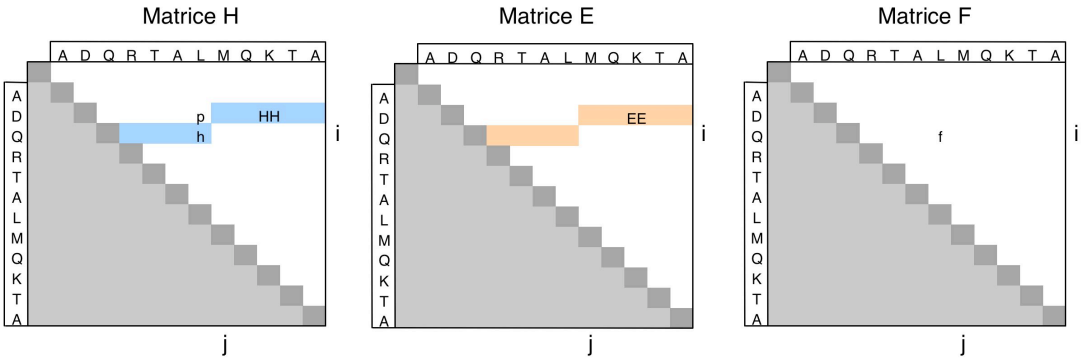


Figure 14 : Matrices H, E et F de l'alignement.
HH est indiqué en bleu et EE en orange.

- *Étape 2*

Une fois la case de fin du meilleur alignement trouvée $H_{I',J'}$, il faut trouver le début de l'alignement $H_{I,J}$. Pour cela, la matrice est recalculée à l'envers en partant de la case de plus haut score. Cependant, la méthode utilisée n'est pas symétrique de la méthode employée précédemment. Plusieurs meilleurs alignements pourraient exister, et une case avec un haut score dans l'algorithme inverse pourrait être le point de départ d'un alignement qui ne finit pas à $H_{I',J'}$. Pour être certain de trouver le bon alignement, les valeurs initiales de H, E, F sont mises à –1 sauf pour la valeur de H à l'origine (p=0 pour i=I'). De plus, 0 est retiré des solutions possibles dans le calcul de H. Dans ce cas, il est possible de s'arrêter quand le coût est supérieur ou égal au score trouvé précédemment.

Une autre solution serait de retenir le point de départ de l'alignement lors du 1er calcul de H, cependant, cette méthode est moins rapide lorsque la longueur des alignement est bien inférieure à la longueur des séquences.

- *Étape 3*

Il est ensuite possible de recalculer l'alignement entre les points de départ et d'arrivée avec l'algorithme de Myers et Miller (Myers and Miller, 1988) (description ci-après). Cette méthode permet d'utiliser le moins de mémoire possible lors du calcul de l'alignement global de deux séquences. Elle minimise la distance au lieu de maximiser le score, quelques arrangements sont donc nécessaires :

$$w_{(a,b)} = s_{\max} - s_{(a,b)}$$
$$g = q$$
$$h = r + \frac{1}{2} s_{\max}$$

Équation 9

$w_{(a,b)}$ est le poids de substitution, g est le poids d'ouverture de gap, et h le poids d'extension.

- *Étape 4*

Pour trouver plusieurs meilleurs alignements successifs, la même méthode est utilisée sauf que les paires déjà alignées sont supprimées des alignements possibles : toutes les positions utilisées lors des précédents alignements sont gardées en mémoire (Waterman and Eggert, 1987). Il est possible d'optimiser le temps et l'espace grâce à l'amélioration proposée par Huang et Miller (Huang and Miller, 1991) et qui sera présentée ci-après.

- *Étape 5*

Pour calculer les répétitions internes, la diagonale est supprimée pour ne pas trouver l'identité, et seulement la moitié de la matrice est calculée car elle est symétrique.

L'algorithme de Myers et Miller (Myers and Miller, 1988) s'inspire de l'algorithme d'Hirschberg (Hirschberg, 1975). Le meilleur alignement passe par la case de plus haut score de la colonne du milieu de la matrice. Pour utiliser moins de

mémoire, il est possible de calculer successivement toutes les meilleures cases du milieu en divisant la matrice en deux à chaque étape. Au début de l'algorithme, la séquence A est divisée en deux (i=m/2), puis il faut chercher dans la séquence B le résidu qui permet d'aligner au mieux B avec $A_{0,i}$. Ainsi toutes les paires de résidus sont alignées récursivement, de part et d'autre du point central, jusqu'à ce que tous les résidus de chaque moitié de la séquence A soient alignés avec la séquence B.

Les différentes étapes sont présentées Figure 15.

1) L'alignement est recalculé entre $H_{I,J}$ et $H_{I',J'}$.
2) La séquence A est divisée en deux à l'indice i (i=m/2)
3) Pour tous les indices j de la séquence B, les scores d'alignement de $A_{1,i}$ avec $B_{1,j}$ de la colonne i (vecteur de longueur n)sont stockés. Pour chaque case de cette colonne, les scores des chemins finissant par un appariement, une délétion ou une insertion sont mémorisés.
4) De la même façon, les scores de l'alignement de A_{i+1}, m avec $B_{j+1,n}$ sont mémorisés. Le calcul est fait en partant de H[m,n] et en remontant la matrice.
5) Le chemin optimal pour aligner les séquences passe forcément par la colonne i. Il est possible de le déterminer en recherchant le minimum parmi toutes les sommes entre une distance de l'étape 2 et une distance de l'étape 3 (cf. Figure 15). Lorsqu'il y a une délétion, il ne faut pas compter 2 fois la pénalité de gap.
6) Ayant ainsi localisé les bornes des deux sous-alignements conduisant à l'alignement optimal ($A_{1,i}$,$B_{1,j}$ et $A_{i+1,m}$ $B_{j+1,n}$), Il faut mémoriser que la lettre A_i de la meilleure case est alignée avec la lettre B_j. Les étapes 1 à 4 sont recommencées pour les sous séquences $A_{1,i}$ et $B_{1,j}$ ainsi que pour $A_{i+1,m}$ et $B_{j+1,n}$. Si les résidus A_i et A_{i+1} sont dans un gap, les étapes 1 à 4 sont recommencées pour les sous-séquences $A_{1,i-1}$ et $B_{1,j}$ et pour $A_{i+1,m}$ et $B_{j,n}$.

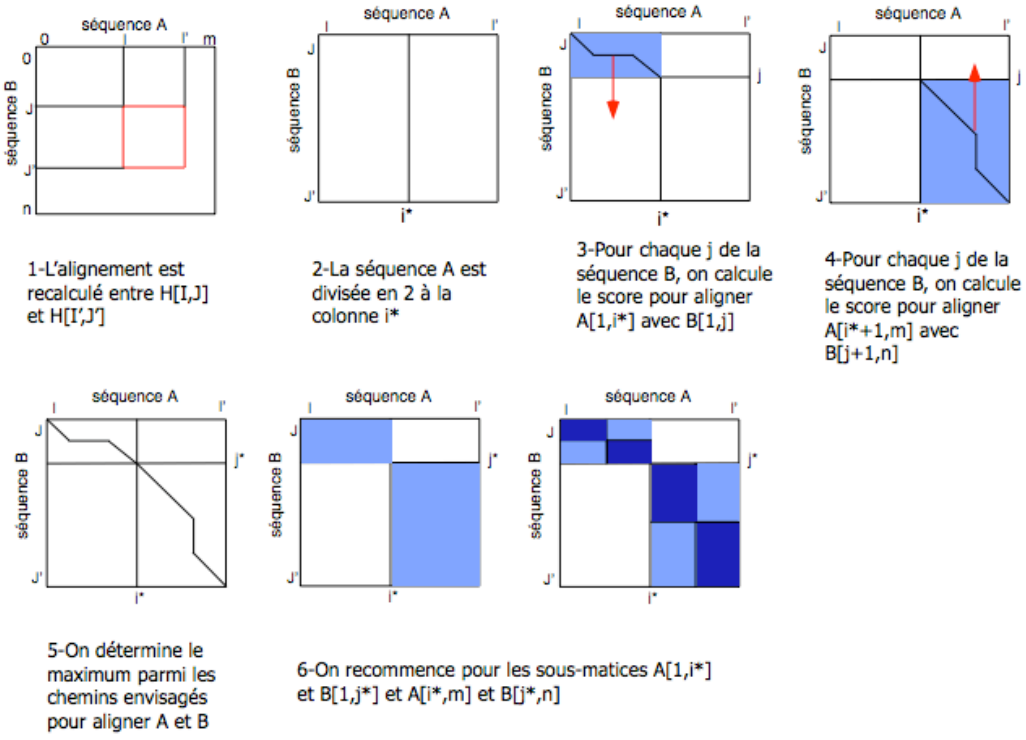


Figure 15 : Schéma de l'algorithme de Myers et Miller.

Pour trouver les k meilleurs alignements, Huang et Miller (Huang and Miller, 1991) proposent une méthode plus rapide et qui prend moins d'espace que celle proposée par Waterman et Eggert (Waterman and Eggert, 1987). Le principe est de mémoriser les k meilleurs scores pendant le calcul de la matrice H. Ces scores doivent provenir de chemins d'alignement disjoints, c'est à dire qui ne se chevauchent pas. Après avoir calculé le 1er meilleur alignement, il faut trouver les chemins qui ont un bon score et qui ont été cachés par le 1er alignement. Pour chaque alignement, l'algorithme va effectuer une 1ère phase de programmation dynamique pour déterminer la zone d'influence de l'alignement précédent à l'intérieur d'une région limitée de la matrice. Ensuite, une 2ème phase permet de recalculer les scores de cette région et de mettre à jour la liste de meilleurs scores.

L'algorithme de Smith et Waterman a une complexité en O (N^2) à la fois en temps et en mémoire, N étant la taille de la séquence. En utilisant l'algorithme de Waterman et Eggert (Waterman and Eggert, 1987) pour trouver les K meilleurs alignements, la complexité en mémoire est la même : O(N^2), la complexité en temps dans le pire des cas est en O ($N^2 \times K$) mais en moyenne est plus faible. L'algorithme proposé par Huang et Miller (Huang and Miller, 1991; Huang et al., 1990) a une

complexité en O(N + K) pour la mémoire et O($N^2 + \sum_{n=1}^{k} L_n^2$) pour le temps, K étant la somme des longueurs des k meilleurs alignements, et L_n est la longueur du $n^{ième}$ alignement.

III.B.2 Adaptation de l'algorithme pour chercher les répétitions aux trois niveaux

III.B.2.a Pourquoi utiliser cet algorithme ?

A notre connaissance, il n'existe aucun algorithme de recherche de répétitions internes utilisable à la fois dans les séquences et les structures des protéines, et à part MATRAS, il n'y a pas d'algorithme qui cherche des répétitions à l'intérieur d'une même structure. De plus, les algorithmes de comparaison de structures sont généralement assez lents à part Yakusa (Carpentier et al., 2005), ils ont besoin en moyenne de 5 à 20 minutes pour comparer une structure à une banque.

L'algorithme SIM est très rapide pour trouver les répétitions contenues dans une séquence, et nous l'avons adapté aux structures en utilisant les angles α.

III.B.2.b Les structures sont décrites par leurs angles α

Pour utiliser l'algorithme SIM sur des structures 3D, il est nécessaire de coder les structures de façon linéaire. Nous avons utilisé les angles α qui ont été utilisés précédemment par Carpentier *et al*. (Carpentier et al., 2005) et qui ont été décrits en premier par Levitt *et al*. (Levitt, 1976; Levitt and Warshel, 1975). L'angle α est l'angle dièdre défini par quatre carbones α successifs et l'angle τ est l'angle plan défini par trois carbones alpha successifs (cf. Figure 16).

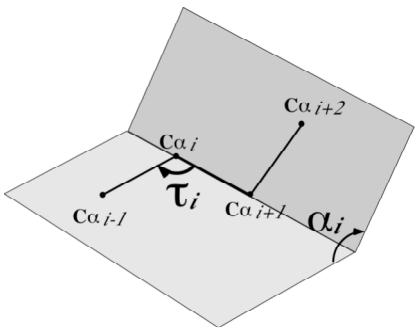


Figure 16 : Angles α et τ.

L'angle α_i est associé au 2ème des quatre carbones α et l'angle τ_i est l'angle formé par les trois premiers carbones alpha.

L'angle τ varie peu autour de 106° donc l'angle α suffit à décrire le squelette carboné. Mathilde Carpentier a vérifié dans sa thèse (Carpentier, 2005) qu'en reconstruisant le squelette carboné d'un fragment d'une protéine à partir de ses angles α discrétisés, il se superpose bien avec le fragment initial.

III.C Systèmes de score et pénalités de gaps

Les scores et les pénalités de gap sont récapitulés dans le Tableau 1. Ils ont été adaptés à chaque niveau d'analyse.

III.C.1 Les Séquences

III.C.1.a Les scores

Pour les séquences nucléiques, un poids positif est généralement utilisé pour les identités, et un poids négatif pour les substitutions. Cependant, ces poids peuvent biaiser l'alignement si les fréquences des nucléotides sont éloignées de 25%. Pour parer à ce problème, nous avons utilisé un score décrit par Achaz *et al*. (Achaz et al., 2007) et qui prend en compte la fréquence des nucléotides. Ainsi, si une séquence est riche en adénine, une répétition riche en adénine aura un poids moindre.

Pour les séquences d'acides aminés, les matrices PAM (Dayhoff et al., 1978) ou BLOSUM (Henikoff and Henikoff, 1992) sont couramment employées et sont basées sur des alignements multiples de séquences. Nous utilisons par défaut la matrice BLOSUM62, mais le programme Swelfe est utilisable avec n'importe quelle matrice de substitution.

III.C.1.b Les pénalités de gap

Les pénalités de gaps par défaut pour les séquences nucléiques sont de –4 pour l'ouverture et de –1 pour l'extension, ce qui correspond à peu près à quatre identités pour l'ouverture, et une identité pour l'extension, et ce qui est conseillé par Achaz *et al*. (Achaz et al., 2007). Pour les acides aminés, la pénalité est de –8 pour l'ouverture de gap et de –3 pour l'extension, ce qui correspond environ à une ou deux identités pour l'ouverture et moins d'une identité pour l'extension.

III.C.2 Les structures

III.C.2.a Les scores testés

De nombreux scores ont été testés pour les structures. Il existe à peu près autant de systèmes de score que de méthodes de comparaison de structure. Nous avons au début de ce travail testé des scores déjà publiés pour la comparaison de deux structures, pour en trouver un qui serait adapté à Swelfe.

Nous avons programmé le score décrit par Gerstein et Levitt (Gerstein and Levitt, 1998) qui a la forme suivante :

$$S = M\left(\sum_{i,j}\frac{1}{1+\left(\frac{d_{a_i,b_j}}{d_0}\right)^2} - \frac{N_{gap}}{2}\right)$$

Équation 10

M est le score maximum d'un match, choisi arbitrairement à 20, d_0 est la distance à laquelle la similarité baisse à la moitié de sa valeur maximale et vaut 5Å, i et j sont les paires de résidus alignés, $d_{ai,bj}$ est la distance entre les Cα de ces deux résidus et N_{gap} est le nombre de gaps.

Nous avons également programmé le TM score (Zhang and Skolnick, 2004) qui est issu du score de Gerstein-Levitt:

$$TM\text{-}score = Max\left[\frac{1}{L_N}\sum_{i=1}^{L_T}\frac{1}{1+\left(\frac{d_i}{d_0}\right)^2}\right]$$

Équation 11

L_N est la taille de la structure native, L_T est le nombre de résidus alignés, d_i est la distance entre la $i^{ème}$ paire de résidus alignés, d_0 est une échelle pour normaliser la différence de match. Max représente la valeur maximale après meilleur alignement dans l'espace.

Avec le paramètre :

$$d_0 = 1{,}24\sqrt[3]{L_N - 15} - 1{,}8$$

Équation 12

le TM-score devient indépendant de la longueur de la protéine.

Cependant ces scores n'étaient pas adaptés à la matrice d'alignement car ils doivent être calculés après alignement optimal, et ne peuvent donc pas servir à construire l'alignement.

De plus, nous souhaitions trouver un score qui dépende des angles α (description linéaire de la séquence) permettant de trouver des duplications. Il fallait donc trouver un score qui puisse donner moins d'importance aux structures secondaires hélices α et feuillets β qui sont communs et très ressemblants du point de vue structural, car ce sont des structures stables et faciles à former, et de ce fait leur ressemblance n'est pas forcément due à une origine évolutive commune. Cette démarche se rapproche du calcul des matrices de type BLOSUM (Henikoff and Henikoff, 1992) ou PAM (Dayhoff et al., 1978) qui sont issues d'alignements multiples.

Nous avions au début utilisé un score (équation 13) qui permet de prendre en compte la différence entre les angles i et j, et donc la similarité structurale d'un groupe de quatre Cα entre les deux structures. $|\Delta angle|$ représente la différence angulaire entre les deux angles α (entre 0 et 180°). La valeur 30° a été choisie car elle permet, en prenant deux angles au hasard dans la PDB d'obtenir une valeur positive dans 25% des cas.

$$S_{i,j} = 30° - |\Delta angle|$$

Équation 13

Nous avons également essayé de créer une matrice de type log-odd en partant d'alignements structuraux multiples calculés par Mathilde Carpentier avec Yakusa (Carpentier et al., 2005), avec la même méthode de calcul que celle utilisée pour créer la matrice BLOSUM (Henikoff and Henikoff, 1992). Cependant, ce score n'a pas donné les résultats escomptés dans la mesure où il était très sévère avec les angles les plus fréquents dans les alignements structuraux (hélices α, feuillets β), et il n'était presque plus possible de trouver ces angles dans les répétitions.

III.C.2.b Le score utilisé

Nous avons finalement créé un score qui permet de prendre en compte à la fois les angles α, et donc la similarité structurale, et la fréquence des angles α dans la PDB afin de minorer les scores des structures trop fréquentes.

$$S_{i,j} = 30° \times \left[(1 - p_i)(1 - p_j)(1 - \beta) + \beta \right] - \left| \Delta_{angle} \right| \quad avec \quad 0 \le \beta \le 1$$

Équation 14

avec $p_i = \frac{f_i - f_{\min}}{f_{\max} - f_{\min}} \qquad p_j = \frac{f_j - f_{\min}}{f_{\max} - f_{\min}}$

Équation 15

et $\beta = 0{,}4$

p_i et p_j sont les fréquences normalisées des angles α_i et α_j dans la PDB, entre 0 et 1 : f_i, f_j sont les fréquences des angles α_i et α_j, f_{max} est la fréquence de l'angle α le plus fréquent et f_{min} la fréquence de l'angle α le moins fréquent. Il y a 360 angles α possibles, donc 360 fréquences d'angles α. Le facteur $(1 - p_i)(1 - p_j)$ est donc élevé pour des angles peu fréquents et faible pour des angles très fréquents. Le facteur $30° \times \left[(1 - p_i)(1 - p_j)(1 - \beta) + \beta \right]$ a de ce fait une valeur d'autant plus grande que les angles sont peu fréquents.

En faisant varier le paramètre β du score, il est possible de donner plus ou moins d'importance aux angles les plus fréquents (cf. Figure 17). Si β=1, la fréquence des angles n'a aucune incidence sur le score alors que si β=0, la fréquence des angles détermine le score. Nous avons choisi de garder la valeur β=0,4 car avec cette valeur intermédiaire nous obtenons des répétitions proportionnellement un peu plus riches en hélices α que la moyenne de la PDB (ce qui est attendu car il est normal que les hélices se retrouvent dans les répétitions), mais moins riches qu'avec une valeur de β plus élevée, ce qui est ce que nous recherchons.

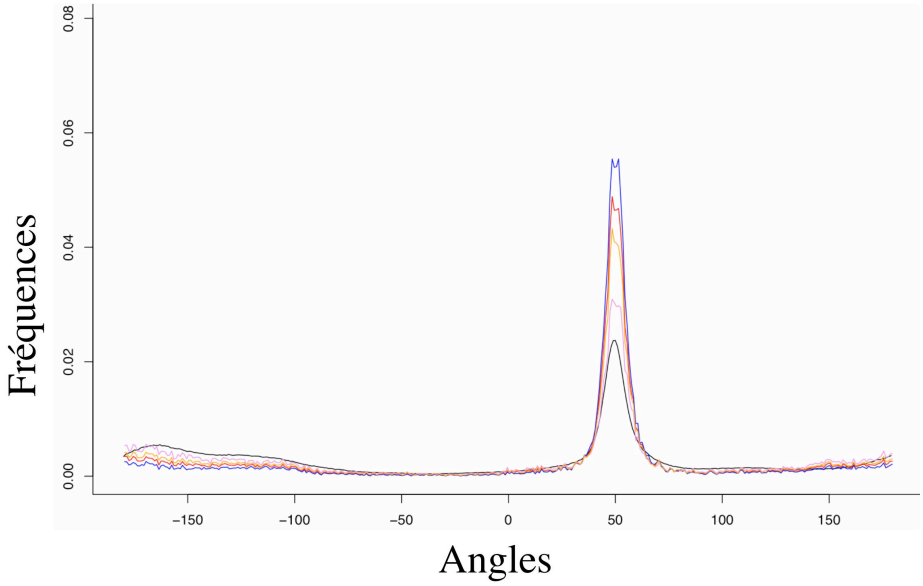


Figure 17 : Fréquences des angles α dans la PDB (en noir) et dans les répétitions.
Les valeurs suivantes du paramètre β sont utilisées: 0,7 (bleu), 0,5 (rouge), 0,4 (orange) et 0,3 (violet). Calculs effectués sur le jeu de données non redondant Cluster50 de la PDB. Pour chaque paramètre β, toutes les répétitions trouvées avec les valeurs par défaut du programme sont conservées et les angles α correspondant aux répétitions trouvées sont calculés. L'histogramme présente la fréquence des angles α présents dans les répétitions pour plusieurs paramètres β.

Notre score donne des résultats comparables à celui de Gerstein et Levitt (Gerstein and Levitt, 1998) (Figure 18) (test de corrélation de Pearson : $r = 0,85$, p-value = 2.10^{-16}).

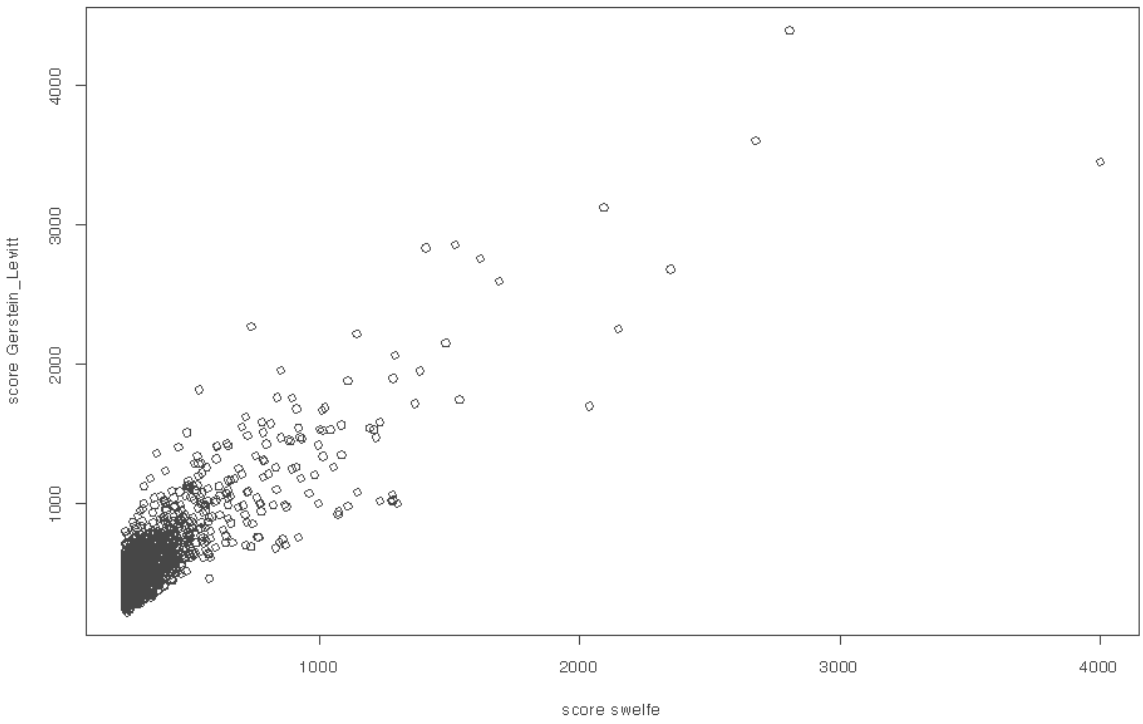


Figure 18 : Score de Swelfe vs. score de Gerstein Levitt.

III.C.2.c Les pénalités de gap

Pour les structures, les pénalités sont de 200° pour l'ouverture de gap et de 50° pour l'extension. Cela correspond à 6 ou 7 identités (le score de deux angles est positif si $\left|\Delta_{angle}\right|$ est inférieur à 30 pour des angles peu fréquents). Cela peut paraître beaucoup mais s'il y a trop de gaps, les structures s'alignent évidemment mal (Figure 19). Cette figure montre qu'en augmentant le poids d'ouverture de gap, le nombre de répétitions augmente, mais le nombre de protéines qui ont des répétitions n'augmente pas, et la taille et le RMSD des répétitions diminuent un peu : donc en fait il y a plus de répétitions par protéine et qui sont plus courtes, certaines répétitions sont donc coupées en deux. Le RMSD diminue un peu, ce qui peut être expliqué par la diminution de taille des répétitions. En augmentant les poids d'ouverture et d'extension de gap en même temps, des résultats similaires sont obtenus.

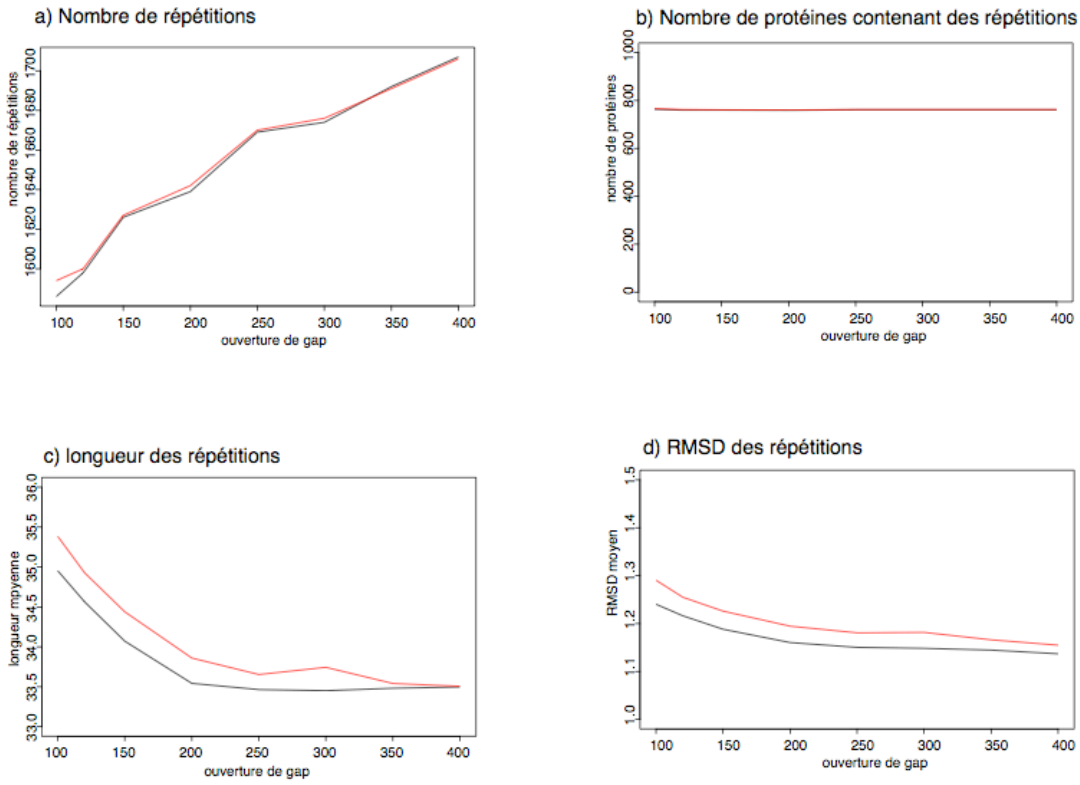


Figure 19 : Poids d'ouverture de gap pour les structures.
(noir : RMSD<5 Å et rouge : RMSD<10Å).

Tableau 1 : Récapitulatif des scores utilisés par défaut à chaque niveau ($S_{i,j}$).
pi, pj : fréquences des acides nucléiques pour les séquences d'ADN et fréquences normalisées des angles αi et αj dans la PDB pour les structures. |Δangle| est la différence angulaire entre les angles i et j (de 0 à 180°). Le paramètre β permet de pondérer le score par les fréquences des angles p_i et p_j.

		Pénalité de gap	
	Substitution/Identité	Ouverture	Extension
Séquence d'ADN	$S_{i,j} = 0{,}5 \times \sigma_{(i,j)} \times \log_4(p_i p_j)$ $\sigma_{(i,j)} = 1$ if $i \neq j$; $\sigma_{(i,j)} = -1$ if $i = j$	-4	-1
Séquence d'acides aminés	Matrice BLOSUM ou PAM	-8	-3
Structure 3D	$S_{i,j} = 30*[(1-p_i)(1-p_j)(1-\beta)+ \beta] - \lvert \Delta angle \rvert$ $\beta = 0{,}4$	-200	-50

III.D La significativité statistique des répétitions trouvées

III.D.1 Les méthodes de Waterman et Vingron pour les séquences

Deux méthodes statistiques ont été implémentées pour estimer la significativité des répétitions trouvées dans les séquences (Waterman and Vingron, 1994).

III.D.1.a Première méthode

La première méthode proposée par Waterman et Vingron consiste à générer un nombre important de séquences aléatoires de même taille et de même composition que la séquence initiale, puis à chercher le meilleur score d'alignement avec Swelfe. Ces scores suivent une distribution de valeurs extrêmes (EVD) de la forme :

$$P(score > t) = 1 - e^{-\gamma \frac{n(n-1)}{2} p^t}$$

Équation 16

t est le score seuil, n est la longueur de la séquence, p et γ sont des paramètres à calculer. Dans la formule le paramètre $\frac{n(n-1)}{2}$ est utilisé (au lieu de nm dans la formule initiale, n et m étant la taille des séquences) car nous cherchons des répétitions internes, donc nous n'utilisons que la moitié de la matrice d'alignement, en excluant sa diagonale, ce qui réduit l'espace des possibilités.

Pour simplifier les calculs et se ramener au calcul d'une p-value, on calcule P(score ≤t)=1-P(score>t) :

$$P = e^{-\gamma \frac{n(n-1)}{2} p^t}$$

Équation 17

Pour chaque score t obtenu avec les séquences aléatoires, la probabilité correspondante est estimée : pour le meilleur score, la probabilité d'obtenir un score strictement supérieur à lui même est de 0 / N, N étant le nombre de séquences aléatoires, la probabilité pour le 2nd score est de 1/N etc.

Il est ensuite possible de tracer la droite $\log(-\log(P))$ en fonction de t.

Une fois la droite tracée (Figure 20), une régression linéaire pondérée est effectuée pour déterminer les paramètres p et γ à partir de l'équation 18 qui est elle même tirée de l'équation 17 :

$$\log(-\log(P)) = \log(\gamma \frac{n(n-1)}{2}) + t\log(p)$$

Équation 18

$\log(p)$ est la pente et $\log(\gamma \frac{n(n-1)}{2})$ est l'ordonnée à l'origine de la droite. Il est nécessaire d'effectuer une régression linéaire pondérée par le nombre de scores, pour chaque valeur de score, parce que la matrice BLOSUM62 contient des valeurs entières, donc les scores de séquences d'acides aminés sont des valeurs entières, et il arrive fréquemment que plusieurs scores aient la même valeur.

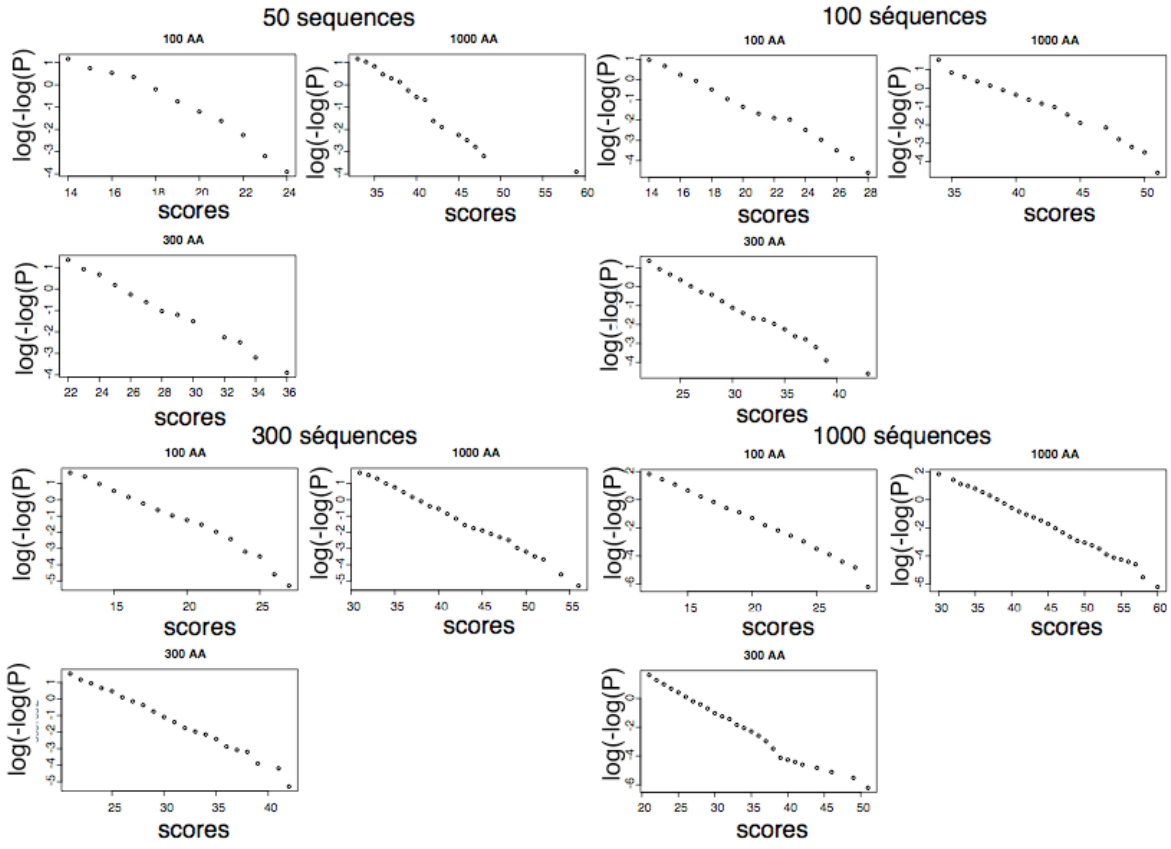


Figure 20 : Valeurs de log(-log(P)) en fonction du score t pour des séquences de 100, 300 et 1000 acides aminés.
Nous avons testé pour 50, 100, 300 et 1000 séquences. Plus les séquences sont longues et plus le nombre de séquences est important, plus la courbe ressemble à une droite, donc plus la régression linéaire pondérée sera précise.

Ensuite, avec la formule précédente et les paramètres calculés précédemment (équation 17), il est possible de déterminer la p-value associée à chaque score. Si cette

p-value est inférieure à un seuil (par défaut 0,01), la répétition obtenue est statistiquement significative.

Nous avons effectué plusieurs tests pour déterminer le nombre de séquences aléatoires à générer pour déterminer au mieux les paramètres manquants (Figure 21).

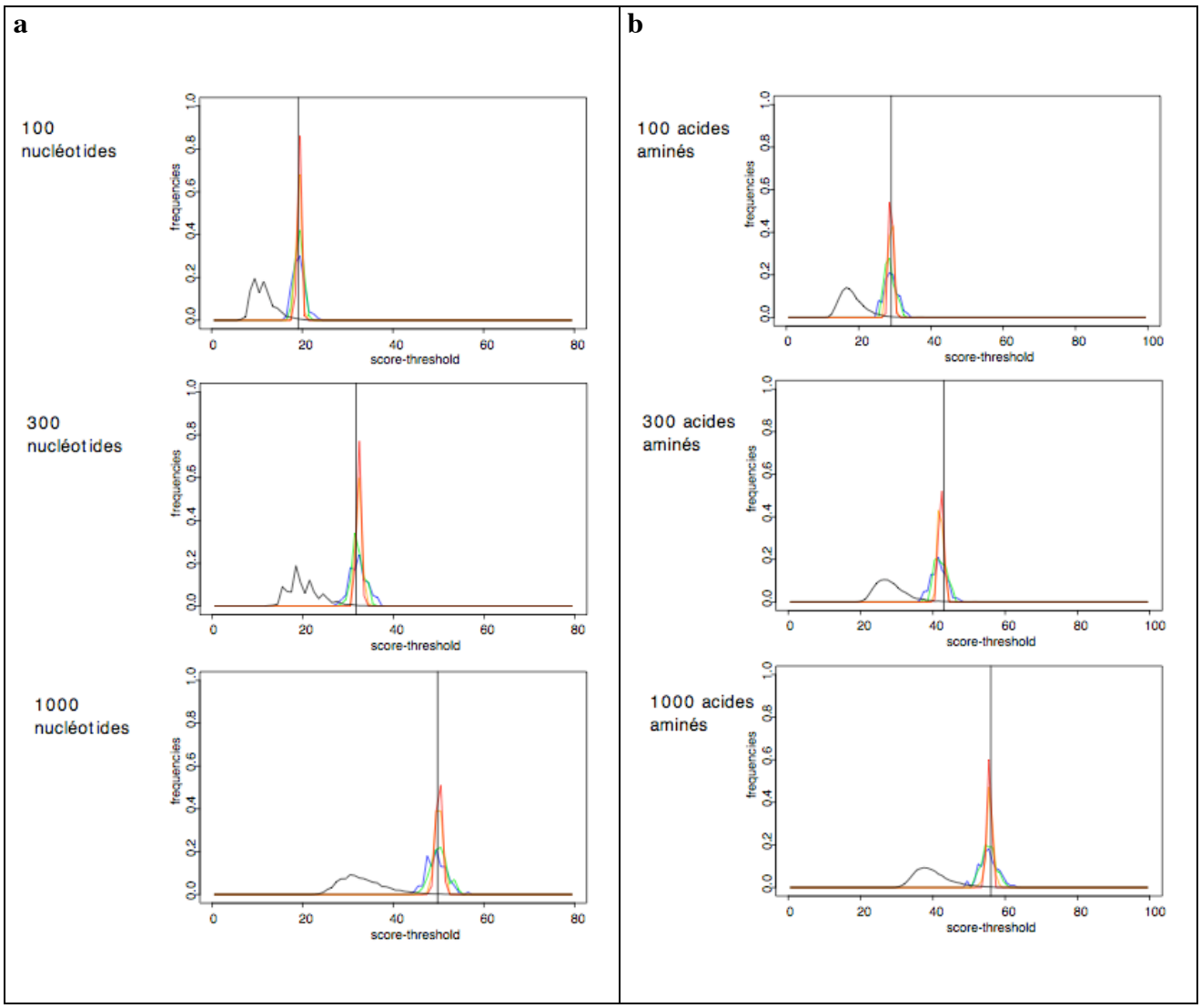


Figure 21 : Nombre de séquences aléatoires pour déterminer au mieux les paramètres de la droite.
Tests effectués pour des séquences de 100, 300, 1000 acides aminés ayant la composition moyenne de SWISS-PROT, ainsi que des séquences de 100, 300, 1000 nucléotides ayant une composition de 25% de chaque base. La méthode statistique a été appliquée sur chaque séquence en prenant un nombre variable de séquences aléatoires. Pour chaque cas, le score seuil à 0,01 est calculé et ce calcul est répété 100 fois. Bleu = 50 séquences, vert = 100 séquences, orange = 500 séquences, rouge = 1000 séquences. Le score seuil obtenu avec une significativité de 0.01 a été calculé cent fois. Courbe noire : distribution des scores sur un échantillon de 10 000 séquences, droite noire verticale : score seuil à 1% sur ce jeu de données. a : nucléotides, b : acides aminés.

Les résultats montrent que 100 séquences aléatoires suffisent à obtenir un seuil correct (courbe verte), mais dans certains cas les résultats sont plus précis avec 1000 séquences aléatoires (courbe rouge), donc les paramètres par défaut utilisent 100 séquences mais pour les analyses, nous utiliserons 1000 séquences aléatoires.

III.D.1.b Méthode de « Declumping »

La seconde méthode est la méthode dite de « declumping », proposée également par Waterman et Vingron (Waterman and Vingron, 1994). L'intérêt de cette méthode est qu'elle utilise une approximation qui nécessite de créer moins de séquences aléatoires et donc prend moins de temps.

Le principe de cette méthode est le suivant : le paramètre $\gamma \frac{n(n-1)}{2} p^t$ peut être estimé par E, le nombre moyen de « clump » ayant un score supérieur à t.

$$E = \gamma \frac{n(n-1)}{2} p^t$$

Équation 19

Pour calculer ce paramètre E, il faut générer des séquences aléatoires comme dans la méthode précédente, sauf qu'au lieu de conserver uniquement le meilleur alignement pour chaque séquence aléatoire, plusieurs meilleurs alignements sont calculés, indépendants les uns des autres (Méthode de Waterman et Eggert (Waterman and Eggert, 1987)). Chacun de ces alignements indépendants est appelé « clump ».

Il est ensuite possible de calculer ce paramètre E pour chaque score t : il s'agit du nombre moyen de « clump » ayant un score supérieur à t divisé par le nombre de séquences aléatoires générées.

Il est ensuite possible de tracer la droite (Figure 22) $\log(E) = \log(\gamma \frac{n(n-1)}{2}) + t \log(p)$, afin de déterminer les paramètres p et γ comme précédemment.

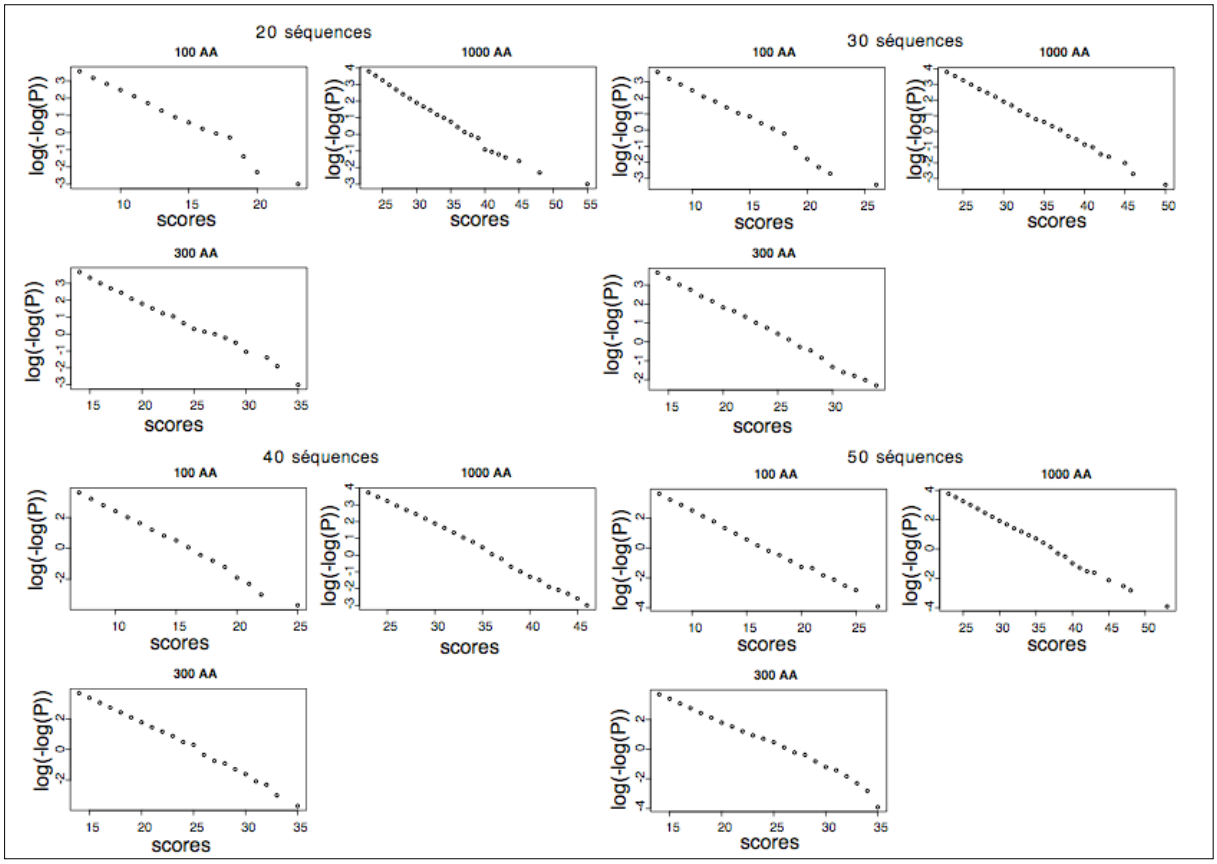


Figure 22 : Valeurs de log(-log(P)) en fonction du score t pour des séquences de 100, 300 et 1000 acides aminés, méthode de declumping.
Nous avons testé pour 20, 30, 40 et 50 séquences.

Ensuite, avec la formule : $P = e^{-\gamma \frac{n(n-1)}{2} p^t}$ il est possible de calculer la p-value associée à chaque score.

Dans ce cas-ci également, nous avons cherché à savoir combien de séquences aléatoires étaient nécessaires pour déterminer au mieux les paramètres, et combien de « clump » devaient être calculés pour chaque séquence. Il faut calculer pour chaque séquence aléatoire plusieurs « clump » jusqu'à atteindre un score seuil. Il n'y aura donc pas le même nombre de « clump » pour chaque séquence aléatoire.

Le score seuil est le score minimal obtenu en cherchant n répétitions dans la matrice d'alignement de la première séquence aléatoire. Pour toutes les séquences aléatoires suivantes, les « clump » seront calculés jusqu'à atteindre ce score seuil. Nous avons calculé le nombre de séquences aléatoires à générer dans les mêmes conditions que pour la première méthode statistique.

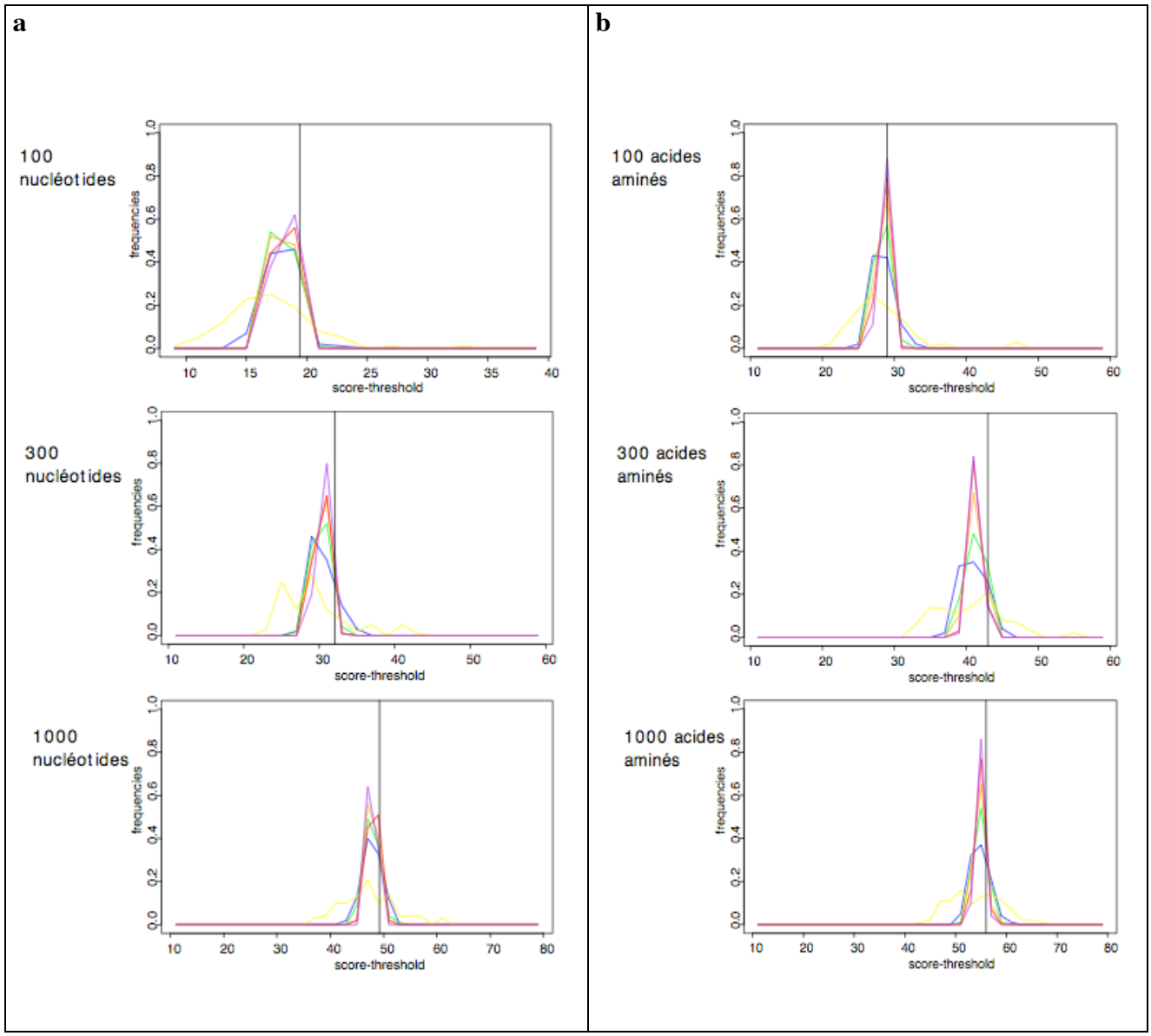


Figure 23 : Nombre de séquences aléatoires pour déterminer au mieux les paramètres de la droite.
Tests effectués pour des séquences de 100, 300, 1000 acides aminés ayant la composition de SWISS-PROT, ainsi que des séquences de 100, 300, 1000 nucléotides ayant une composition de 25% de chaque base. La méthode statistique a été appliquée sur chaque séquence en prenant un nombre variable de séquences aléatoires. Pour chaque cas, le score seuil à 1% est calculé et ce calcul est répété 100 fois. Jaune = 1 séquence, bleu = 10 séquences, vert = 20 séquences, orange = 30 séquences, rouge = 50 séquences, violet = 80 séquences). Le score seuil obtenu avec une significativité de 0.01 a été calculé cent fois. Droite noire verticale : score seuil 1% sur un échantillon de 10 000 séquences. a : nucléotides, b : acides aminés.

La Figure 23 montre que 20 séquences aléatoires permettent d'avoir une estimation correcte du score seuil, et cette valeur est un bon compromis entre le temps de calcul et la précision du seuil. Des estimations ont également été faites pour savoir le nombre de scores à calculer pour chaque séquence aléatoire, et 50 semble être une bonne valeur (c'est le nombre de scores pour la 1ère séquence aléatoire, les autres

nombres de scores dépendent du score le plus faible obtenu avec la 1ère séquence aléatoire). Cependant cette méthode paraît sous estimer les probabilités (biais à gauche sur la Figure 23), nous avons donc conservé la première méthode pour les analyses présentées par la suite.

III.D.2 Le calcul du RRMSD pour les structures

III.D.2.a Le RMSD

Le RMSD (Root Mean Square Deviation) est une mesure de comparaison de structures très utilisée pour comparer deux structures ou deux fragments de structures. Pour le calculer, il faut superposer au mieux les deux structures puis calculer la racine carrée de la moyenne des carrés des distances entre les carbones α correspondants entre les deux structures.

$$RMSD = \sqrt{\frac{\sum_{i=1}^{N} (x_i - x_i')^2 + (y_i - y_i')^2 + (z_i - z_i')^2}{N}}$$

Équation 20

N est le nombre de carbones α à superposer pour chaque structure, et les carbones α_i ont pour coordonnées (x_i,y_i,z_i) pour la structure A et (x_i',y_i',z_i') pour la structure B.

Les structures sont superposées de façon à obtenir le RMSD minimal. Plusieurs méthodes sont possibles pour calculer cette superposition. Nous avons utilisé la méthode basée sur les quaternions de Zucker et Somorjai (Zuker and Somorjai, 1989).

Le RMSD est calculé pour chaque répétition trouvée afin de vérifier que les copies de la répétition sont bien superposables (le codage en angles α n'est pas complètement équivalent aux coordonnées des Cα). Il a été utilisé comme filtre après alignement avant que nous ne programmions le calcul du RRMSD dans Swelfe (par défaut RMSD < 4,5Å).

III.D.2.b Le RMSD Relatif (RRMSD)

Le RMSD Relatif, ou RRMSD a été proposé par Betancourt *et al*. (Betancourt and Skolnick, 2001).

Le RRMSD est indépendant de la longueur des fragments protéiques à aligner, et permet donc de comparer des alignements de longueurs différentes. En effet, le RMSD

est très dépendant de la longueur des répétitions. Des répétitions courtes ont donc souvent des RMSD meilleurs que des répétitions plus longues. Nous avons donc finalement utilisé le RRMSD comme filtre après calcul des alignements par la matrice SIM pour vérifier que les fragments obtenus se superposent bien, indépendamment de leur longueur. Ce filtre permet à certaines longues répétitions dont le RMSD était supérieur au seuil (4,5Å) d'être conservées.

Le calcul du RRMSD est un calcul de RMSD divisé par un RMSD obtenu aléatoirement entre deux fragments structuraux α et β de même longueur N choisis aléatoirement.

$$RRMSD_{\alpha\beta} = \frac{RMSD_{\alpha\beta}}{< RMSD_{\alpha\beta} >}$$

Équation 21

Le RMSD moyen entre deux polypeptides de rayons de giration $R_{g\alpha}$ et $R_{g\beta}$ pris au hasard est :

$$< RMSD_{\alpha\beta} >= R^2_{g\alpha} + R^2_{g\beta} - 2C(N)R_{g\alpha}R_{g\beta}$$

Équation 22

La fonction C(N), qui ne dépend que de la taille de la protéine, a été calculée par Betancourt et Skolnick à partir d'une banque de fragments de près de 1300 structures aléatoires non homologues. Ils ont trouvé la relation suivante :

$$C(N) \approx 0.42 - 0.05(N-1)e^{-(N-1)/4.7} + 0.63e^{-(N-1)/37}$$

Équation 23

Le rayon de giration mesure la compacité d'une protéine et est calculé par :

$$R^2_{g\alpha} = \frac{1}{N}\sum_{i=0}^{N}\left[(x_i - x_c)^2 + (y_i - y_c)^2 + (z_i - z_c)^2\right]$$

Équation 24

N est la longueur de la protéine et le centre de masse de la protéine a pour coordonnées (x_c, y_c, z_c). Voici le détail de leurs calculs pour la fonction C(N). Le RMSD entre α et β se calcule par :

$$RMSD^2_{\alpha\beta} = \frac{1}{N}\sum_{i=1}^{N}\left(r_{\alpha,i} - Qr_{\beta,i}\right)^2 = R^2_{g\alpha} + R^2_{g\beta} - 2\left(\frac{\sum_{i=1}^{N} r_{\alpha,i} \times Qr_{\beta,i}}{\sqrt{\sum_{i=1}^{N} r^2_{\alpha,i} \sum_{i=1}^{N} r^2_{\beta,i}}}\right) R_{g\alpha} R_{g\beta}$$

Équation 25

Q est la matrice de rotation qui aligne au mieux les vecteurs. $r_{\alpha,i}$ et $r_{\beta,i}$ sont les coordonnées des résidus des fragments α et β à la position i.

C(N) est défini par :

$$C(N) \equiv \left\langle \frac{\sum_{i=1}^{N} r_{\alpha,i} \times Qr_{\beta,i}}{\sqrt{\sum_{i=1}^{N} r^2_{\alpha,i} \sum_{i=1}^{N} r^2_{\beta,i}}} \right\rangle_{\alpha\beta}$$

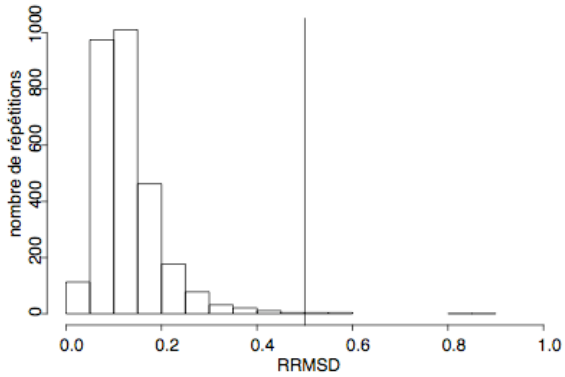
Équation 26

A partir de la banque de fragments de près de 1300 structures aléatoires indépendantes (moins de 30% d'identité), les auteurs ont calculé un « coefficient de corrélation entre les structures » qui est une mesure, liée au RMSD, de similarité entre deux structures après superposition optimale. En traçant le nombre de résidus en fonction de ce coefficient, ils ont obtenu une courbe dont l'équation est ci-dessus (équation 23).

Le RRMSD permet également de calculer une p-value sur l'alignement des répétitions structurales. Dans leur article, Betancourt et Skolnick (Betancourt and Skolnick, 2001) indiquent que deux structures ayant un RRMSD de 0 sont identiques et deux structures ayant un RRMSD de 1 sont aussi différentes que deux structures prises au hasard. Pour les protéines de plus de 100 acides aminés de long, la distribution de probabilité devient gaussienne avec un écart type de 0,11. La probabilité que deux structures aléatoires aient un RRMSD<0,5 est de 10^{-6}. Pour des chaînes plus courtes, la distribution s'éloigne d'une gaussienne et pour deux chaînes de 20 résidus, un RRMSD <0,5 a une probabilité de 10^{-3} d'apparaître par hasard. Comme les répétitions structurales intragéniques trouvées avec Swelfe ont généralement une longueur d'au moins 20 acides aminés, nous avons mis comme filtre après alignement un seuil de RRMSD de 0,5. Cela correspond à une p-value de 10^{-3}.

La plupart des répétitions trouvées satisfont ce filtre, ce qui montre que l'utilisation des angles α est une bonne approximation du squelette carboné de la protéine (Figure 24).

a)



b)

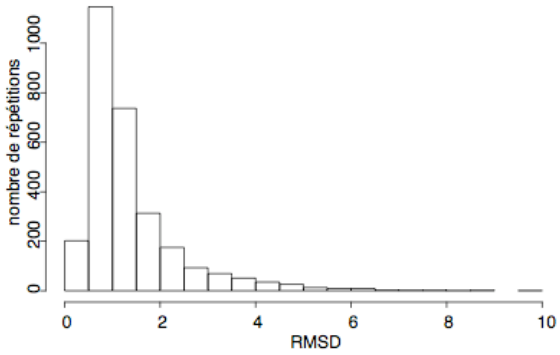


Figure 24 : RRMSD et RMSD des répétitions structurales.
a) RRMSD des répétitions structurales pour un score > 250. La plupart des répétitions sont en dessous du seuil de 0,5. b) : RMSD des répétitions structurales après application du seuil RRMSD<0,5. La plupart des répétitions ont un RMSD < 4,5Å, sauf certaines longues répétitions.

III.E Les banques de données de séquences et de structures

Nous avons besoin de données existant à la fois en séquences d'ADN, en séquences d'acides aminés et en structure 3D.

III.E.1 Les données utilisées

Les liens vers les séquences d'acides aminés présents dans les fichiers PDB sont en nombre insuffisants (Tableau 2). Près de 40% des chaînes de la PDB n'ont pas de lien vers une banque de séquence. De plus, il n'y a pas de lien direct entre les structures de la PDB et les séquences nucléiques correspondantes.

Tableau 2 : Liens vers les banques présents dans les fichiers PDB.
(données d'août 2008).

Nom de banque	Nombre de références dans la PDB
Uniprot (UNP)	75181
Genbank (GB)	2418
Protein Identification Ressource (PIR)	197
EMBL	106
TrEMBL	6
NDB	2
Swiss-prot (SWS)	1
Total	**77 911**
Nombre de chaînes	**128 107**
PDB (fichiers PDB se référençant eux-mêmes)	3042

Pour chercher les séquences d'ADN correspondant aux structures 3D, nous avons écrit un programme spécifique. Nous avons utilisé les structures de la PDB, en sélectionnant les séquences d'après le champ SEQRES des fichiers PDB (fichier de séquences correspondant aux structures PDB) pour conserver les séquences protéiques de plus de 50 acides aminés et avec moins de 10% de X (résidus indéterminés). Cependant, les séquences contenues dans le champ SEQRES ne correspondent pas toujours aux acides aminés du fichier PDB, il peut y avoir des résidus qui diffèrent, et la taille peut être différente. De ce fait, nous utilisons les séquences d'acides aminés correspondant aux structures sélectionnées précédemment directement dans le fichier PDB dans les champs ATOM et HETATM. Les acides aminés non standards sont remplacés par les acides aminés les plus proches lorsque c'est possible (cf Tableau 3).

D'autre part, les séquences CDS de TrEMBL non redondantes sont récupérées puis traduites en séquences d'acides aminés.

Tableau 3 : Correspondance entre les acides aminés spéciaux et les acides aminés les plus proches correspondants.

Acides aminés des fichiers PDB	Noms entiers	Acides aminés les plus proches
ASX	ASP/ASN	D
GLX	GLU/GLN	E
CEA	s-hydroxy-cysteine	C
HYP	hydroxyproline	P
PCA	5-pyrrolidone-2-carboxylic_acid	E
UNK	unknown	X
ACD	acidic unknown	X
HSE	homoserine	S
HYL	hydroxylysine	K
ALB	beta-alanine	A
ALI	aliphatic unknown	X
ARO	aromatic unknown	X
BAS	basic unknown	X
MSE	selenomethionine	M
CSE	selenocysteine	C

III.E.2 Quickhit : le problème de la recherche des gènes associés aux structures de la PDB

Ensuite le programme Quickhit est utilisé. Il est basé sur un automate et a pour objectif de trouver les séquences correspondant aux structures. Le principe est le suivant : il cherche des graines (répétitions exactes) d'une taille de 10 acides aminés présentes dans la séquence requête (PDB) et la séquence cible (CDS). Les graines doivent être éloignées au moins de 10 positions. Si une structure a au moins deux graines en commun avec une séquence ADN, un alignement exact est réalisé. Le programme sort les couples qui ont plus de 70% de similarité.

Seuls les couples qui ont plus de 90% d'identité sont conservés pour faire la banque. Ensuite, les structures sont converties en angles α. Les séquences et les structures sont éventuellement coupées pour se ramener à la taille de la séquence/structure la plus courte. Il arrive fréquemment que les séquences soient plus longues que les structures. S'il y a des gaps dans une séquence/structure, ils sont

remplacés par des Z (caractère non alignable). Si l'alignement a plus de 5% de gaps, il n'est pas conservé.

Cette banque contient actuellement 93 136 séquences et structures.

Nous avions au départ utilisé le programme CD-HIT (Li and Godzik, 2006) pour trouver les séquences protéiques des CDS de TrEMBL qui sont les plus similaires possibles des séquences protéiques extraites des fichiers PDB, néanmoins, à la fin du calcul, il manquait dans le meilleur des cas encore plus de 18% des structures du jeu de données Cluster50 (en enlevant les structures trop petites ou contenant trop de caractères indéterminés) et la fabrication de cette banque était longue (plus d'une semaine pour effectuer tous les calculs). Nous avons finalement écrit le programme Quickhit, qui permet de couvrir 85% des protéines de la PDB de plus de 50 acides aminés et contenant moins de 10% de résidus inconnus. Ce programme est beaucoup plus rapide et fabrique la banque en un à deux jours. Pour certaines structures de la PDB, il est difficile, voire impossible, de trouver les séquences nucléiques qui leur correspondent . Les causes en sont que certaines structures sont synthétisées de façon artificielle, d'autres sont très petites (moins de 50 résidus), d'autres encore ont un nombre important de résidus indéterminés.

III.F Affinement de la méthode

Plusieurs options ont été ajoutées au programme Swelfe pour le rendre plus performant.

III.F.1 Allongement des répétitions structurales

Les extrémités des alignements structuraux ne sont pas toujours déterminées précisément et l'observation visuelle des répétitions montre parfois que l'alignement pourrait être prolongé d'un côté ou de l'autre. Pour remédier à ce problème, nous avons ajouté une option inspirée de ProSup (Lackner et al., 2000). Cette options permet d'allonger l'alignement si les carbones α à ses extrémités sont suffisamment proches. Pour cela, nous calculons la distance euclidienne dans l'espace entre deux carbones α de part et d'autre de l'alignement avec la formule :

$$\text{distance}(A,B) = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2 + (z_a - z_b)^2}$$

Équation 27

La valeur seuil de l'élongation est nommée e (en général 1Å), I et J sont les positions de début de l'alignement et I' et J' sont les positions de fin de l'alignement.

Pour $I' < i \leq m$ et $J' < j \leq n$, tant que distance$(A_i, B_j) \leq e$, l'alignement est allongé d'un carbone α aux positions A_i et B_j. Le calcul est également effectué pour $I > i \geq 0$ et $J > j \geq 0$.

Bien sûr, il n'est pas possible de prendre en compte d'éventuels gaps avec cette méthode.

III.F.2 Recouvrement des alignements

Lorsque l'algorithme SIM est appliqué à deux séquences différentes A et B, il est tout à fait possible, que les positions de l'alignement de $A_{I,I'}$ et de $B_{J,J'}$ se recouvrent, c'est à dire que I≤J ≤I' ou J≤ I≤ J', par exemple si les séquences sont très similaires sur des positions proches.

Par contre, si ce recouvrement est observé lorsque A et B sont la même séquence, (recherche de répétitions internes, qui est le cas qui nous intéresse), il a une signification biologique différente : il s'agit d'une séquence qui contient plusieurs copies répétées en tandem, au moins trois. Dans ces cas, les copies 1 et 2 sont alignées avec les copies 2 et 3, il y a donc un recouvrement au niveau de la 2ème copie.

Nous avons donc ajouté une option permettant d'accepter ou refuser un recouvrement partiel des copies des répétitions trouvées (entre 0 et 1). Ce paramètre est mis à 0 pour ne trouver que des répétitions non chevauchantes, et permet d'obtenir dans plusieurs alignement successifs la copie 1 alignée avec la copie 2, la copie 2 avec la copie 3 et la copie 3 avec la copie 1.

III.F.3 Suppression des répétitions successives chevauchantes

L'algorithme de Huang et Miller (Huang and Miller, 1991) permet de calculer plusieurs meilleurs alignements successifs. Cependant, pour certaines séquences très répétées, il arrive que les alignements successifs se chevauchent, c'est-à-dire que le 2nd meilleur alignement commence quelques positions après le 1er et lui ressemble beaucoup. Pour résoudre ce problème, nous avons écrit un script python à utiliser

comme filtre après le programme qui permette de ne garder qu'un alignement (le meilleur) si plusieurs alignements se chevauchent plus qu'une valeur seuil définie par l'utilisateur (par défaut 0,5 fois la longueur de l'alignement obtenu).

III.F.4 Calcul de l'angle de rotation pour superposer les deux copies d'une répétition structurale

Nous avons inclus dans Swelfe, pour la suite des analyses, le calcul de l'angle de rotation nécessaire pour superposer deux copies d'une répétition structurale. Cela est nécessaire pour déterminer les répétitions dont les copies sont symétriques à 180° (symétrie C2). Ce calcul est basé sur le calcul de superposition des deux copies de la répétition structurale de Zucker et Somorjai (Zuker and Somorjai, 1989). Ensuite, un changement de repère permet de se placer dans un nouveau repère dont l'axe x est l'axe de la rotation principale qui permet de superposer les deux copies de la répétition. Il est ainsi plus facile de calculer l'angle de rotation dans le nouveau repère.

Les étapes sont donc les suivantes :

1. La superposition optimale des deux copies de la répétition est calculée avec l'algorithme de Zucker *et al*. (Zuker and Somorjai, 1989), afin d'obtenir une matrice de rotation.
2. Les valeurs propres et vecteurs propres de la matrice de rotation sont calculés. La partie réelle des valeurs propres différentes de 1 correspond au cosinus de l'angle de rotation et la partie imaginaire correspond au sinus de cet angle. Il est donc possible d'en déduire l'angle qui permet de superposer les deux copies de la répétition.

III.G Comparaison des résultats de Swelfe et DALI et temps d'exécution

Pour confirmer la validité de notre approche au niveau des répétitions structurales, nous avons comparé les résultats de notre programme avec ceux de DALI. Swelfe a été modifié pour prendre en entrée deux structures différentes car DALI ne peut pas chercher des répétitions structurales internes. Nous allons donc comparer une structure contre une banque de données. Les jeux de données tests sont issues de l'article de Novotny *et al*. (Novotny et al., 2004) où les auteurs comparent les résultats de plusieurs serveurs de comparaison de structures.

Il serait difficile de calculer la sensibilité et la spécificité des répétitions trouvées en structures car il faudrait pour cela pouvoir définir quels sont les faux positifs et les faux négatifs dans les répétitions structurales, et nous ne disposons pas de critères statistiques adaptés pour cela.

La banque de données utilisée est Cluster90 (issue de la PDB, structures présentant moins de 90% d'identité entre les séquences). Elle contient 17 293 structures.

III.G.1 Les cas difficiles

Ce jeu de données contient 11 paires de structures similaires, mais non triviales à trouver par les algorithmes classiques de recherche de similarité de structure, qui sont issues pour 10 d'entres elles de l'article de Fisher *et al*. (Fischer et al., 1996). En donnant en entrée une protéine de chaque paire au programme, celui-ci doit retrouver l'autre dans une banque de structures (Cluster90) au meilleur rang possible. Les résultats ici (Tableau 4) montrent que DALI donne des résultats un peu meilleurs que Swelfe.

Tableau 4 : Résultats pour les cas difficiles.
Ce tableau contient la position à laquelle est trouvée la cible parmi tous les résultats dans la banque de structures. NS : non significatif. (en bleu = dans les 100 meilleurs, en vert : autre rangs significatifs).

		Swelfe	DALI
Requête	**Cible**	**Rang**	**Rang**
1bgeB	2gmfA	**NS**	**43**
3hlaB	2rheA	**NS**	**615**
2azaA	1pazA	**NS**	**331**
1cewI	1molA	**30**	**35**
1fxiA	1ubqA	**215**	**NS**
1cidA	2rheA	**199**	**NS**
1crlA	1edeA	**351**	**271**
1tenA	3hhrB	**92**	**NS**
1tieA	4fgfA	**NS**	**177**
2simA	1nsbA	**683**	**74**
1g61A	1jdwA	**508**	**21**

III.G.2 La famille des cyclophilines

La famille des cyclophilines contient plusieurs protéines de cette famille. Il faut ici donner en entrée une protéine de cette famille au programme et il doit trouver les protéines de la même famille au meilleur rang possible. Les résultats (Tableau 5) montrent qu'ici Swelfe donne de meilleurs résultats que DALI.

Tableau 5 : Résultats pour la famille des cyclophilines.
Le rang de chaque cible parmi les résultats est indiqué (dans les 10 meilleurs en bleu, dans les 20 meilleurs en vert, identité en rouge).

A-Swelfe	Cible							
Requête	1awq	1cyn	1qoi	1lop	1qng	2rmc	1dyw	1ihg
1awq	1	11	23	33	10	9	6	21
1cyn	15	1	23	33	17	2	13	19
1qoi	18	15	1	34	5	21	13	3
1lop	14	12	29	1	27	11	31	19
1qng	15	18	20	34	1	17	2	8
2rmc	14	3	23	33	17	1	12	19
1dyw	13	16	23	34	2	15	1	12
1ihg	24	16	11	44	6	18	5	1

B-DALI	Cible							
Requête	1awq	1cyn	1qoi	1lop	1qng	2rmc	1dyw	1ihg
1awq	1	114	111	140	96	126	87	121
1cyn	117	1	103	139	104	4	101	105
1qoi	106	96	1	137	5	120	98	87
1lop	126	19	31	1	37	30	133	119
1qng	85	119	109	140	1	124	7	107
2rmc	114	5	107	139	110	1	103	115
1dyw	77	117	119	140	7	123	1	107
1ihg	112	80	64	139	18	119	16	1

III.G.3 Les « few-SSE »

Il s'agit de protéines qui ont peu de structures secondaires. Comme notre programme n'est pas basé sur les structures secondaires, il ne devrait pas être pénalisé. Il s'agit encore une fois de donner en entrée une protéine au programme et que celui-ci trouve toutes les autres protéines au meilleur rang possible. Ici encore (Tableau 6), il semble que Swelfe se comporte mieux que DALI.

Tableau 6 : Résultats pour les « few-SSE ».
Le rang de chaque cible parmi les résultats est indiqué (en bleu : 10 meilleurs rangs, en vert : 20 meilleurs et en rouge : identité).

A-Swelfe	Cible			
Requête	1b2i	1cea	1pml	5hpg
1b2i	1	8	14	6
1cea	10	1	8	3
1pml	NS	9	1	10
5hpg	12	2	11	1

B-DALI	Cible			
Requête	1b2i	1cea	1pml	5hpg
1b2i	1	9	38	20
1cea	30	1	33	NS
1pml	39	16	1	9
5hpg	39	4	38	1

III.G.4 Conclusion

Ces résultats montrent que Swelfe donne des résultats corrects pour tous les jeux de données testés. Ces résultats sont comparables à ceux de DALI et ne sont pas plus mauvais dans la plupart des cas. Cela montre donc que la méthode que nous avons utilisée pour comparer les structures est valide.

III.G.5 Temps d'exécution

Swelfe est relativement rapide. Nous avons effectué des tests en utilisant un MacPro Xeon et en analysant les 9537 protéines de l'ensemble Cluster50 de la PDB pour lesquelles nous avons les séquences d'ADN et d'acides aminés correspondantes. Le programme a mis moins d'une minute pour trouver les répétitions structurales ou d'acides aminés, et 5 minutes pour les répétitions nucléiques. En ajoutant les évaluations statistiques qui ralentissent le programme, à cause du nombre de séquences à aléatoires à générer et pour lesquelles on effectue l'alignement, le programme met 30 minutes pour les séquences d'acides aminés et 20h pour les séquences nucléiques. Ces calculs ont été faits avec la méthode de « declumping » de Waterman et Vingron et avec 20 séquences aléatoires. Swelfe utilise environ 16 Mo Ram pour la banque d'ADN, qui est la plus grande.

III.H La mise à disposition du programme

III.H.1 Le site Internet

J'ai développé un site web pour mettre Swelfe à la disposition de la communauté, il est disponible à l'adresse : http://bioserv.rpbs.jussieu.fr/cgi-bin/swelfe.

Il est hébergé par RPBS. La Ressource Parisienne en Bioinformatique Structurale est, comme son nom l'indique, un serveur qui héberge des outils de bioinformatique structurale. Il a pour but de mettre à la disposition de la communauté des logiciels spécifiques et innovants. Il contient maintenant plus de 20 programmes d'analyse de séquence et de structure, de drug design, etc.

Swelfe peut être téléchargé et/ou utilisé en ligne, il y a deux pages d'aide qui expliquent tous les paramètres qui peuvent être utilisés pour chaque cas. Voici la page d'accueil de Swelfe (Figure 25).

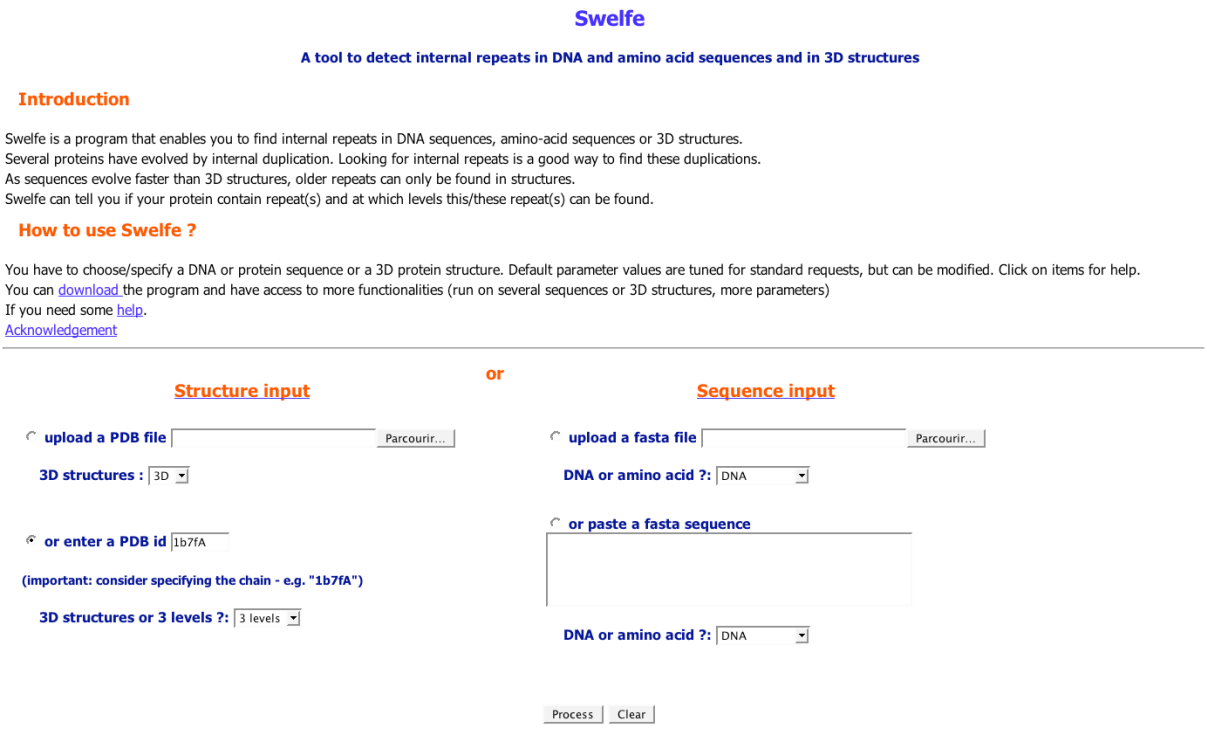


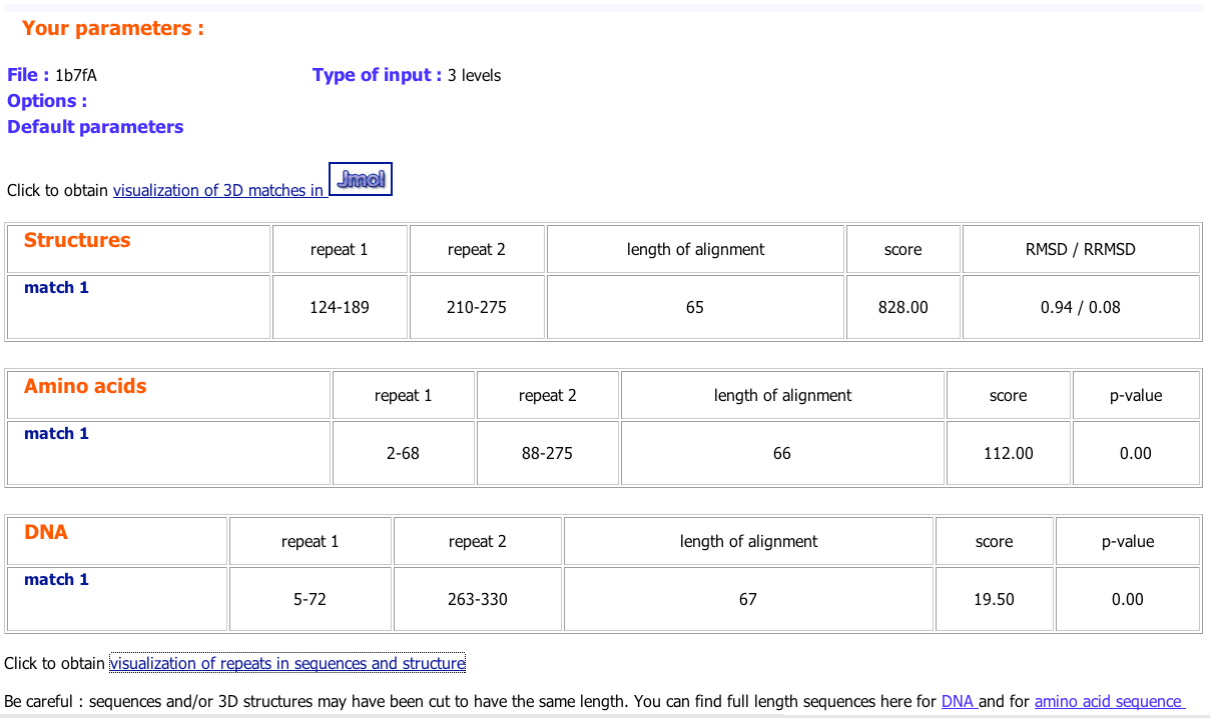
Figure 25 : Page d'accueil du site web.

La requête peut être soit une séquence (ADN ou acides aminés), soit une structure 3D. Si l'utilisateur choisit de donner un nom PDB, il peut demander au programme de chercher directement les séquences d'ADN et d'acides aminés correspondantes qui sont

dans la banque qui a été pré-calculée (cf. partie III.E). Le programme calcule ensuite les répétitions présentes aux 3 niveaux.

III.H.2 L'affichage des résultats aux trois niveaux

La page de résultats est présentée sur la Figure 26. Si l'option « 3 levels » a été sélectionnée, il est possible de voir les positions des répétitions aux trois niveaux en même temps.



Your parameters :

File : 1b7fA **Type of input :** 3 levels
Options :
Default parameters

Click to obtain visualization of 3D matches in Jmol

Structures	repeat 1	repeat 2	length of alignment	score	RMSD / RRMSD
match 1	124-189	210-275	65	828.00	0.94 / 0.08

Amino acids	repeat 1	repeat 2	length of alignment	score	p-value
match 1	2-68	88-275	66	112.00	0.00

DNA	repeat 1	repeat 2	length of alignment	score	p-value
match 1	5-72	263-330	67	19.50	0.00

Click to obtain visualization of repeats in sequences and structure

Be careful : sequences and/or 3D structures may have been cut to have the same length. You can find full length sequences here for DNA and for amino acid sequence

Figure 26 : Page de résultats du site web.

- *Trois niveaux en parallèle*

Un avantage du site web par rapport au téléchargement du programme est que les alignements obtenus sont comparables facilement. Si l'option « 3 levels » est sélectionnée, le programme affiche en parallèle les répétitions trouvées aux trois niveaux (Figure 27).

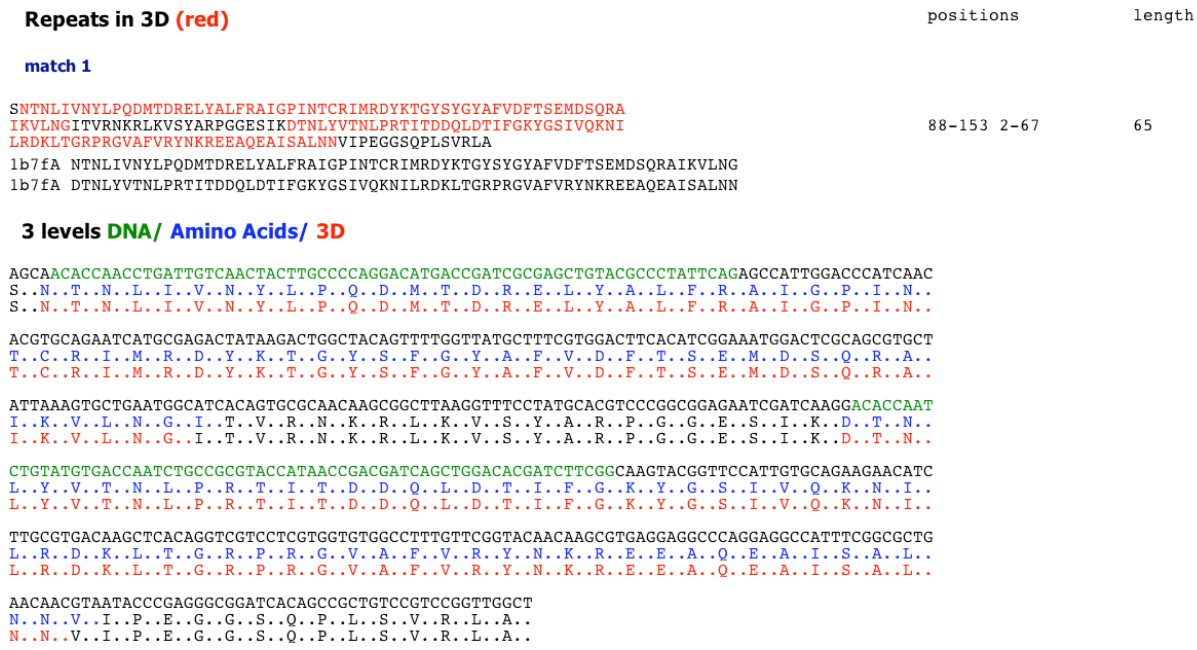


Figure 27 : Alignement aux trois niveaux.

- *Jmol*

Les répétitions trouvées en structures sont affichées en 3D avec Jmol. Les répétitions sont visibles en couleur (Figure 28). Il est possible de visualiser successivement plusieurs répétitions.

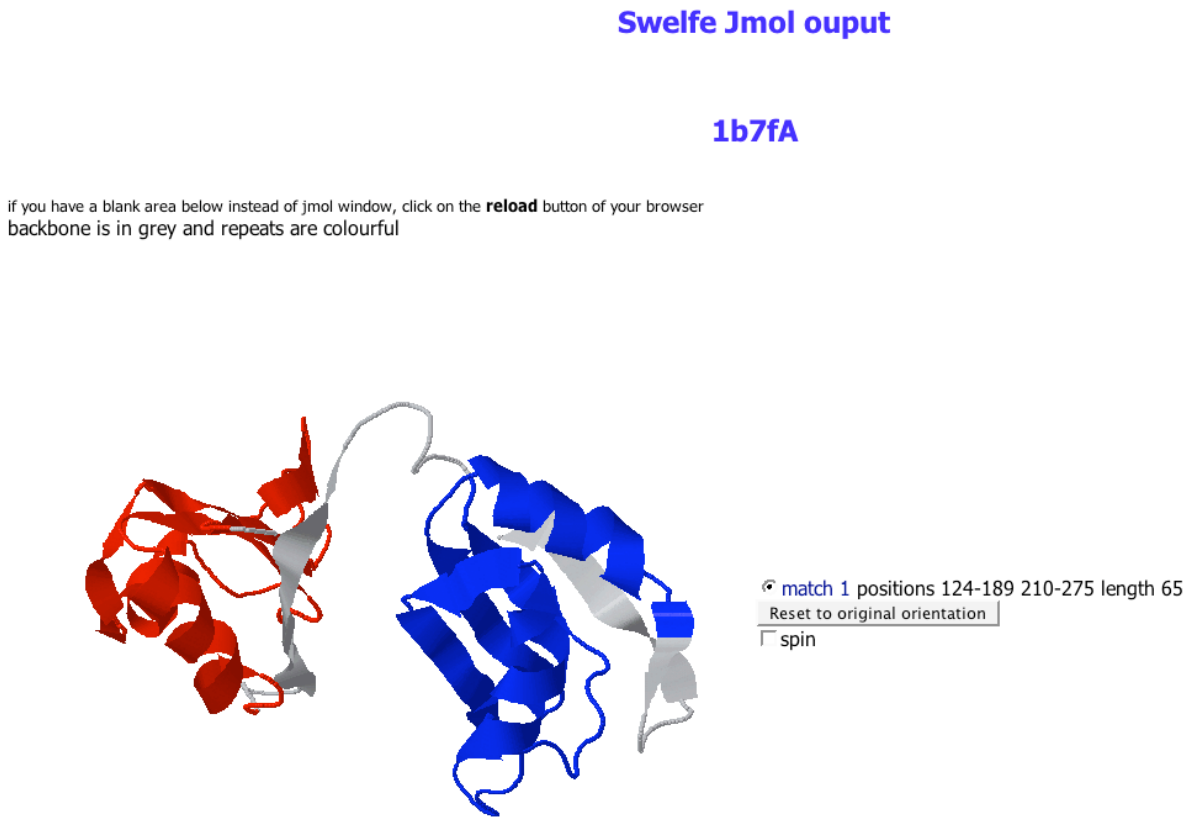


Figure 28 : Visualisation de l'alignement avec Jmol.

IV Comparaison des répétitions trouvées dans les séquences et les structures des protéines

Les répétitions trouvées par Swelfe dans les séquences d'ADN, les séquences d'acides aminés et les structures tridimensionnelles des protéines sont présentées et analysées ici.

IV.A Les données utilisées

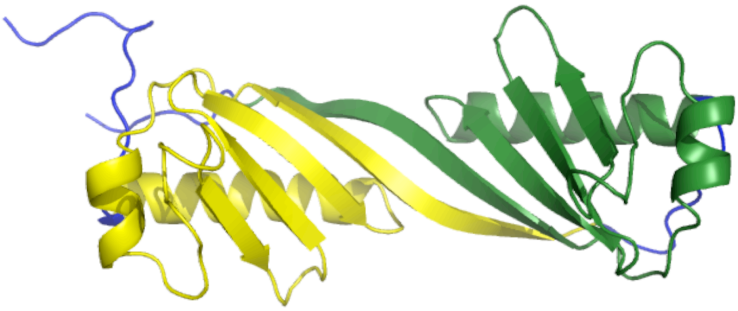
Nous sommes partis d'un jeu de données non redondant de séquences et de structures, issues de l'ensemble Cluster50[10] de la PDB (moins de 50% d'identité de séquence). Les séquences nucléiques et protéiques ont été extraites des CDS de TrEMBL. Ce jeu de données contient 7952 protéines aux trois niveaux et est issu de la banque contenant les protéines aux trois niveaux décrite précédemment (III.E).

Les paramètres de Swelfe ont été choisis de façon à obtenir un nombre important de répétitions qui satisfont des critères précis de significativité. Les répétitions sélectionnées pour les séquences nucléiques et protéiques ont une p-value < 0,01 calculée à partir de 1000 séquences aléatoires. Les répétitions structurales ont un score supérieur à 250° (correspond à 7 ou 8 identités) et un RMSD < 4,5Å. Au moment où ces calculs ont été faits, le RRMSD (Relative RMSD) n'était pas inclus dans Swelfe.

IV.B Les répétitions trouvées à chaque niveau

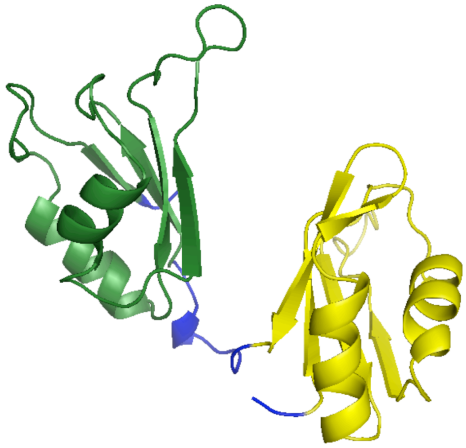
Swelfe a permis de trouver des répétitions dans 727 structures (9,1%), 654 séquences d'acides aminés (8,2%) et 264 séquences nucléiques (3,3%). Des exemples de répétitions sont montrés sur les Figures 29 et 30. Comme attendu, les répétitions trouvées dans les séquences d'ADN sont plus courtes qu'aux autres niveaux, car les séquences d'ADN évoluent plus vite. Cependant, de façon étonnante au premier abord, les protéines pour lesquelles des répétitions sont trouvées aux trois niveaux en parallèle ne sont pas majoritaires.

[10] Le sous ensemble non redondant Cluster50 est créé à l'aide du programme CD-HIT (Li and Godzik, 2006). Ce programme cherche des mots exacts de taille fixes communs à au moins deux séquences des structures de la PDB, puis aligne les séquences partageant plusieurs mots. Le taille et le nombre de mots exacts dépend du seuil de similarité entre les séquences d'un même groupe (entre 50 et 95% d'identité).



GATGAGATCCCGTATAAAGCAGTCGTAAATATAGAGAATATCGTTGCCACAGTGACTTTGGATCAAACATTGGATTTATATGCGATGGAAAGAAGCGTACCA
D..E..I..P..Y..K..A..V..V..N..I..E..N..I..V..A..T..V..T..L..D..Q..T..L..D..L..Y..A..M..E..R..S..V..P..
D..E..I..P..Y..K..A..V..V..N..I..E..N..I..V..A..T..V..T..L..D..Q..T..L..D..L..Y..A..M..E..R..S..V..P..
AACGTTGAATATGATCCTGATCAATTCCCAGGATTAATATTTAGGCTTGAATCTCCCAAGATAACCTCATTAATATTTAAATCAGGAAAAATGGTCGTTACT
N..V..E..Y..D..P..D..Q..F..P..G..L..I..F..R..L..E..S..P..K..I..T..S..L..I..F..K..S..G..K..M..V..V..T..
N..V..E..Y..D..P..D..Q..F..P..G..L..I..F..R..L..E..S..P..K..I..T..S..L..I..F..K..S..G..K..M..V..V..T..
GGAGCTAAAAGTACAGATGAGCTAATAAAGGCTGTAAAACGAATTATAAAAACCCTTAAAAAATATGGAATGCAACTAACAGGAAAACCTAAGATACAAATA
G..A..K..S..T..D..E..L..I..K..A..V..K..R..I..I..K..T..L..K..K..Y..G..M..Q..L..T..G..K..P..K..I..Q..I..
G..A..K..S..T..D..E..L..I..K..A..V..K..R..I..I..K..T..L..K..K..Y..G..M..Q..L..T..G..K..P..K..I..Q..I..
CAAAACATAGTCGCATCAGCTAATCTGCACGTTATAGTTAACCTTGATAAAGCAGCATTCCTGCTAGAGAATAACATGTACGAACCAGAGCAGTTCCCAGGT
Q..N..I..V..A..S..A..N..L..H..V..I..V..N..L..D..K..A..A..F..L..L..E..N..N..M..Y..E..P..E..Q..F..P..G..
Q..N..I..V..A..S..A..N..L..H..V..I..V..N..L..D..K..A..A..F..L..L..E..N..N..M..Y..E..P..E..Q..F..P..G..
CTAATATATAGAATGGATGAGCCCAGAGTTGTTCTATTAATTTTTAGCAGTGGTAAAATGGTTATTACAGGAGCTAAGAGAGAAGATGAAGTTCATAAGGCT
L..I..Y..R..M..D..E..P..R..V..V..L..L..I..F..S..S..G..K..M..V..I..T..G..A..K..R..E..D..E..V..H..K..A..
L..I..Y..R..M..D..E..P..R..V..V..L..L..I..F..S..S..G..K..M..V..I..T..G..A..K..R..E..D..E..V..H..K..A..
GTTAAAAAAATATTCGATAAACTGGTAGAGTTAGATTGTGTAAAGCCCGTTGAAGAAGAAGAGTTAGAA
V..K..K..I..F..D..K..L..V..E..L..D..C..V..K..P..V..E..E..E..E..L..E..
V..K..K..I..F..D..K..L..V..E..L..D..C..V..K..P..V..E..E..E..E..L..E..

Figure 29 : Les répétitions trouvées dans la protéine 1mp9 chaîne A, protéine de liaison à l'ADN de l'archée mésothermophile *Sulfolobus acidocaldarius*.
Protéine de 193 acides aminés. RMSD : 1,53 Å; RRMSD : 0,11. En haut : répétition structurale obtenue (les deux copies sont coloriées en jaune et en vert). En bas : la répétition d'ADN est indiquée en vert, la répétition d'acides aminés en bleu et la répétition structurale reportée sur la séquence d'acides aminés en rouge.



AGCAACACCAACCTGATTGTCAACTACTTGCCCCAGGACATGACCGATCGCGAGCTGTACGCCCTATTCAGAGCCATTGGACCCATCAACACGTGCAGAATCATGCGAGAC
S..N..T..N..L..I..V..N..Y..L..P..Q..D..M..T..D..R..E..L..Y..A..L..F..R..A..I..G..P..I..N..T..C..R..I..M..R..D..
S..N..T..N..L..I..V..N..Y..L..P..Q..D..M..T..D..R..E..L..Y..A..L..F..R..A..I..G..P..I..N..T..C..R..I..M..R..D..
TATAAGACTGGCTACAGTTTTGGTTATGCTTTCGTGGACTTCACATCGGAAATGGACTCGCAGCGTGCTATTAAAGTGCTGAATGGCATCACAGTGCGCAACAAGCGGCTT
Y..K..T..G..Y..S..F..G..Y..A..F..V..D..F..T..S..E..M..D..S..Q..R..A..I..K..V..L..N..G..I..T..V..R..N..K..R..L..
Y..K..T..G..Y..S..F..G..Y..A..F..V..D..F..T..S..E..M..D..S..Q..R..A..I..K..V..L..N..G..I..T..V..R..N..K..R..L..
AAGGTTTCCTATGCACGTCCCGGCGGAGAATCGATCAAGGACACCAATCTGTATGTGACCAATCTGCCGCGTACCATAACCGACGATCAGCTGGACACGATCTTCGGCAAG
K..V..S..Y..A..R..P..G..G..E..S..I..K..D..T..N..L..Y..V..T..N..L..P..R..T..I..T..D..D..Q..L..D..T..I..F..G..K..
K..V..S..Y..A..R..P..G..G..E..S..I..K..D..T..N..L..Y..V..T..N..L..P..R..T..I..T..D..D..Q..L..D..T..I..F..G..K..
TACGGTTCCATTGTGCAGAAGAACATCTTGCGTGACAAGCTCACAGGTCGTCCTCGTGGTGTGGCCTTTGTTCGGTACAACAAGCGTGAGGAGGCCCAGGAGGCCATTTCG
Y..G..S..I..V..Q..K..N..I..L..R..D..K..L..T..G..R..P..R..G..V..A..F..V..R..Y..N..K..R..E..E..A..Q..E..A..I..S..
Y..G..S..I..V..Q..K..N..I..L..R..D..K..L..T..G..R..P..R..G..V..A..F..V..R..Y..N..K..R..E..E..A..Q..E..A..I..S..
GCGCTGAACAACGTAATACCCGAGGGCGGATCACAGCCGCTGTCCGTCCGGTTGGCT
A..L..N..N..V..I..P..E..G..G..S..Q..P..L..S..V..R..L..A..
A..L..N..N..V..I..P..E..G..G..S..Q..P..L..S..V..R..L..A..

Figure 30 : Les répétitions trouvées dans la protéine 1b7f chaîne A, protéine sxl-léthale de *Drosophila melanogaster*.
Protéine de 167 acides aminés. RMSD : 0,97 Å; RRMSD : 0,09. En haut : répétition structurale obtenue (les deux copies sont coloriées en jaune et en vert). En bas : la répétition d'ADN est indiquée en vert, la répétition d'acides aminés en bleu et la répétition structurale reportée sur la séquence d'acides aminés en rouge.

IV.C Comparaison des répétitions trouvées aux différents niveaux

La Figure 31 présente le nombre de protéines contenant des répétitions à un ou plusieurs niveaux, indépendamment du chevauchement possible de ces répétitions. Par exemple trente-six protéines contiennent des répétitions détectées par Swelfe dans leur séquence protéique et leur séquence nucléique, mais pas dans leur structure 3D. Sur les 7952 protéines à chaque niveau, 3,3% des séquences nucléiques, 8,2% des séquences protéiques et 9,1% des structures contiennent des répétitions à au moins un niveau. De façon surprenante, il y a un nombre important de protéines qui ne contiennent des répétitions que dans leur séquence nucléique ou protéique, et le recouvrement des trois niveaux est assez faible. Des explications à ces observations seront proposées plus loin.

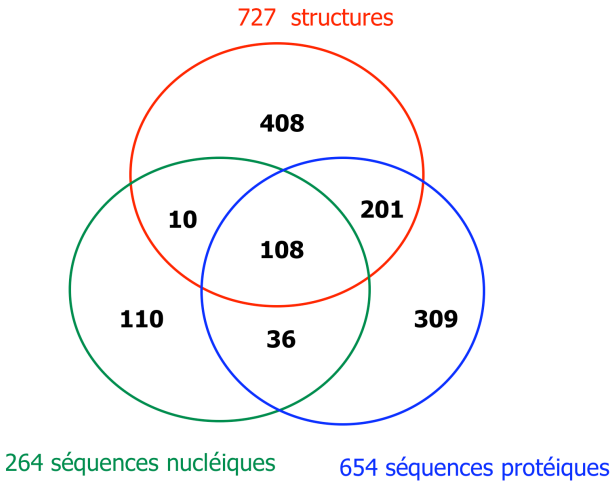
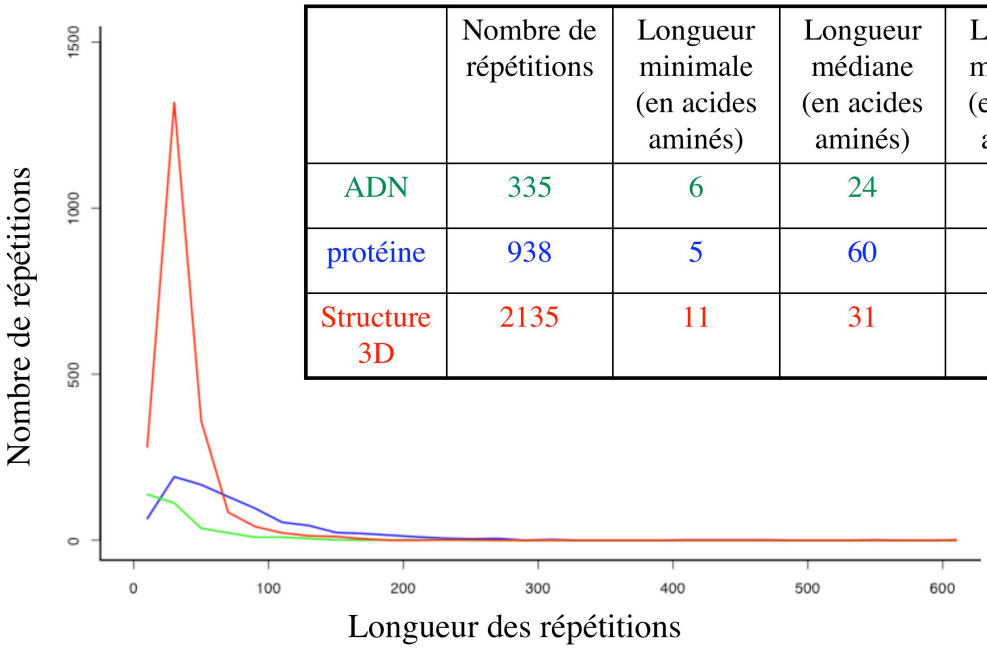


Figure 31 : Diagramme de Venn du nombre de protéines contenant des répétitions à chaque niveau.

De plus, les répétitions trouvées sont assez hétérogènes en terme de longueur. La Figure 32 montre que les répétitions structurales sont très nombreuses, mais assez courtes, les répétitions d'acides aminés sont longues et relativement nombreuses, tandis que les répétitions nucléiques sont peu nombreuses et courtes.



	Nombre de répétitions	Longueur minimale (en acides aminés)	Longueur médiane (en acides aminés)	Longueur maximale (en acides aminés)
ADN	335	6	24	224
protéine	938	5	60	606
Structure 3D	2135	11	31	244

Figure 32 : Longueur des répétitions (vert : ADN, bleu : acides aminés, rouge : structures 3D).

Nous avons ensuite voulu vérifier si les répétitions trouvées à plusieurs niveaux se chevauchent, c'est à dire si elles sont situées au même endroit dans la séquence. Le

chevauchement a été calculé avec deux méthodes différentes. Les deux copies de la répétition seront notée A et A' à un niveau et B et B' à un autre niveau.

La première méthode (Figure 33) considère que des répétitions sont chevauchantes à deux niveaux à deux conditions : les deux copies de la répétitions se chevauchent sur au moins cinq acides aminés (ou 15 nt) de long, et que les décalages de position de début des répétitions soient les mêmes, c'est à dire que : $\left|(D_B - D_A) - (D_{B'} - D_{A'})\right| \leq nombre\ de\ gaps$

Équation 28

D_A, $D_{A'}$, D_B, $D_{B'}$ sont les positions de début des répétitions (cf Figure 33). Si ces deux conditions sont respectées, les répétitions sont considérées comme chevauchantes.

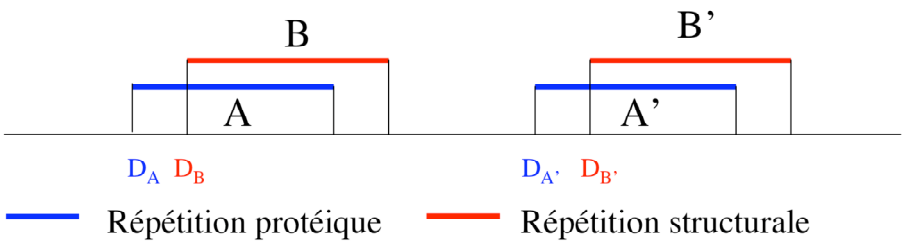


Figure 33 : Schéma de répétitions chevauchantes à deux niveaux (méthode 1).

Une autre méthode (Figure 34) permet de calculer le chevauchement des répétitions aux trois niveaux par le calcul du pourcentage de la séquence recouvert par des répétition à un, deux, ou trois niveaux. Pour faire ce calcul, seules les protéines ayant des répétitions à au moins un niveau ont été prises en compte.

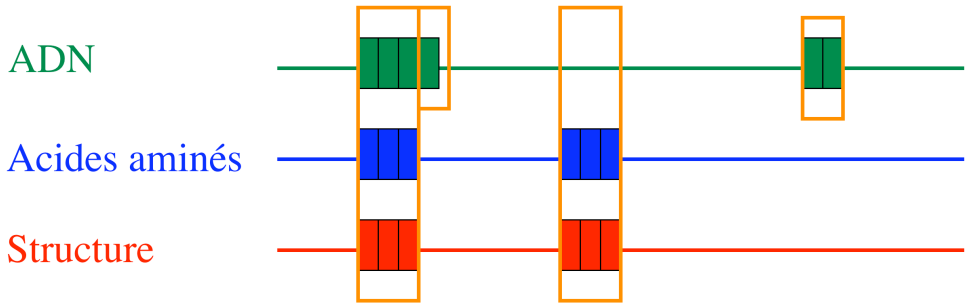


Figure 34 : Schéma de répétitions chevauchantes à plusieurs niveaux (méthode 2). Chaque rectangle de couleur représente un acide aminé (ou 3 nucléotides dans le cas de l'ADN). On comptabilise les acides aminés présents dans une répétition à un ou plusieurs niveaux pour toutes les séquences. Ce nombre est divisé par la taille totale de toutes les séquences.

Les résultats de ces deux analyses sont présentés sur la Figure 35.

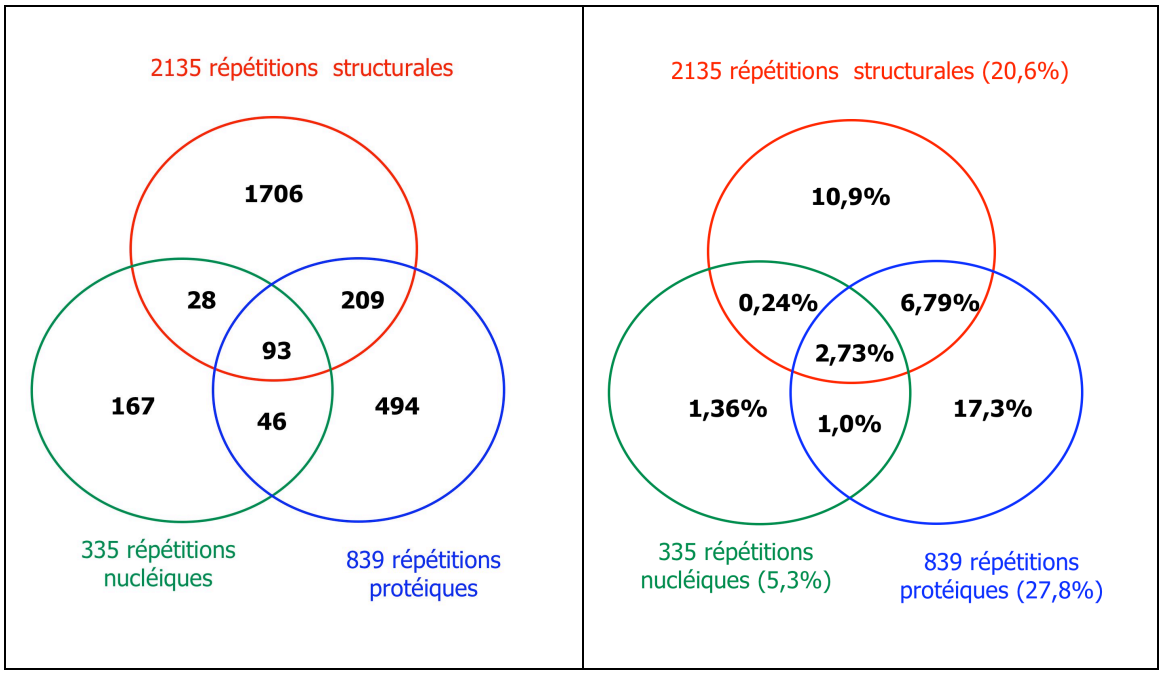


Figure 35 : Diagramme de Venn du recouvrement des répétitions aux trois niveaux (méthode 1 à gauche, méthode 2 à droite).
Seules les protéines contenant des répétitions à au moins un niveau sont prises en compte. A gauche : le total n'est pas toujours égal au nombre de répétitions, car certaines répétitions à un niveau correspondent à plusieurs répétitions à un autre niveau, en particulier, plusieurs répétitions structurales correspondent souvent à une seule répétition en acides aminés.

La Figure 35 (gauche) présente le nombre de répétitions chevauchantes à chaque niveau. Par exemple, il y a 93 répétitions qui sont trouvées à la fois dans les séquences nucléiques, les séquences protéiques et les structures tridimensionnelles. La Figure 35 (droite) présente le pourcentage de recouvrement des répétitions pour les protéines qui contiennent au moins une répétition à un niveau. Le pourcentage indiqué à coté du nombre de répétitions total est la somme des recouvrement des séquences/structures à un niveau. Par exemple, les répétitions structurales couvrent en moyenne 20,6% des structures qui ont au moins une répétition à un niveau. Les répétitions présentes aux trois niveaux couvrent en moyenne 2,73% des séquences/structures, les répétitions trouvées seulement dans les séquences d'ADN et d'acides aminés couvrent en moyenne 1,0% des séquences/structures.

La Figure 35 indique qu'une part importante des répétitions se chevauche sur au moins deux niveaux. Cela est vrai en particulier pour les répétitions trouvées en ADN qui sont souvent trouvées également en acides aminés et/ou en structure. Pour les autres répétitions en ADN, deux explications sont possibles. Tout d'abord, il est possible qu'il y ait un décalage du cadre de lecture entre la 1ère copie et la 2ème copie de la répétition, mais cela ne représente qu'une minorité de répétitions. Ensuite, la Figure 35 semble indiquer que les répétitions nucléiques qui ne sont présentes qu'à un niveau sont souvent petites (nombre important de répétitions Figure 35 à gauche, mais faible recouvrement de la protéine Figure 35 à droite), cela est confirmé par la Figure 36. Il est possible que les petites répétitions nucléiques ne puissent pas être retrouvées en acides aminés. En effet, un codon d'ADN apporte plus d'informations qu'un acide aminé : il y a 64 codons possibles et seulement 20 acides aminés, et donc en utilisant le même seuil statistique, une répétition exacte d'ADN pourra dans certains cas être significative pour une taille plus petite que la même répétition traduite en acides aminés. De plus, les répétitions en ADN tiennent compte de la composition de la séquence. Ainsi une séquence répétée en nucléotides peu fréquents pourra n'être trouvée qu'à ce niveau. Concernant les répétitions trouvées dans les séquences nucléiques et les structures mais pas dans les séquences protéiques, elles sont très rares et le chevauchement observé est largement dû au hasard.

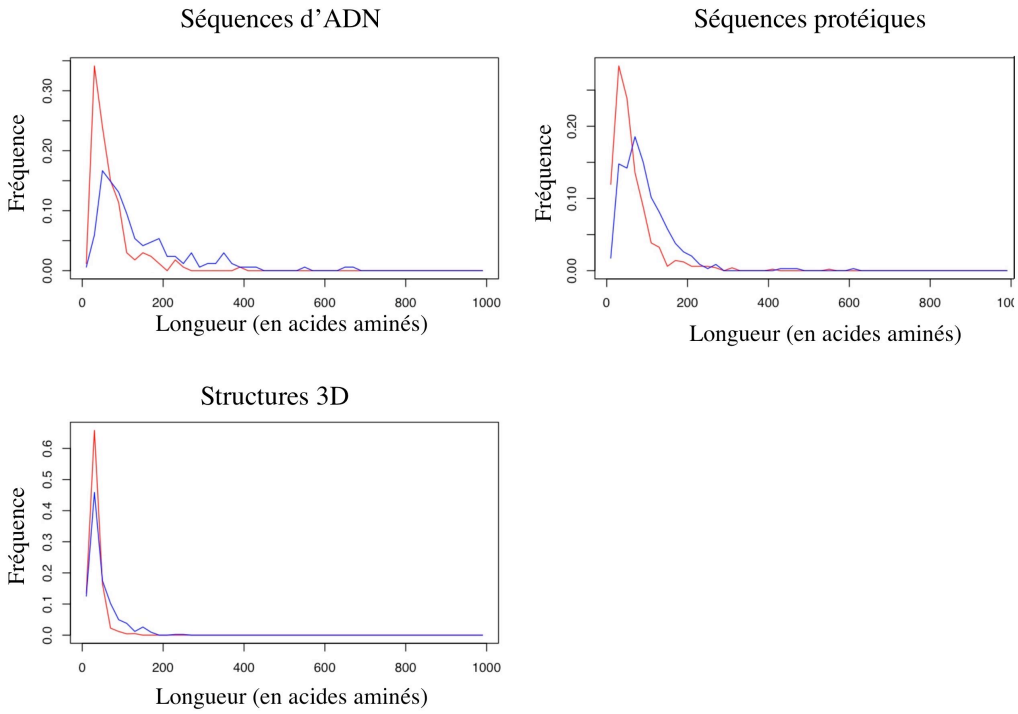


Figure 36 : Longueur des répétitions trouvées à un ou plusieurs niveaux.
Les courbes rouges représentent les répétitions qui sont également trouvées aux autres niveaux et les courbes bleues, les répétitions qui ne sont trouvées qu'à ce niveau.

Il existe également des différences entre les répétitions trouvées en acides aminés et en structures. Les répétitions trouvées dans les séquences d'acides aminés sont souvent longues, et il arrive que plusieurs répétitions structurales d'une protéine correspondent à une seule répétition observée en acides aminés. De plus les gaps peuvent être trouvés fréquemment dans les séquences d'ADN ou d'acides aminés, alors qu'ils sont beaucoup moins fréquents dans les structures. En effet, le poids de gap dans les alignements structuraux est proportionnellement plus élevé que pour les séquences (équivaut à 6 ou 7 angles α presque identiques contre 1 à 2 identités pour les séquences pour l'ouverture de gap). Certaines répétitions protéiques contenant beaucoup de gaps pourraient donc ne pas être retrouvées dans les structures. La plupart des gaps sont présents dans les boucles des structures (Pascarella and Argos, 1992), mais Swelfe ne différencie pas les boucles des autres structures secondaires et nous avons donc gardé un poids de gap élevé. De plus, le score de structure est assez pénalisant pour les motifs structuraux fréquents comme les hélices α et les feuillets β, et les répétitions structurales riches en hélices α ou feuillet β sont donc pénalisées, alors qu'elles ne le sont pas au niveau des séquences d'acides aminés. De plus, dans la PDB, les fragments d'au moins 12 acides aminés identiques ont une structure similaire (Rooman et al., 1990), et les fragments de moins de 12 acides aminés identique peuvent avoir des

structures différentes. Les petites répétitions protéiques pourraient donc ne pas être trouvées en structure. La Figure 36 indique que les répétitions protéiques qui ne sont pas trouvées aux autres niveaux sont plus petites. Une répétition significative à un niveau ne l'est pas forcément à un autre.

Enfin, il y a un nombre important de répétitions qui ne sont présentes qu'en structure. Ce résultat est attendu car les structures évoluent moins vite que les séquences. Le score de similarité de séquence des répétitions qui ne sont trouvées qu'au niveau structural est faible et est en moyenne négatif (Figure 37). Cela signifie soit que les séquences ont évolué et que leur similarité n'est plus détectable, soit que ces structures se ressemblent par convergence ou par hasard. Le score pénalise les structures secondaires fréquentes afin de limiter les répétitions trouvées par hasard. La Figure 32 montre que la plupart des répétitions structurales sont petites, et la Figure 35 indique que les répétitions qui ne sont trouvées qu'à ce niveau sont souvent petites (nombre important mais faible recouvrement de la protéine).

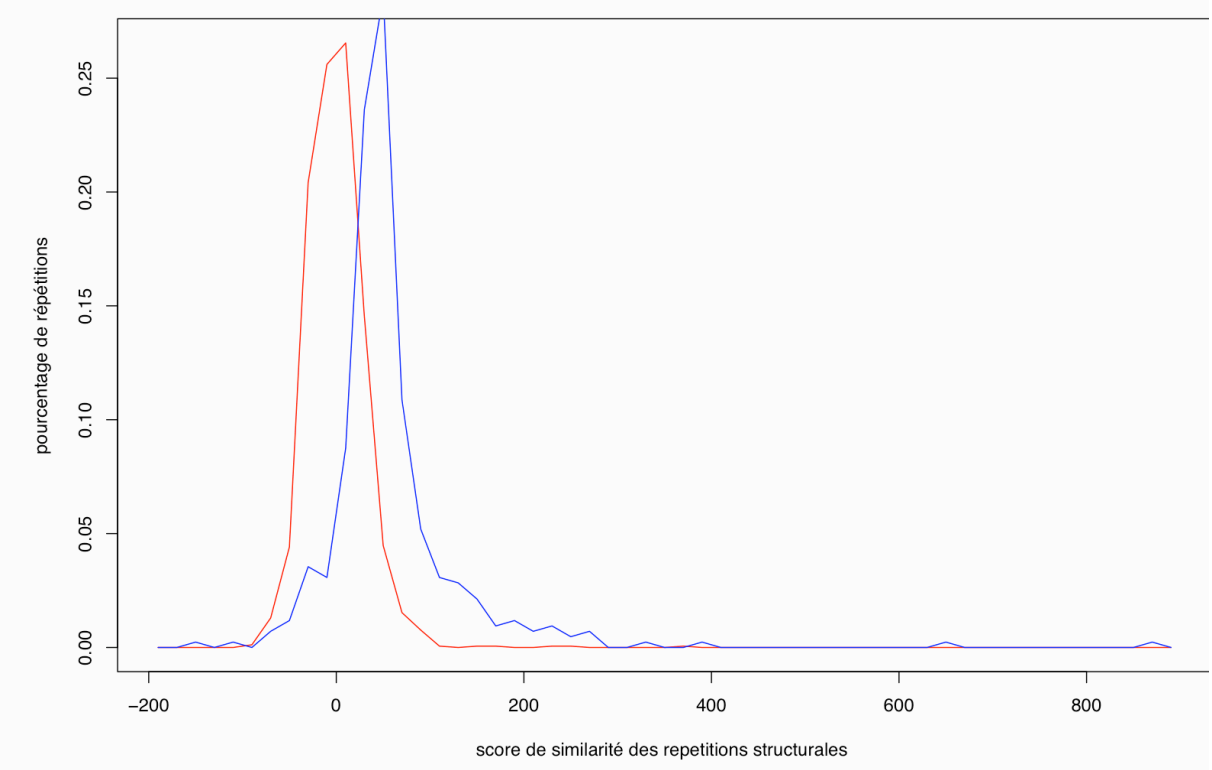


Figure 37 : Score de similarité de séquence des répétitions structurales.
Les répétitions trouvées seulement à ce niveau sont en rouge et les répétitions trouvées aux autres niveaux sont en bleu. Le calcul de similarité de séquence est effectué à partir de l'alignement structural, et utilise la matrice BLOSUM62.

IV.D Comparaison avec les domaines

De nombreuses études de duplications intragéniques ont été faites au niveau des domaines (par exemple (Apic et al., 2001b ; Bjorklund et al., 2006; Ekman et al., 2005)). Pour savoir si l'étude de répétitions trouvées *ab initio* par Swelfe aurait pu donner les même résultats qu'avec une étude de répétitions de domaines, nous avons comparé les répétitions obtenues par Swelfe aux répétitions de domaines Pfam. Nous avons utilisé le programme de détection des domaines par HMM de Pfam: HMMER2. Les domaines de Pfam sont définis par un profil HMM et ce programme permet de chercher les profils de tous les domaines dans une banque de séquence.

Le nombre de répétitions trouvées par chaque méthode est présenté dans le Tableau 7. Le chevauchement entre les répétitions de domaines Pfam et les répétitions trouvées par Swelfe est présenté dans le Tableau 8. Il est calculé avec la première méthode utilisée pour calculer le chevauchement des répétitions à plusieurs niveaux (chevauchement de 5 acides aminés, même décalage de phase). Il peut arriver que plusieurs répétitions de Swelfe correspondent à un domaine Pfam ou inversement que plusieurs domaines Pfam correspondent à une répétition trouvée par Swelfe. Les répétitions trouvées par Swelfe dans les séquences protéiques et les structures couvrent environ la moitié des répétitions de domaines, par contre, les répétitions de domaines correspondent à moins de 15% des répétitions trouvées par Swelfe à chaque niveau.

Tableau 7 : Nombre de répétitions de domaines et nombre de répétitions aux trois niveaux trouvées par Swelfe.

	Nombre de répétitions	Nombre de protéines contenant des répétitions
Pfam	232	214
Répétitions nucléiques	335	264
Répétitions d'acides aminés	839	654
Répétitions structurales	2135	727

Tableau 8 : Chevauchement entre les répétitions aux trois niveaux et les répétitions de domaines.

	Acides aminés	ADN	Structures
Nombre de domaines Pfam dupliqués correspondant à une répétition de Swelfe	132	50	111
Nombre de répétitions trouvées par Swelfe correspondant à un domaine Pfam dupliqué	124	50	141
% des répétitions de domaines correspondant aux répétitions trouvées par Swelfe	56,9%	21,5%	47,8%
% des répétitions trouvées par Swelfe correspondant à des répétitions de domaines	14,8%	14,9%	6,6%

Une étude plus précise des différences observées montre que les répétitions d'acides aminés qui ne correspondent pas à des domaines sont plus petites (Figure 38).

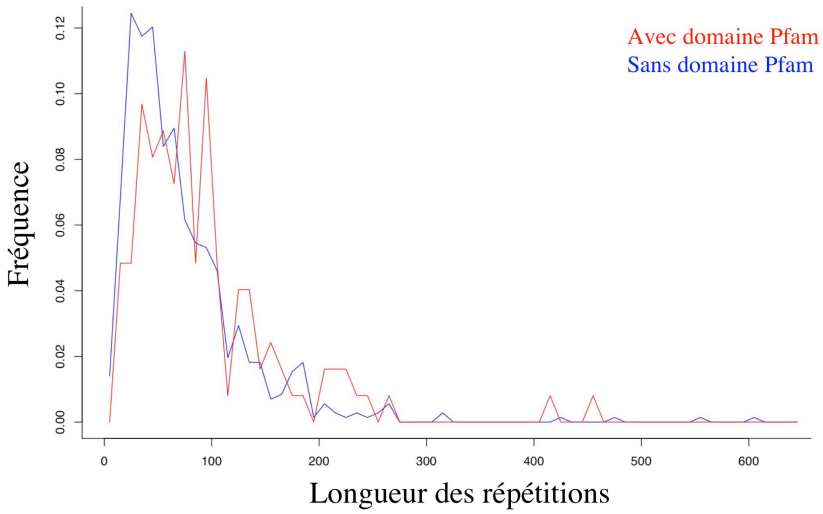


Figure 38 : Taille des répétitions d'acides aminés trouvées par Swelfe correspondant (en rouge) ou non (en bleu) à des répétitions de domaines.

Les répétitions d'acides aminés correspondant à des domaines semblent avoir à peu près la même taille que les domaines (Figure 39), alors que les répétitions structurales sont plus petites (Figure 40).

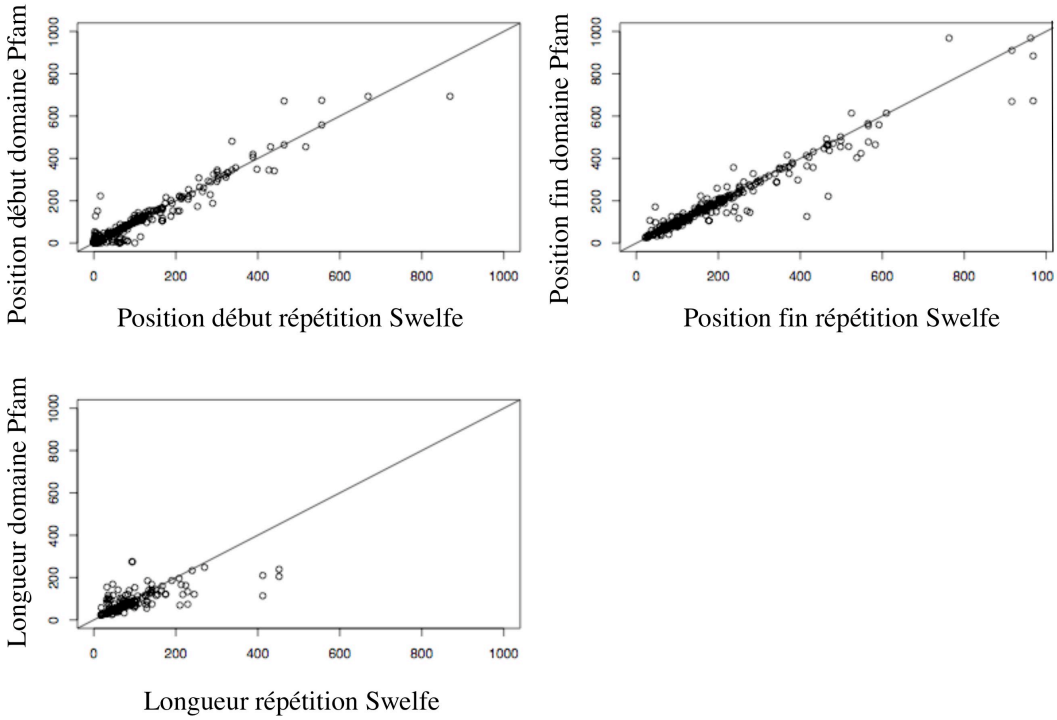


Figure 39 : Comparaison des répétitions d'acides aminés trouvées par Swelfe et des duplications de domaines Pfam concernant la position dans la séquence et la longueur.

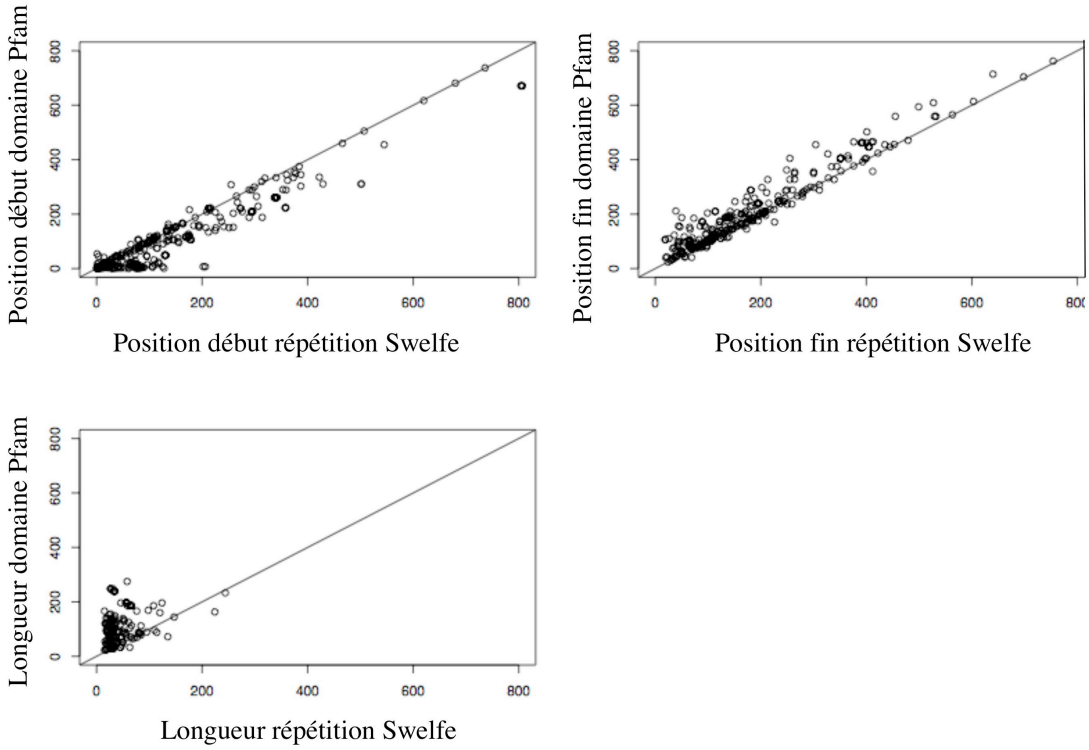


Figure 40 : Comparaison des répétitions structurales trouvées par Swelfe et des duplications de domaines Pfam concernant la position dans la séquence et la longueur.

Cette comparaison des répétitions trouvées par Swelfe et des domaines Pfam montre que les domaines ne couvrent pas toutes les répétitions trouvées aux trois

niveaux, en particulier aux niveaux des répétitions nucléiques et structurales. Cela peut s'expliquer par le fait que les profil HMM de Pfam sont créés à partir d'alignement de domaines protéiques. Il semble également que les domaines Pfam ne discernent pas les répétitions protéiques petites. De plus, les domaines Pfam ne couvrent pas toutes les protéines connues : en 2005, ils couvraient en moyenne 70% des gènes d'un génome (Lee et al., 2005). De ce fait, il est probable que certaines répétitions protéiques ne sont pas trouvées par Pfam. Cette comparaison indique que les résultats trouvés ici n'auraient pas pu être tous trouvés par une étude de domaines.

IV.E Conclusion

Swelfe a trouvé un nombre important de répétitions aux trois niveaux : environ 10% des protéines du jeu de données non redondant contiennent une répétition à au moins un niveau. Les différences entre les trois niveaux sont cependant assez importantes. Beaucoup de répétitions sont trouvées seulement au niveau des structures ou des séquences protéiques, les causes possibles ont été discutées ci-dessus. Ces différences montrent bien l'intérêt de l'étude des répétitions aux trois niveaux en parallèle. La comparaison avec les domaines Pfam indique que les études de répétitions de domaines ne permettraient pas de trouver autant de répétitions, et que Pfam retrouve particulièrement difficilement les répétitions nucléiques ou structurales.

V Étude des répétitions structurales pseudo-symétriques à 180°

Parmi les répétitions structurales, nous avons observé un certain nombre de répétitions dont les copies présentent une symétrie à 180°. Nous avons décidé de les étudier plus en détails afin de quantifier ce type de répétition, et de comprendre comment ces protéines se situeraient par rapport à d'autres éléments symétriques comme les homo-dimères.

V.A Les répétitions longues

Nous avons besoin d'un jeu de données non redondant pour chercher les répétitions longues. En effet, la PDB compte beaucoup de variants de la même protéine : d'après les données des statistiques actuelles de la PDB, les 53 521 structures peuvent être regroupées en 18 803 chaînes ayant moins de 90% d'identité de séquence entre elles. De plus, le nombre de repliements nouveaux (déterminés par SCOP) ajoutés à la PDB chaque année décroît depuis 2004 : le nombre de structures croît beaucoup plus vite que le nombre de nouvelles structures. Nous avons tout d'abord utilisé, pour une partie des analyses, le sous-ensemble Cluster50 de la PDB, dont les structures ont moins de 50% de séquence d'identité entre elles, qui est créé à partir de CD-HIT (Li and Godzik, 2006). Ce sous-ensemble comporte néanmoins un peu de redondance, et en créant la banque aux trois niveaux nous avons trouvé environ une centaine de structures qui correspondent à la même séquence protéique qu'une autre structure (soit environ 1,2% de redondance en ne prenant en compte que les séquences à priori identiques). Nous avons donc cherché si d'autres jeux de données non redondant pouvaient résoudre ce problème, et nous avons donc utilisé Astral. Les jeux non redondants d'Astral sont faits à partir du découpage en domaines des protéines. Le représentant de chaque groupe est choisi en fonction de la qualité de la structure. Plusieurs jeux de données non redondant sont proposés. Le jeu de données basé sur les familles SCOP (1 représentant par famille) n'était pas satisfaisant car certaines protéines n'étaient pas représentées. Nous avons choisi le jeu de données Astral dont les structures ont moins de 50% d'identité de séquence. Comme ce jeu de données est basé sur les structures, il ne sera pas possible de trouver dans ce jeu de données à la fois une protéine avec un domaine, et une protéine avec deux fois ce même domaine. Cela aurait pu être handicapant pour

la recherche de répétitions, mais des calculs rapides ont montré que nous obtenions le même pourcentage de structures contenant des répétitions en utilisant Cluster50 ou Astral, nous avons donc conservé ce dernier jeu de données.

Nous avons utilisé les structures de la base de données Astral (Chandonia et al., 2004) dont les domaines ont moins de 50% d'identité de séquence. Ce jeu de données est basé sur les domaines, mais comme nous nous intéressons aux protéines entières, nous avons conservé toute la structure à chaque fois (8657 structures). Ce jeu de données non redondant permet de ne pas prendre en compte les structures trop proches. Nous souhaitions un jeu de données moins redondant, et nous avons supprimé les structures avec plus de 40% d'identité de séquence entre elles. Les répétitions structurales longues ont été cherchées avec Swelfe grâce aux paramètres suivants : longueur de la répétition supérieure à 50 acides aminés, poids d'ouverture de gap de 70 et d'extension de 30, score>200, RRMSD<0,5. 172 protéines contiennent des répétitions longues soit 2% des protéines du jeu de données de départ. Le fait de sélectionner les répétitions longues permet de diminuer le risque que ces répétitions soient trouvées par hasard.

V.B Quel est le nombre de répétitions structurales pseudo-symétriques ?

V.B.1 Angle de rotation pour superposer les deux copies d'une répétition

L'angle permettant de superposer les deux copies de la répétition est calculé (III.F.4). La Figure 41 indique que 61 répétitions, soit environ 35% des longues répétitions structurales, sont pseudo-symétriques à 180° (valeur absolue de l'angle de rotation comprise entre 170° et 180°), ce qui correspond à une répétition avec une symétrie C2 (symétrie de rotation autour d'un axe d'un angle 360°/N, ici N=2). Il y a également des pics plus petits à 60°, 90°, 120° qui pourraient correspondre à des symétries C3, C4 ou C6. Au total 51% des protéines ont une symétrie (en prenant en compte les protéines qui ont une rotation d'une valeur absolue de ±5° autour de 60°, 90°, 120°, et entre 170° et 180°).

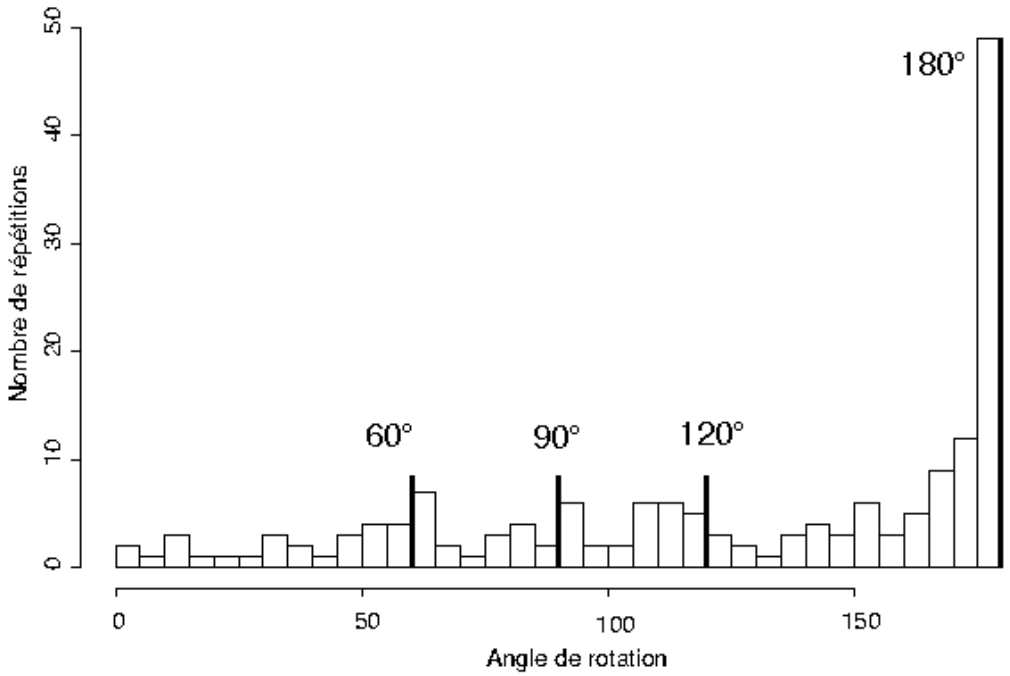


Figure 41 : Angles de rotation permettant de superposer les deux copies de la répétition (valeur absolue).

V.B.2 Définition de trois catégories : les protéines tout α, les protéines très répétées, les autres protéines

En observant les protéines contenant des répétitions longues, deux types de protéines particulières se détachent : les protéines riches en hélices α et les protéines très répétées. Les protéines restantes formeront la 3ème catégorie.

- *Les protéines riches en hélices α*

Les protéines dont plus de 80% de la structure est composée d'hélices α (angle α entre 40° et 65°) composent cette première catégorie. Les hélices α sont une structure secondaire très fréquente qui ne provient pas forcément d'une duplication. Le score de similarité de séquence est faible (Figure 42). La différence de score entre les répétitions des protéines riches et non riches en hélices α est significative (test de Wilcoxon unilatéral, p-value<0,01). Ces protéines ont été formées par une duplication ancienne, par un phénomène de convergence, ou plus vraisemblablement et simplement par le hasard, étant donné la fréquence de cette structure secondaire.

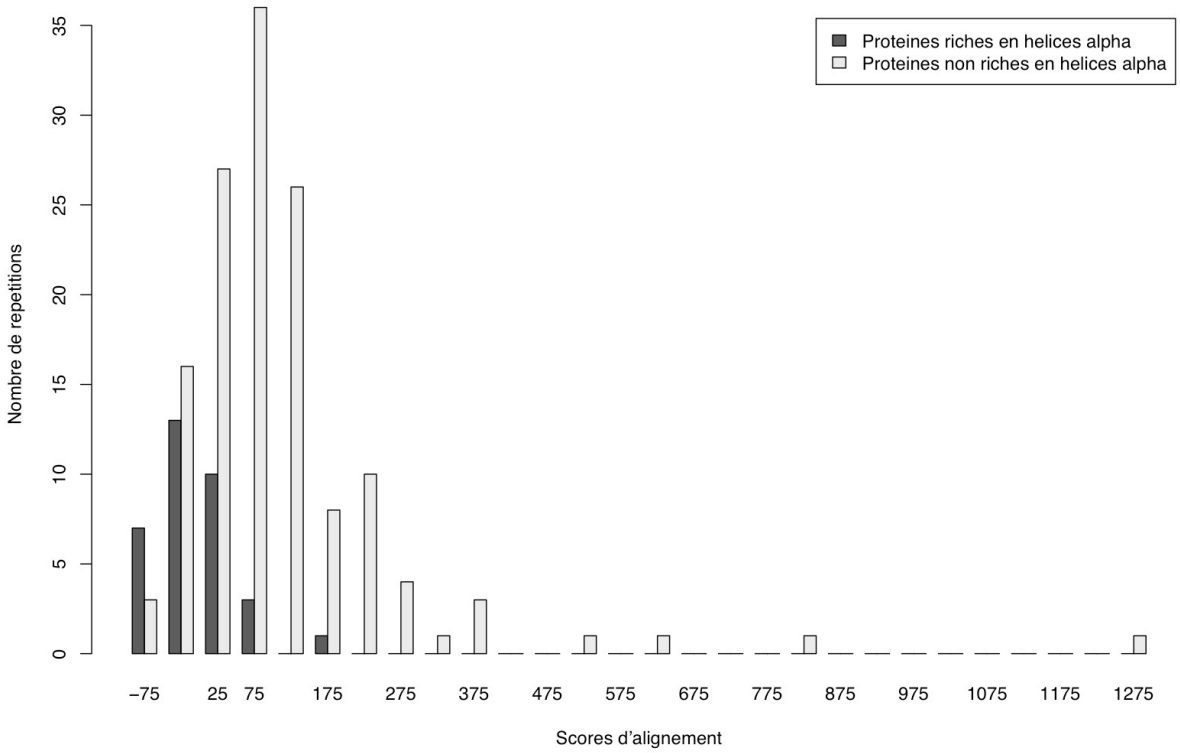


Figure 42 : Score de similarité de séquence (matrice BLOSUM62) des copies des répétitions des protéines riches et non riches en hélices α.

- *Les protéines très répétées*

Certaines protéines ont une structure très répétée. C'est le cas par exemple des hélices β (β propellers), des protéines riches en leucine ou des ankyrines (Figure 11). Certaines de ces protéines sont aussi très riches en hélices α et ont été classées dans la 1ère catégorie. Les autres ont été déterminées visuellement avec Pymol, et contiennent au moins six motifs répétés. Ces protéines ont probablement été créées par plus d'un événement de duplication et ne comportent que rarement une symétrie à 180° (Figure 43).

- *Les autres protéines*

Les protéines restantes sont placées dans la catégorie « autres protéines ». Ce sont les protéines les plus susceptibles d'avoir évolué par un seul événement de duplication, et sont donc des candidats préférentiels pour la recherche de protéines avec une répétition pseudo-symétrique. Cette classe compte 88 protéines dont 35 ont une répétition pseudo-symétrique à 180° (40%) (Figure 43).

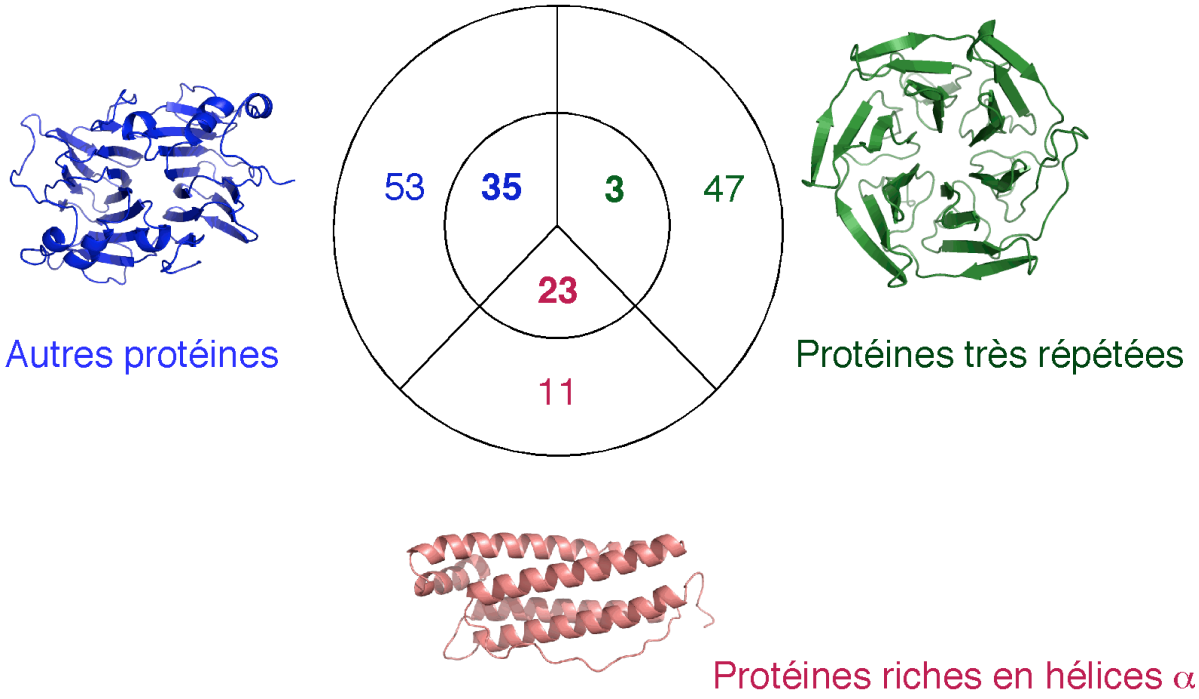


Figure 43 : Nombre de protéines de chaque catégorie et nombre de protéines dont les deux copies de la répétitions peuvent être superposées par une rotation de 180° autour d'un axe(au centre en gras).

- *Les protéines avec une répétition symétrique C3*

Il existe deux exemples de répétitions avec trois éléments répétés trouvés par Swelfe, un exemple est montré Figure 44. Ces répétitions pourraient remplacer des trimères.

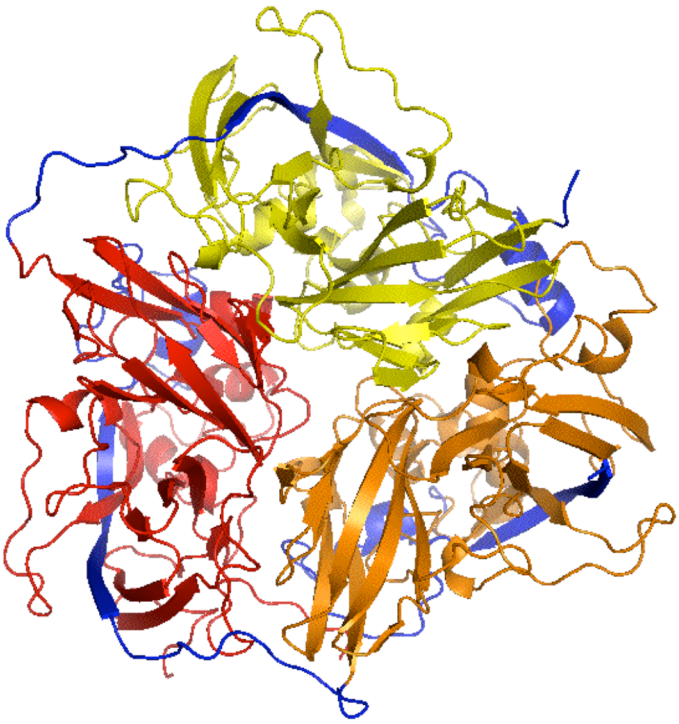


Figure 44 : La protéine 2j5w chaîne A (Ceruloplasmine d'*Homo sapiens*) présente une répétition dont les copies sont symétrique à 120° (en jaune, orange et rouge).

V.B.3 Comparaison avec les domaines de Pfam

Pour savoir si les mêmes résultats auraient pu être obtenus avec une étude de domaines, nous avons comparé les répétitions trouvées dans le jeu de données « autres protéines » avec les domaines Pfam (Finn et al., 2008) (Tableau 9). Une répétition trouvée par Swelfe et une répétition de domaines Pfam correspondent s'il y a au moins 70% de recouvrement entre les copies des répétitions. Cinq cas sont possibles :

- chaque copie de la répétition correspond à un domaine, et ces domaines ont le même nom (ligne 1) ;
- un domaine Pfam contient les deux copies de la répétition (ligne 2) ;
- chaque copie de la répétition correspond à deux domaines Pfam (ligne 3) ;
- chaque copie de la répétition correspond à un domaine Pfam, les deux domaines Pfam font partie du même « clan » (ligne 4) ;
- la ligne 5 inclut les protéines pour lesquelles les répétitions et les domaines Pfam ne se chevauchent pas assez ou pas du tout, ou des protéines qui n'ont aucun domaine Pfam (un cas).

Il s'avère qu'une bonne partie des répétitions trouvées correspondent à des répétitions de domaines (55 cas) mais il reste un peu moins d'un tiers des cas qui ne sont pas des répétitions de domaines Pfam. Cette analyse montre que les répétitions que nous avons trouvées ne sont pas en contradiction avec les répétitions de domaines Pfam, mais que notre étude ne pouvait pas se limiter à une étude des répétitions des domaines Pfam, dans la mesure où il n'est pas possible de calculer une symétrie structurale pour ces dernières.

Tableau 9 : Correspondances entre les domaines Pfam et les répétitions.

	Total	Symétrique	Non symétrique
La duplication de domaines correspond aux répétitions	55	21	34
Domaines Pfam contenant 2 copies d'une répétition	14	6	8
Chaque répétition correspond à 2 domaines	2	1	1
2 domaines presque identiques correspondent aux répétitions	3	1	2
Autres cas	14	6	8

V.C Est ce que les protéines avec une répétition pseudo-symétrique de type C2 peuvent remplacer des dimères ?

V.C.1 Recherche de protéines contenant une ou au moins trois copies de la répétition

Pour savoir si les protéines contenant une répétition dont les deux copies sont symétrique pourraient être des analogues fonctionnels de protéines ne contenant qu'une copie de cette répétition dans la même espèce ou dans une autre espèce, nous avons cherché des protéines ne contenant qu'une copie de la répétition, ou au moins trois copies, à la fois dans les bases de données de séquences et de structures.

La banque que nous avons utilisée, Astral, est une banque basée sur les domaines, et donc le jeu de données non redondant contient rarement à la fois une protéine avec

deux domaines identiques, et une protéine avec un seul de ces domaines. Cluster50 est basé seulement sur la similarité de séquences et contient des protéines similaires seulement sur une partie de leur chaîne. Il est donc plus susceptible de contenir à la fois des protéines avec une et deux copies de la répétition.

Pour les structures, la répétition en angles α a été cherchée contre Astral50 (Chandonia et al., 2004) et Cluster50 (Berman et al., 2000) avec Swelfe en comparaison deux à deux[11]. Les protéines qui avaient une différence de taille de moins de la moitié de la taille de la répétition par rapport à la protéine de départ ont été supprimées pour ne pas trouver de protéines identiques ou très proches. Enfin, une vérification visuelle avec Pymol a permis de confirmer si les protéines correspondaient bien à la moitié de la protéine initiale ou contenaient plusieurs répétitions dont une partie se superpose à la protéine initiale.

En ce qui concerne les séquences (Figure 45), les séquences des répétitions ont été cherchées contre les CDS de TrEMBL (Kulikova et al., 2007). Un premier filtre est effectué avec Blast (Altschul et al., 1990) pour diminuer la taille de la banque. Swelfe compare les séquences des répétitions avec la banque ainsi réduite. Enfin, un alignement global avec des poids de gap nuls aux extrémités est effectué (Matrice BLOSUM62 (Henikoff and Henikoff, 1992), ouverture de gap 1,2, extension de gap 0,8). Les séquences qui ont plus de 40% de similarité avec la séquence de départ et une différence de taille de plus de la moitié de la taille de la répétition avec la séquence d'origine sont conservées. Ensuite, une dernière vérification est effectuée :

1. Pour les protéines ne contenant qu'une copie, les positions de l'alignement « end-gap free[12] » sont vérifiées : il faut que les positions qui correspondent à une répétition aient moins de 20% de gaps et que les positions qui correspondent à l'autre répétition aient plus de 80% de gaps.

2. Pour les protéines qui ont des répétitions multiples, une copie de la répétition est cherchée avec Swelfe contre ces protéines et tous les alignements possibles sont calculés. Ensuite, les alignements successifs non chevauchants sont sélectionnés s'ils correspondent au moins aux 3/4 de la longueur de la répétition d'origine.

[11] Swelfe peut comparer deux séquences ou deux structures comme indiqué dans le paragraphe (III.G).
[12] Dans cet alignement, les gaps aux extrémité des séquences ont un poids nul.

Enfin, dans les deux cas, les protéines trop similaires (plus de 40% de similarités) sont retirées des analyses, à la fois pour les protéines ayant une et au moins trois copies de la répétition, afin de conserver des variants en nombre de répétition non redondants.

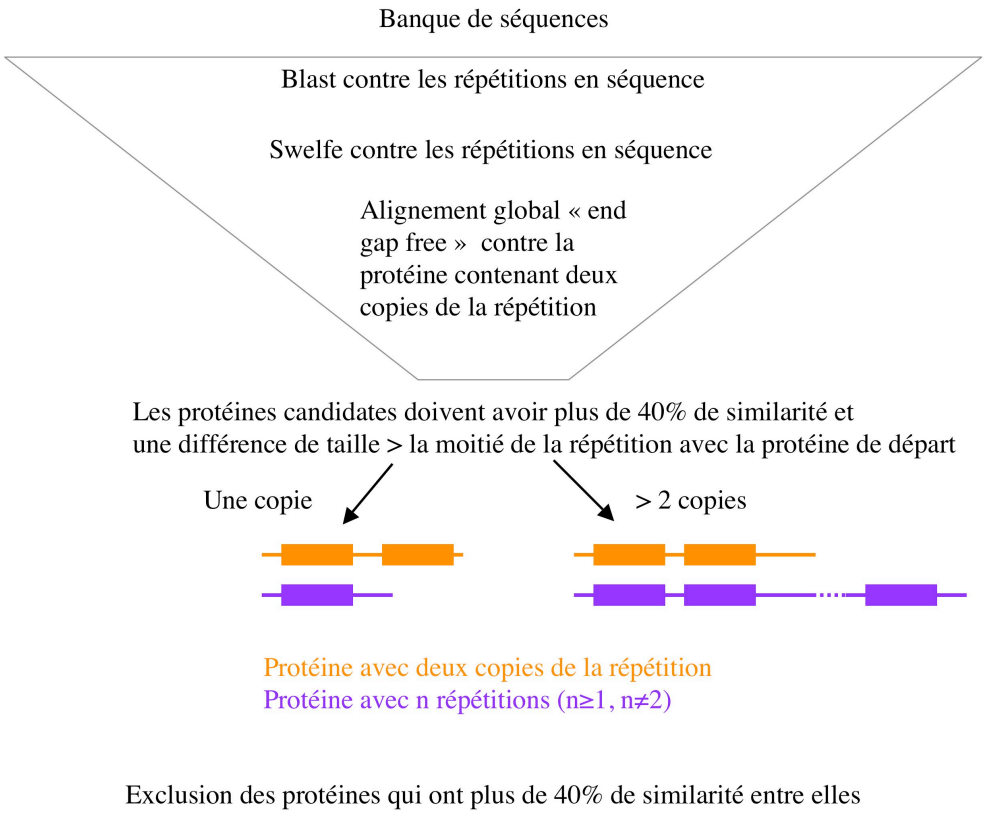


Figure 45 : Schéma de la méthode de recherche des protéines avec une ou au moins trois copies de la répétition en séquence.

Les résultats sont présentés dans le Tableau 10. Il existe un nombre important de protéines avec une seule répétition, sauf dans les données d'Astral, comme attendu.

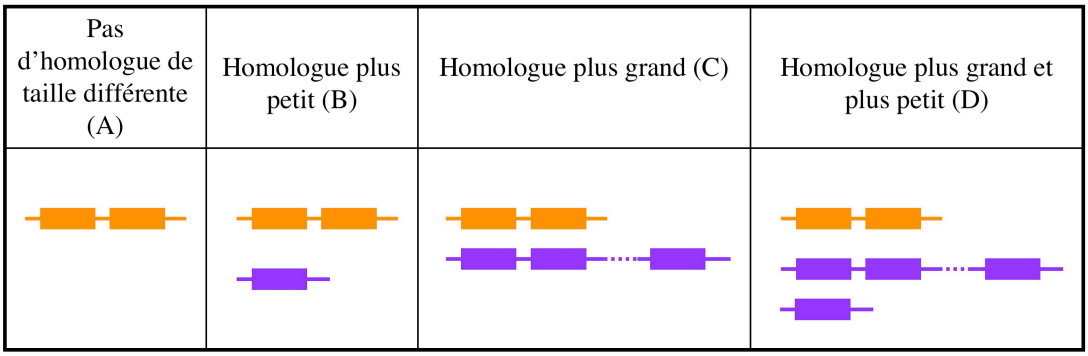


Figure 46 : Schéma de figures contenant une ou plusieurs copies de la répétition.

Tableau 10 : Nombre de protéines avec une ou au moins trois copies de la répétition, dans les CDS et dans la PDB (Astral 50 et Cluster50).
Les colonnes correspondent au schéma de la Figure 46. La colonne de droite représente les protéines pour lesquelles il existe une protéine avec une seule copie de la répétition, une protéine avec deux copies (celle à partir de laquelle est faite l'analyse), et une protéine (ou plus) contenant au moins trois copies.

	A	B	C	D
PDB (Astral 50)				
Protéines symétriques	32	3	0	0
Protéines non symétriques	49	3	1	0
PDB (Cluster50)				
Protéines symétriques	23	11	1	0
Protéines non symétriques	35	16	1	1
CDS				
Protéines symétriques	20	14	1	0
Protéines non symétriques	22	13	14	4

De plus, les protéines avec une répétition pseudo-symétrique pour lesquelles il existe des protéines analogues avec plus de deux copies de la répétition sont peu nombreuses, et ces protéines sont presque toutes non symétriques. D'autre part, en regroupant les données des trois banques pour chaque protéine, les résultats montrent la présence de quatorze protéines existant en au moins trois variants en nombre de copies (par exemple, une copie, deux copies et trois copies). Parmi ces protéines, seule une présente une symétre.

Il existe un exemple de protéine dont les copies sont symétriques à 180° pour lequel il existe une protéine avec quatre copies de la répétition (Figure 47).

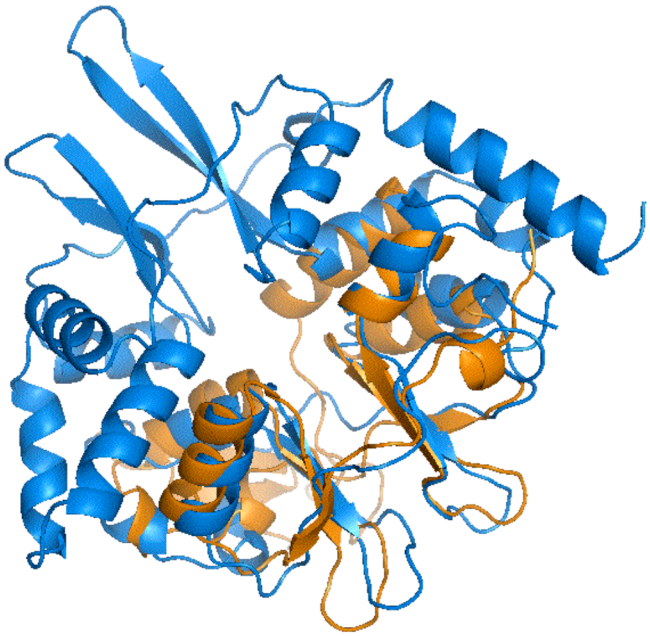


Figure 47 : Exemple de protéine pseudo-symétrique à 180° pour lequel il existe une protéine avec quatre copies de la répétition.
En orange : 1vpm chaîne A (protéine hypothétique de *Thermoplasma acidophilum*), en bleu : 2oox chaîne G (protéine hypothétique de *Schizosaccharomyces pombe*). La protéine 1vpmA présente une duplication symétrique à 180°. La protéine 2ooxG est le double de la protéine 1vpmA.

V.C.2 Les caractéristiques des protéines avec un nombre de copies différent de deux

Nous avons constaté que la taille des structures de la PDB correspond rarement à la taille du gène et cette différence peut être assez grande pour certaines protéines. En effet, les protéines de la PDB sont limitées en taille, et souvent tronquées, du fait de la difficulté à résoudre la structure des grosses protéines. Nous avons comparé la taille des protéines pseudo-symétriques et non symétriques dans la PDB et dans les CDS les plus proches, provenant de la banque expliquée au paragraphe (III.E) (Figure 48). Les tailles ne sont pas significativement différentes pour les protéines dans la PDB (test de Wilcoxon, p-value = 0,23). Par contre, la taille des CDS des protéines est significativement différente (test Wilcoxon, p-value = 0,0035). La taille des CDS des protéines non symétriques est donc plus élevée que celle des protéines pseudo-symétriques.

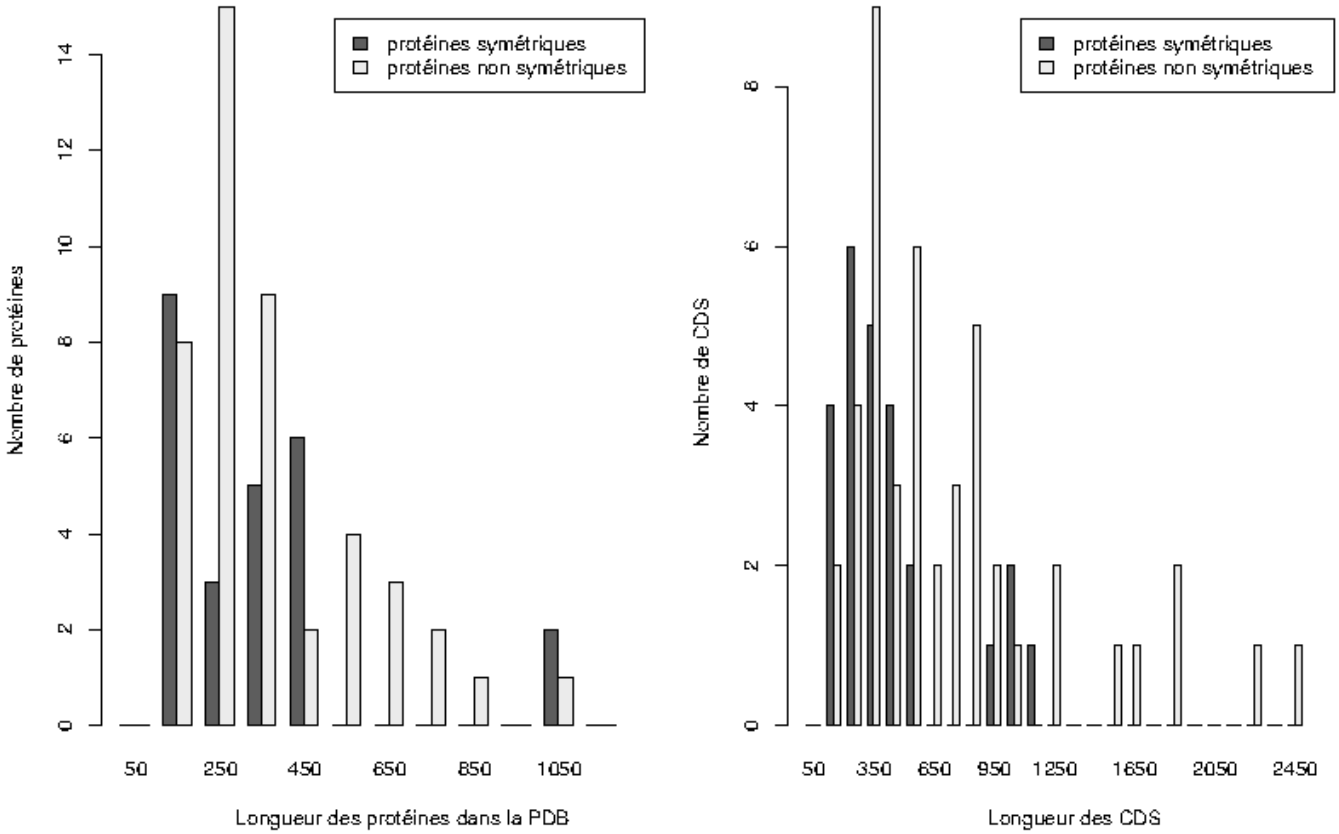


Figure 48 : Longueur des protéines dans la PDB et dans les CDS.
Seules les protéines qui ont un score d'alignement BLOSUM positif sont prises en compte.

Swelfe a été utilisé pour calculer le nombre de répétitions dans les gènes des protéines : trois protéines avec une répétition pseudo-symétrique contiennent au moins trois copies de la répétition, et quinze protéines non symétriques contiennent au moins trois copies. Cette différence est significative à 5% (chi2 : p-value = 0,03). Cela indique qu'une partie des gènes des protéines contenant une répétition non symétrique contient en fait au moins trois copies de la répétition.

Toutes ces analyses nous amènent à proposer une explication possible de la différence entre les protéines pseudo-symétriques et certaines protéines non symétriques. Les protéines pseudo-symétriques peuvent correspondre à une protéine avec une seule copie de la répétition, mais rarement à une protéine avec plusieurs copies. Les protéines non symétriques, pour une partie d'entre elles, peuvent exister avec un nombre variable de répétitions, et ont donc une taille plus grande en moyenne. Ces protéines pourraient donc exister avec un nombre variable de copies de la répétition, et ne s'agenceraient donc pas de façon symétrique.

V.C.3 Est ce que certaines protéines avec une répétition symétrique peuvent remplacer des homo-dimères ?

Les protéines contenant deux copies de la répétition symétriques à 180° ressemblent à des homo-dimères formés d'une seule protéine. Nous avons cherché à savoir si les protéines contenant une seule copie de la répétition pouvaient être des homodimères, qui pourrait dans certains cas être remplacés par une protéine contenant deux copies de la répétition symétriques à 180°. Nous avons donc comparé les degrés d'oligomérisation de ces deux types de protéines, pour savoir si celles qui ne contiennent qu'une copie de la répétition ont une structure quaternaire double de celle des protéines contenant deux copies symétriques de la répétition. Pour cela, nous avons utilisé les données de la banque PQS qui est une banque de structures quaternaires basée sur les structures de la PDB (Tableau 11). Cette banque ne contient pas d'information sur les monomères car elle ne peut pas trancher entre les protéines monomériques et les protéines oligomériques mais dont un seul monomère est présent dans le fichier PDB. De ce fait, un nombre important de protéines n'a pas de données dans PQS. Parmi les 35 protéines symétriques à 180°, il y a trois protéines dimériques contenant deux copies de la répétition qui sont similaires à trois autres protéines tétramériques qui ne contiennent qu'une copie de la répétition (Tableau 11 et Figure 49). Un exemple est présenté Figure 50. Il y a également un cas de trimère composé de deux copies de la répétition qui ressemble à un hexamère composé d'une copie de la répétition. Enfin, il existe deux cas pour lesquels il n'y a pas d'informations PQS, et qui pourraient donc être des monomères, et qui sont similaires à deux dimères contenant une seule copie de la répétition. Ces protéines représentent au total six protéines avec une répétition symétrique C2 qui pourraient remplacer des homo-dimères.

Tableau 11 : Protéines avec deux copies de la répétition, et dont l'état quaternaire est le double de protéines avec une seule copie de la répétition (issus de la banque PQS des dimères).

La structure quaternaire des protéines avec une ou deux copie de la répétition ont été comparées, lorsqu'elles étaient disponibles.

Protéines avec 2 copies de la répétition		Protéines avec une copie de la répétition	
Nom PDB	Etat quaternaire	Nom PDB	Etat quaternaire
1alnA	Homo dimère	1r5tA,1uwzA, 2fr5A	Homo tétramère
1bd7A	Homo dimère	1ha4A	Homo tétramère
1ddzA	Homo dimère	1i6oA	Homo tétramère
2gvhA	Homo trimère	1vpmA,1y7uA,1yliA	Homo hexamère
1gttA	Non annoté (Monomère ?)	1sawA	Homo dimère
1o7fA	Non annoté (Monomère ?)	1u12A	Homo dimère

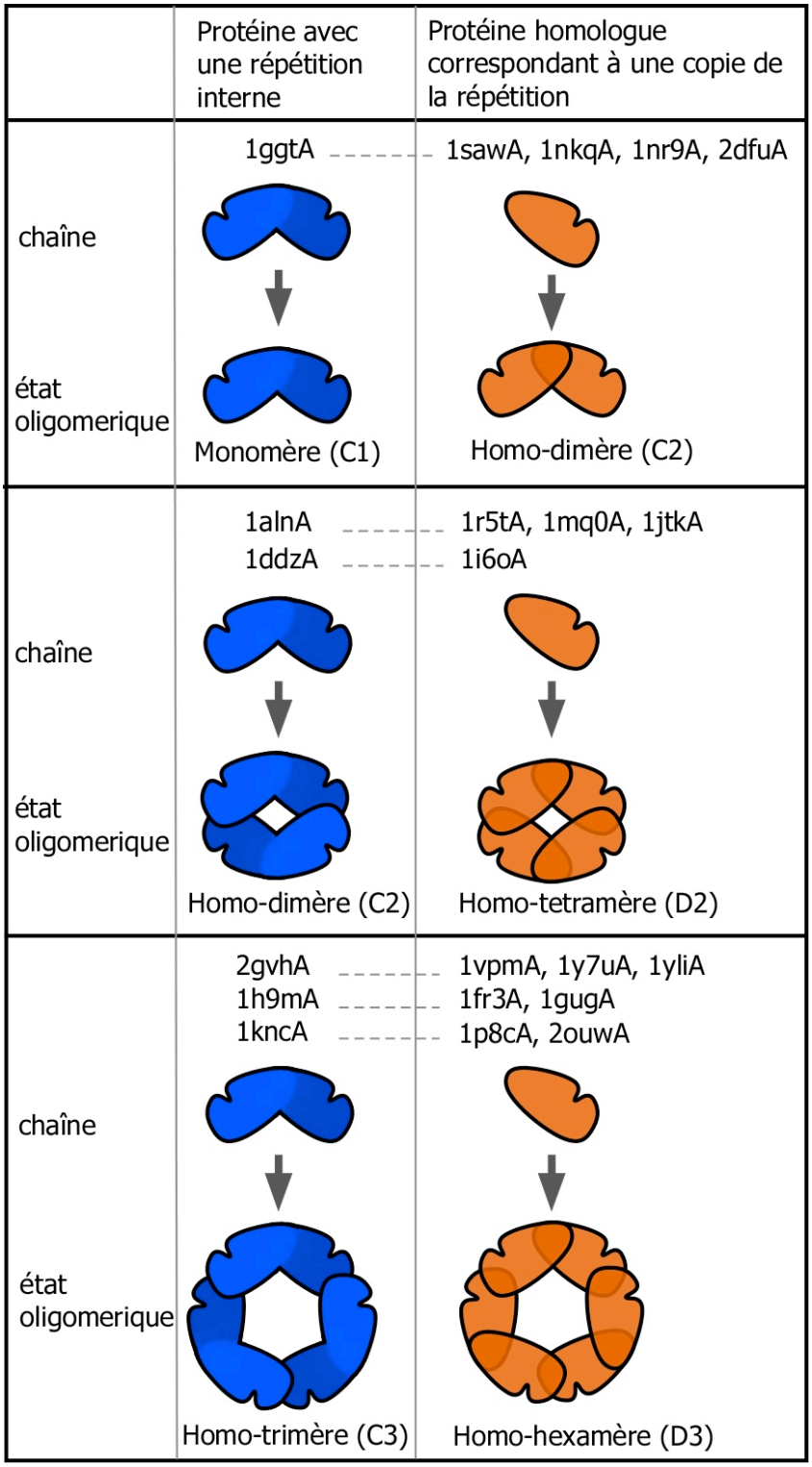


Figure 49 : Schéma des protéines avec deux copies de la répétition, et dont l'état quaternaire est le double de protéines avec une seule copie de la répétition.

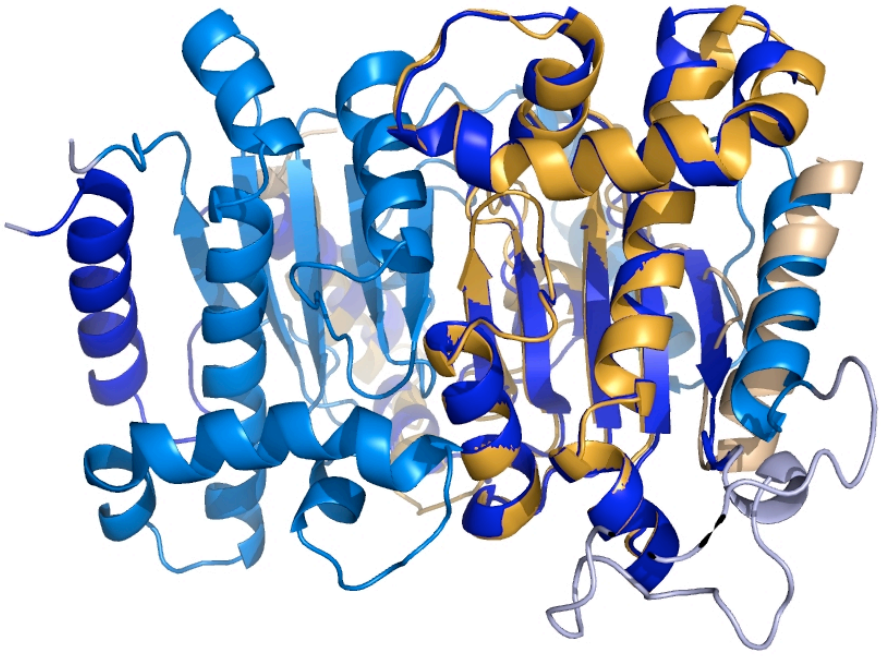


Figure 50 : Exemple de protéine avec deux copies de la répétition et une symétrie de 180° (en bleu) qui correspond à une protéine qui n'a qu'une copie de la répétition (en orange). 1ddz chaîne A (en bleu, anhydrase β-carbonique de *Porphyridium purpureum*) contient deux éléments répétés symétriques (en bleu et bleu marine) et a une structure quaternaire dimérique. 1i6o chaîne A (en orange, anhydrase β-carbonique d'*Escherichia coli*) a une structure quaternaire tétramerique et correspond (en orange vif) à la moitié de la chaîne A de 1ddz.

Dans l'hypothèse où une protéine avec une répétition pseudo-symétrique C2 pourrait remplacer fonctionnellement un homo-dimère, il est possible que les deux protéines ne co-existent pas chez la même espèce. Pour tester cette hypothèse, nous avons regardé les espèces dans la PDB chez lesquelles étaient présentes les protéines avec une ou deux copies de la répétition. De plus, comme les données de Cluster50 sont non redondantes, l'ensemble du groupe dont la protéine est le représentant a été pris en compte. Nous avons comparé toutes les espèces deux à deux et compté le nombre de fois où les deux formes de la protéine co-existent dans au moins une espèce. Il y a quatre protéines pseudo-symétriques qui co-existent avec leur équivalent mono-copie chez la même espèce contre quinze protéines non symétriques. Les résultats sont aux limites de la significativité statistique (test χ^2, p-value = 0,059). De plus ces résultats sont incomplets, et il faudrait faire une recherche dans les génomes entiers pour vérifier cette hypothèse.

V.D Les protéines contenant des répétitions structurales ont elle une fonction particulière ?

Nous avons testé si les enzymes sont sous représentées parmi les protéines contenant des répétitions pseudo-symétriques. En effet, un certain nombre d'homo-dimères peuvent réguler leur fonction par association ou dissociation de leurs monomères. Si les protéines avec deux copies de la répétition sont des analogues fonctionnels de ces homo-dimères mais ne peuvent pas associer ou dissocier les copies de la répétition, par exemple à cause de la taille du linker, cela pourrait modifier la régulation de leur fonction. Par contre, si les protéines avec deux copies de la répétition ont la possibilité de s'associer ou se dissocier, cela ne sera pas désavantageux.

Nous avons donc extrait les numéros EC de la base de données d'enzymes de PDBsum (Laskowski et al., 2005). Les protéines ayant un numéro EC ont été considérées comme des protéines enzymatiques, et les protéines n'ayant pas de numéro comme des protéines non enzymatiques. Des comparaisons ont été effectuées entre les protéines contenant des répétitions pseudo-symétriques et non symétriques à 180°, à la fois dans le jeu de données « autres protéines » et dans toutes les protéines contenant des répétitions. Les protéines contenant des répétitions ont également été comparées avec l'ensemble du jeu de données de départ pour savoir si les protéines contenant des répétitions étaient moins fréquemment des enzymes. Tous les tests de χ^2 effectués se sont révélés non significatifs. Il n'y a donc pas de différence observable dans la répartition des enzymes entre les protéines avec ou sans répétition, qu'elles soient pseudo-symétriques ou non.

Tableau 12 : Comparaison des protéines enzymatiques et non enzymatiques parmi les protéines contenant des répétitions.

Symétrie à 180° (« autres protéines »)	Enzymes	Non enzymes
Symétrique	10	25
Non symétrique	15	35
Symétrie à 180° (toutes les répétitions)	Enzymes	Non enzymes
Symétrique	11	50
Non symétrique	25	86
Astral50 vs répétitions	Enzymes	Non enzymes
Astral 50 sans répétitions	2687	5932
Astral 50 avec répétitions	25	63

V.E Un modèle pour les protéines avec une répétition structurale symétrique à 180°

Deux catégories principales se dégagent parmi les protéines étudiées. Une partie des protéines présente une répétition symétrique, ces protéines correspondent pour certaines à des homo-dimères et semblent moins fréquemment trouvées dans la même espèce que les protéines contenant des répétitions non symétriques. Il est donc possible que les protéines avec une répétition symétrique C2 puissent remplacer certains homo-dimères. D'un autre coté, il existe des répétitions non symétriques, dont certaines existent avec un nombre variable d'éléments répétés. Ces protéines sont plus longues que celles contenant une répétition symétrique, et peuvent présenter au moins trois copies de la même répétition. Les protéines avec une et deux copies peuvent co-exister chez la même espèce.

Les duplications internes seraient peu délétères si elles arrivent dans des protéines qui possèdent un nombre variable de répétitions, qui se replient indépendamment. D'un autre coté, les protéines qui formaient des homo-dimères avant la duplication, peuvent continuer à adopter ce repliement après duplication sans dommage majeur, sauf si le fait de se dimériser modifie la fonction du dimère. Si ces répétitions sont moins délétères, il serait logique de les observer plus fréquemment.

Est ce que les protéines avec une répétition symétrique C2 seraient plus avantageuses que les homo-dimères dans certains cas ? Dans l'introduction, nous avons vu que les homo-dimères présentent certains avantages : ils permettent de former une grosse protéine, plus stable et moins sensible aux dégradations, à partir de l'assemblage de monomères. Ces monomères sont plus petits, donc synthétisés avec moins d'erreurs, leur assemblage peut dans certains cas modifier l'activité ou réguler la fonction de la protéine.

Cependant les protéines avec une répétition pseudo-symétrique C2 pourraient présenter d'autres avantages. En effet, certains homo-oligomères sont instables à faible concentration, et les protéines contenant une répétition pseudo-symétrique devraient être plus stables. Les homo-oligomères peuvent également poser des problèmes d'agrégation s'ils sont mal repliés (Ding et al., 2002), mais ils subiraient une plus forte pression de sélection contre l'agrégation que les protéines composées d'un seul

monomère (Chen and Dokholyan, 2008). De plus, les protéines avec une répétition pseudo-symétrique seraient plus libres d'évoluer que les vrais dimères. En effet, une mutation dans le gène d'un monomère aura des conséquences sur tous les monomères de l'assemblage (Monod et al., 1965), alors qu'une mutation dans une protéine avec une répétition pseudo-symétrique aura des conséquences moins importantes, et les mutations ponctuelles pourraient aboutir à une rupture de la symétrie. Une protéine avec une répétition pseudo-symétrique C2 est plus libre d'évoluer et pourrait ne conserver qu'une symétrie partielle. D'autre parte, les répétitions partielles symétriques pourraient être un moyen de créer des régions de symétrie sans avoir besoin de créer un homo-dimère au préalable. Les protéines avec une répétition dont les copies sont symétrique peuvent être stabilisées par cette symétrie et pourraient avoir moins de chances d'évoluer vers un état avec plus de deux répétitions.

Il est difficile de savoir qui des dimères ou des protéines avec une répétition symétrique C2 présente le plus d'avantages. La création de protéines avec une répétition symétrique pourrait dans un premier temps ne pas être contre si elle n'est pas délétère pour la protéine, et pourrait dans certains cas être avantageuse pour l'organisme, par exemple pour des protéines présentes en faible concentration. Elle offre également à la protéine des possibilités d'évolution différentes de celles d'un oligomère composé de plusieurs sous unités identiques. Les exemples de protéines contenant une répétition dont les deux copies sont symétriques sont peu fréquentes comparées aux homo-dimères. Est-ce que c'est parce que la formation de ces répétitions est rare (il faut que la duplication ait lieu au bon endroit pour que ce ne soit pas délétère) ou parce que les protéines qui forment des homo-dimères sont plus avantageuses ? Il est difficile de répondre à cette question avec les données actuelles.

VI Conclusion et perspectives

L'objectif de ce travail de thèse était de mieux comprendre pourquoi et comment certaines répétitions, qui sont créées au niveau des gènes et répercutés sur la séquence d'acides aminés puis sur la structure des protéines, sont conservées par les organismes. Pour cela, nous avons identifié et étudié les répétitions dans les gènes, les séquences protéiques et les structures tridimensionnelles des protéines.

VI.A Résumé

Nous avons développé un programme, Swelfe, qui cherche les répétitions à ces trois niveaux. Ce programme utilise l'algorithme de programmation dynamique proposé par Huang et Miller (Huang and Miller, 1991; Huang et al., 1990) et a été adapté aux structures en les représentant de façon linéaire sous la forme de leurs angles α. Les scores et les tests de significativité des répétitions obtenues ont été adaptés pour chaque niveau. Swelfe a été comparé à DALI (Holm and Sander, 1993) pour valider la méthode au niveau des répétitions structurales. Nous avons créé une banque contenant les séquences d'ADN et d'acides aminés correspondant aux structures de la PDB afin de pouvoir comparer les répétitions obtenues aux différents niveaux. Ce programme est disponible pour la communauté à l'adresse http://bioserv.rpbs.jussieu.fr/swelfe et peut être téléchargé ou utilisé en ligne. Il a fait l'objet d'une publication (Abraham et al., 2008).

Nous avons ensuite cherché les répétitions dans un ensemble non redondant de séquences nucléiques, séquences protéiques et structures tridimensionnelles. Swelfe a permis de trouver un nombre important de répétitions : environ 10% des protéines contiennent des répétitions à au moins un niveau. Cependant, le recouvrement des répétitions aux trois niveaux est assez faible et beaucoup de répétitions ne sont trouvées qu'à un seul niveau. Un partie des raisons expliquant ces différences ont été discutées. Ces résultats devraient faire l'objet d'un prochain article.

L'étude des structures contenant des répétitions longues a permis de mettre en évidence qu'environ un tiers de ces protéines contiennent une répétition dont les deux copies sont symétriques à 180°, comme le sont deux monomères dans un homo-dimère.

L'analyse de ces protéines contenant une répétition symétrique montre que certaines pourraient effectivement remplacer des dimères. Cette étude fait l'objet d'un article actuellement en préparation et dont le manuscrit est en annexe.

Je vais maintenant présenterles questions que soulèvent ces résultats. Je développerai ensuite les limites de la méthode et les solutions qui pourraient y être apportées. Cette étude ouvre de nouvelles perspectives et je prévois de poursuivre certaines analyses au cours des prochains mois.

VI.B Discussion

Une des conclusions étonnantes de ce travail est que les répétitions qui sont trouvées aux trois niveaux sont assez différentes. Cela peut paraître étonnant dans la mesure où les structures sont issues des séquences d'acides aminés, qui sont elles-mêmes issues des séquences nucléiques. Les recherches de répétitions internes sont habituellement faites à un seul niveau à la fois. Nos résultats montrent que l'information n'est pas la même à tous les niveaux, et que la recherche de répétitions à un seul niveau ne permet pas une étude exhaustive. Les petites répétitions sont celles pour lesquelles les différences sont les plus grandes, et ne sont pas trouvées à tous les niveaux en même temps. Les séquences nucléiques, protéiques et les structures ne sont pas conservées de la même façon : à cause de la redondance du code génétique et des contraintes fonctionnelles, les séquences protéiques sont conservées plus longtemps que les séquences nucléiques, et les structures sont conservées plus longtemps que les séquences protéiques. De plus, les scores et les seuils statistiques sont adaptés à chaque niveau, mais ne permettent pas toujours de trouver les mêmes répétitions. Certaines répétitions peuvent être significatives à un niveau mais pas à un autre. Par exemple, la composition de la séquence n'est prise en compte qu'au niveau des séquences nucléiques. La matrice BLOSUM62 tient compte des substitutions d'acides aminés les plus fréquentes dans les protéines. Le score au niveau des structures pénalise les angles α les plus fréquents, qui correspondent à des structures secondaires qui peuvent être plus conservées que les boucles. Les poids de gaps sont très défavorisés pour les répétitions structurales. Une étude a montré que les gaps sont souvent plus présents dans les boucles des structures (Pascarella and Argos, 1992). Les angles / acides aminés /

nucléotides correspondant aux hélices α et feuillets β devraient avoir un poids de gap plus élevé à tous les niveaux car ils sont défavorisés dans ces structures secondaires. L'information apportée par les séquences nucléiques et protéiques et les structures est donc assez différente et il paraît important d'utiliser toute cette information pour mieux comprendre l'évolution des répétitions.

Une autre analyse qui soulève encore beaucoup de questions est celle des répétitions structurales longues. Environ le tiers des copies de ces répétitions est symétrique à 180°. Le fait que les deux copies de la répétition soient symétriques pourrait augmenter leur stabilité. Ces protéines pourraient également offrir des perspectives évolutives différentes de celles des homo-dimères. Ces protéines sont peu nombreuses comparées aux homo-dimères. Est-ce dû au fait que ces répétitions sont rares, la probabilité que la duplication ait lieu à un endroit non délétère paraît faible, ou au fait qu'elles sont moins avantageuses que les homo-dimères ? Est ce que leur fonction pourrait être différente ? Ces protéines évoluent-elles différemment des homo-dimères ? Il n'est pas possible de répondre pour le moment à ces questions.

VI.C Améliorations, perspectives

Ce travail ouvre plusieurs perspectives. Comme indiqué précédemment, les gaps seraient moins fréquents dans les hélices α et feuillets β. Swelfe utilise un poids de gap constant quel que soient les structures secondaires. Nous avons donc gardé un poids de gap élevé pour les structures. Il pourrait être intéressant d'accorder un poids de gap moins important pour les angles α ne correspondant ni aux hélices α ni aux feuillets β afin de favoriser les gaps dans les boucles. De plus, les boucles sont les parties les plus flexibles de la protéine, et cette flexibilité n'est pas du tout prise en compte par Swelfe. Diminuer le poids de gap à ces endroits permettrait peut-être de donner un peu de souplesse à l'alignement au niveau des boucles.

Il serait également intéressant de rassembler les informations contenues dans les séquences et les structures afin de pouvoir trouver les répétitions. En effet les répétitions les plus anciennes peuvent ne plus être détectées en séquence, et certaines structures secondaires très fréquentes ne sont pas forcément issues de duplication. Prendre en compte l'ensemble des informations permettrait de trouver les répétitions de façon plus

exhaustive et plus fiable. Il serait aussi possible d'appliquer des poids de gaps différents pour les séquences codant pour des hélices α et feuillets β, ou des boucles.

Un biais de nos analyses est bien sûr dû à l'utilisation de données de la PDB. En effet, plusieurs études ont montré que la composition de la PDB est différente de la composition des génomes et des banques de données de séquences protéiques. Les protéines de la PDB sont significativement moins longues que celles trouvées dans huit génomes microbiens, et les deux jeux de données présentent des compositions en acides aminés différentes (Gerstein, 1998). De plus, il y a un biais par rapport aux fonctions des protéines : les protéines membranaires, de signalisation, désordonnées ou contenant des régions de faible complexité sont significativement sous-représentées dans la PDB, en comparaison aux protéines de SWISS-PROT. Inversement, les ponts disulfure, les sites de liaison au métal et les sites impliqués dans une activité enzymatique sont surreprésentés (Peng et al., 2004). Tous ces biais sont dus à plusieurs raisons notamment la facilité à obtenir les cristaux de certaines protéines (les protéines membranaires sont difficiles à résoudre car elles nécessitent une bicouche lipidique ou un substitut amphiphile) et les objectifs des groupes de recherche qui déterminent les structures (certains s'intéressent à un organisme modèle, à une voie métabolique, à des protéines liées à une maladie particulière). La PDB contient donc peu de protéines membranaires, qui sont fréquentes en séquence et connues pour contenir des répétitions (Mande and Rao, 2006). Elle contient également peu de protéines contenant des répétitions internes annotées « REPEAT » par SWISS-PROT, qui peuvent correspondre à des structures de faible complexité (Peng et al., 2004). Les résultats que nous avons obtenus sous estiment très probablement le nombre de répétitions dans les séquences car ils sont basés sur les données de la PDB.

Pour contrer ce biais, il serait intéressant de faire une étude des répétitions dans les génomes. Nous n'avons pas encore pu le faire par manque de temps mais nous projetons de rechercher prochainement ces répétitions dans espèces proches d'entérobactéries. Ce travail permettrait de calculer la fréquence des répétitions intragéniques dans les génomes, leur localisation, comprendre s'il s'agit de gènes essentiels ou non, transférés ou non, s'il s'agit d'événements de duplications récents… Il serait également possible de faire une étude en prenant en compte des banques de

domaines protéiques et les banques structurales pour quantifier les répétitions de domaines.

Concernant les répétitions pseudo-symétriques, il serait intéressant de faire une étude dans plusieurs organismes pour savoir si les protéines contenant une répétition pseudo-symétrique C2 existent chez la même espèce que les homo-dimères dont chaque monomère contient une copie de la répétition. Cependant, il faudrait dans ce cas faire une étude sur des génomes entiers, ce qui est long, et il n'y a peut être pas assez de génomes proches eucaryotes séquencés pour étudier certaines familles de protéines (les eucaryotes sont les organismes qui contiennent le plus de répétitions). Il pourrait aussi être intéressant d'étudier comment les structures s'adaptent aux répétitions dans les cas où ces répétitions ne sont pas symétriques à 180°. Les mutations qui arrivent dans les protéines avec une répétition dont les copies sont symétriques peuvent entraîner une rupture de symétrie et deux copies peuvent évoluer indépendamment. Il serait intéressant d'étudier comment la protéine peut perdre une partie de sa symétrie. Une autre analyse que nous n'avons pu faire est de chercher si les répétitions entraînent un changement de fonction de la protéine. Cette analyse n'est pas très simple car elle suppose de connaître la fonction de la protéine avant et après duplication, et les fonctions sont souvent assignées par comparaison avec la protéine la plus proche. Il est donc possible que des fonctions identiques soient assignées à ces deux protéines.

Les amplifications sont majoritairement délétères dans les protéines, car elles risquent de changer la séquence d'acides aminés, et le repliement, et peuvent modifier la fonction. Certaines répétitions sont conservées dans les organismes, et elles sont probablement adaptatives. Les répétitions anciennes peuvent être difficiles à trouver si les séquences ont trop divergé, et certaines similarités structurales sont dues à des structures secondaires très fréquentes. Leur étude aux trois niveaux permet de pallier à ces problèmes. Parmi les longues répétitions, qui sont à priori les plus délétères, nous avons trouvé environ 30% de répétitions dont les copies sont symétriques et qui pourraient remplacer certains homo-dimères. Il existe également certaines protéines qui ont un nombre variable de répétitions longues. Certains types d'amplifications semblent donc viables pour les protéines. Ainsi, elles constituent dans certains cas une opportunité pour la création de nouvelles structures et fonctions.

Références

1 Abraham, A.L., Rocha, E.P. and Pothier, J. (2008) Swelfe: a detector of internal repeats in sequences and structures. *Bioinformatics*, **24**, 1536-1537.
2 Achaz, G. (2002) Etude de la dynamique des génomes: les répétitions intrachromosomiques. *Génétique*. Université Paris 6, Paris.
3 Achaz, G., Boyer, F., Rocha, E.P., Viari, A. and Coissac, E. (2007) Repseek, a tool to retrieve approximate repeats from large DNA sequences. *Bioinformatics*, **23**, 119-121.
4 Achaz, G., Coissac, E., Viari, A. and Netter, P. (2000) Analysis of intrachromosomal duplications in yeast Saccharomyces cerevisiae: a possible model for their origin. *Mol Biol Evol*, **17**, 1268-1275.
5 Alberts, B., Bray, D., Lewis, J., Raff, M., Toberts, K. and Watson, J. (1994) *Molecular Biology of the Cell*. Arland Publishing inc., New-York, London.
6 Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J Mol Biol*, **215**, 403-410.
7 Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, **25**, 3389-3402.
8 Anderson, P. and Roth, J. (1981) Spontaneous tandem genetic duplications in Salmonella typhimurium arise by unequal recombination between rRNA (rrn) cistrons. *Proc Natl Acad Sci U S A*, **78**, 3113-3117.
9 Anderson, R.P. and Roth, J.R. (1977) Tandem genetic duplications in phage and bacteria. *Annu Rev Microbiol*, **31**, 473-505.
10 Andrade, M.A., Perez-Iratxeta, C. and Ponting, C.P. (2001) Protein repeats: structures, functions, and evolution. *J Struct Biol*, **134**, 117-131.
11 Andre, I., Strauss, C.E., Kaplan, D.B., Bradley, P. and Baker, D. (2008) Emergence of symmetry in homooligomeric biological assemblies. *Proc Natl Acad Sci U S A*, **105**, 16148-16152.
12 Anfinsen, C.B. (1973) Principles that govern the folding of protein chains. *Science*, **181**, 223-230.
13 Apic, G., Gough, J. and Teichmann, S.A. (2001a) Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J Mol Biol*, **310**, 311-325.
14 Apic, G., Gough, J. and Teichmann, S.A. (2001b) An insight into domain combinations. *Bioinformatics*, **17 Suppl 1**, S83-89.
15 Apic, G., Huber, W. and Teichmann, S.A. (2003) Multi-domain protein families and domain pairs: comparison with known structures and a random model of domain recombination. *J Struct Funct Genomics*, **4**, 67-78.
16 Aravind, L., Tatusov, R.L., Wolf, Y.I., Walker, D.R. and Koonin, E.V. (1998) Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles. *Trends Genet*, **14**, 442-444.
17 Arnaiz, O., Cain, S., Cohen, J. and Sperling, L. (2007) ParameciumDB: a community resource that integrates the Paramecium tetraurelia genome sequence with genetic data. *Nucleic Acids Res*, **35**, D439-444.
18 Aroul-Selvam, R., Hubbard, T. and Sasidharan, R. (2004) Domain insertions in protein structures. *J Mol Biol*, **338**, 633-641.
19 Aury, J.M., Jaillon, O., Duret, L., Noel, B., Jubin, C., Porcel, B.M., Segurens, B., Daubin, V., Anthouard, V., Aiach, N., Arnaiz, O., Billaut, A., Beisson, J., Blanc,

I., Bouhouche, K., Camara, F., Duharcourt, S., Guigo, R., Gogendeau, D., Katinka, M., Keller, A.M., Kissmehl, R., Klotz, C., Koll, F., Le Mouel, A., Lepere, G., Malinsky, S., Nowacki, M., Nowak, J.K., Plattner, H., Poulain, J., Ruiz, F., Serrano, V., Zagulski, M., Dessen, P., Betermier, M., Weissenbach, J., Scarpelli, C., Schachter, V., Sperling, L., Meyer, E., Cohen, J. and Wincker, P. (2006) Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. *Nature*, **444**, 171-178.

20 Avise, J.C. and Kitto, G.B. (1973) Phosphoglucose isomerase gene duplication in the bony fishes: an evolutionary history. *Biochem Genet*, **8**, 113-132.

21 Axe, D.D., Foster, N.W. and Fersht, A.R. (1996) Active barnase variants with completely random hydrophobic cores. *Proc Natl Acad Sci U S A*, **93**, 5590-5594.

22 Bahadur, R.P., Chakrabarti, P., Rodier, F. and Janin, J. (2003) Dissecting subunit interfaces in homodimeric proteins. *Proteins*, **53**, 708-719.

23 Bashton, M. and Chothia, C. (2002) The geometry of domain combination in proteins. *J Mol Biol*, **315**, 927-939.

24 Benson, G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res*, **27**, 573-580.

25 Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) The Protein Data Bank. *Nucleic Acids Res*, **28**, 235-242.

26 Betancourt, M.R. and Skolnick, J. (2001) Universal similarity measure for comparing protein structures. *Biopolymers*, **59**, 305-309.

27 Bjorklund, A.K., Ekman, D. and Elofsson, A. (2006) Expansion of protein domain repeats. *PLoS Comput Biol*, **2**, e114.

28 Bjorklund, A.K., Ekman, D., Light, S., Frey-Skott, J. and Elofsson, A. (2005) Domain rearrangements in protein evolution. *J Mol Biol*, **353**, 911-923.

29 Blundell, T.L. and Srinivasan, N. (1996) Symmetry, stability, and dynamics of multidomain and multicomponent protein systems. *Proc Natl Acad Sci U S A*, **93**, 14243-14248.

30 Blundell, T.L. and Wood, S.P. (1975) Is the evolution of insulin Darwinian or due to selectively neutral mutation? *Nature*, **257**, 197-203.

31 Bornberg-Bauer, E., Beaussart, F., Kummerfeld, S.K., Teichmann, S.A. and Weiner, J., 3rd. (2005) The evolution of domain arrangements in proteins and interaction networks. *Cell Mol Life Sci*, **62**, 435-445.

32 Bourque, G., Zdobnov, E.M., Bork, P., Pevzner, P.A. and Tesler, G. (2005) Comparative architectures of mammalian and chicken genomes reveal highly variable rates of genomic rearrangements across different lineages. *Genome Res*, **15**, 98-110.

33 Branden, C. and Tooze, J. (1999) *Introduction to Protein Structure*. Garland Publishing, New York.

34 Bridges, C.B. (1935) Salivary chromosome maps with a key to the banding of the chromosomes of Drosophila melanogaster. *Journal of Heredity*, **26**, 60-64.

35 Britten, R.J. and Davidson, E.H. (1969) Gene Regulation for Higher Cells - a Theory. *Science*, **165**, 349-&.

36 Brunet, F.G., Crollius, H.R., Paris, M., Aury, J.M., Gibert, P., Jaillon, O., Laudet, V. and Robinson-Rechavi, M. (2006) Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Mol Biol Evol*, **23**, 1808-1816.

37 Carpentier, M. (2005) Méthodes de détection des similarités structurales : caractérisation des motifs conservées dans les familles de structures pour l'annotation des génomes. *Sciences de la Vie*. Université Paris 6, Paris.
38 Carpentier, M., Brouillet, S. and Pothier, J. (2005) YAKUSA: a fast structural database scanning method. *Proteins*, **61**, 137-151.
39 Chandonia, J.M., Hon, G., Walker, N.S., Lo Conte, L., Koehl, P., Levitt, M. and Brenner, S.E. (2004) The ASTRAL Compendium in 2004. *Nucleic Acids Res*, **32**, D189-192.
40 Chen, Y. and Dokholyan, N.V. (2008) Natural selection against protein aggregation on self-interacting and essential proteins in yeast, fly, and worm. *Mol Biol Evol*, **25**, 1530-1533.
41 Chiti, F., Taddei, N., Bucciantini, M., White, P., Ramponi, G. and Dobson, C.M. (2000) Mutational analysis of the propensity for amyloid formation by a globular protein. *Embo J*, **19**, 1441-1449.
42 Chothia, C. (1991) *Biological asymmetry and handedness*. Chichester New York , Wiley -- 1991.
43 Chothia, C. and Gerstein, M. (1997) Protein evolution. How far can sequences diverge? *Nature*, **385**, 579, 581.
44 Chothia, C., Gough, J., Vogel, C. and Teichmann, S.A. (2003) Evolution of the protein repertoire. *Science*, **300**, 1701-1703.
45 Chothia, C. and Lesk, A.M. (1986) The relation between the divergence of sequence and structure in proteins. *Embo J*, **5**, 823-826.
46 Consortium, T.U. (2007) The Universal Protein Resource (UniProt). *Nucleic Acids Res*, **35**, D193-197.
47 Cornish-Bowden, A.J. and Koshland, D.E., Jr. (1971) The quaternary structure of proteins composed of identical subunits. *J Biol Chem*, **246**, 3092-3102.
48 Corpet, F. (1988) Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res*, **16**, 10881-10890.
49 Critchlow, S.E. and Jackson, S.P. (1998) DNA end-joining: from yeast to man. *Trends Biochem Sci*, **23**, 394-398.
50 Curcio, M.J. and Derbyshire, K.M. (2003) The outs and ins of transposition: from mu to kangaroo. *Nat Rev Mol Cell Biol*, **4**, 865-877.
51 Dandekar, T., Schuster, S., Snel, B., Huynen, M. and Bork, P. (1999) Pathway alignment: application to the comparative analysis of glycolytic enzymes. *Biochem J*, **343 Pt 1**, 115-124.
52 Davis, J.C. and Petrov, D.A. (2004) Preferential duplication of conserved proteins in eukaryotic genomes. *PLoS Biol*, **2**, E55.
53 Dayhoff, M., Schwartz, R.M. and Orcutt, B.C. (1978) *A model of evolutionary change in proteins*, Washington.
54 de Souza, S.J., Long, M., Schoenbach, L., Roy, S.W. and Gilbert, W. (1996) Intron positions correlate with module boundaries in ancient proteins. *Proc Natl Acad Sci U S A*, **93**, 14632-14636.
55 de Souza, S.J., Long, M., Schoenbach, L., Roy, S.W. and Gilbert, W. (1997) The correlation between introns and the three-dimensional structure of proteins. *Gene*, **205**, 141-144.
56 Delcher, A.L., Salzberg, S.L. and Phillippy, A.M. (2003) Using MUMmer to identify similar regions in large sequence sets. *Curr Protoc Bioinformatics*, **Chapter 10**, Unit 10 13.

57 DePristo, M.A., Weinreich, D.M. and Hartl, D.L. (2005) Missense meanderings in sequence space: a biophysical view of protein evolution. *Nat Rev Genet*, **6**, 678-687.
58 Dermitzakis, E.T. and Clark, A.G. (2001) Differential selection after duplication in mammalian developmental genes. *Mol Biol Evol*, **18**, 557-562.
59 Dianov, G.L., Kuzminov, A.V., Mazin, A.V. and Salganik, R.I. (1991) Molecular mechanisms of deletion formation in Escherichia coli plasmids. I. Deletion formation mediated by long direct repeats. *Mol Gen Genet*, **228**, 153-159.
60 Ding, F., Dokholyan, N.V., Buldyrev, S.V., Stanley, H.E. and Shakhnovich, E.I. (2002) Molecular dynamics simulation of the SH3 domain aggregation suggests a generic amyloidogenesis mechanism. *J Mol Biol*, **324**, 851-857.
61 Dobson, C.M. (2003) Protein folding and misfolding. *Nature*, **426**, 884-890.
62 Doolittle, R.F. (1992) Stein and Moore Award address. Reconstructing history with amino acid sequences. *Protein Sci*, **1**, 191-200.
63 Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.
64 Ekman, D., Bjorklund, A.K., Frey-Skott, J. and Elofsson, A. (2005) Multi-domain proteins in the three kingdoms of life: orphan domains and other unassigned regions. *J Mol Biol*, **348**, 231-243.
65 Enright, A.J. and Ouzounis, C.A. (2001) Functional associations of proteins in entire genomes by means of exhaustive detection of gene fusions. *Genome Biol*, **2**, RESEARCH0034.
66 Ferreiro, D.U. and Komives, E.A. (2007) The plastic landscape of repeat proteins. *Proc Natl Acad Sci U S A*, **104**, 7735-7736.
67 Ferris, S.D. and Whitt, G.S. (1979) Evolution of the differential regulation of duplicate genes after polyploidization. *J Mol Evol*, **12**, 267-317.
68 Finn, R.D., Tate, J., Mistry, J., Coggill, P.C., Sammut, S.J., Hotz, H.R., Ceric, G., Forslund, K., Eddy, S.R., Sonnhammer, E.L. and Bateman, A. (2008) The Pfam protein families database. *Nucleic Acids Res*, **36**, D281-288.
69 Fischer, D., Elofsson, A., Rice, D. and Eisenberg, D. (1996) Assessing the performance of fold recognition methods by means of a comprehensive benchmark. *Pac Symp Biocomput*, 300-318.
70 Fisher, R.A. (1930) *The Genetical Theory of Natural Selection*. Clarendon Press.
71 Flores, M., Brom, S., Stepkowski, T., Girard, M.L., Davila, G., Romero, D. and Palacios, R. (1993) Gene amplification in Rhizobium: identification and in vivo cloning of discrete amplifiable DNA regions (amplicons) from Rhizobium leguminosarum biovar phaseoli. *Proc Natl Acad Sci U S A*, **90**, 4932-4936.
72 Fondon, J.W., 3rd and Garner, H.R. (2004) Molecular origins of rapid and continuous morphological evolution. *Proc Natl Acad Sci U S A*, **101**, 18058-18063.
73 Fong, J.H., Geer, L.Y., Panchenko, A.R. and Bryant, S.H. (2007) Modeling the evolution of protein domain architectures using maximum parsimony. *J Mol Biol*, **366**, 307-315.
74 Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.L. and Postlethwait, J. (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, **151**, 1531-1545.
75 Forman, J.R., Qamar, S., Paci, E., Sandford, R.N. and Clarke, J. (2005) The remarkable mechanical strength of polycystin-1 supports a direct role in mechanotransduction. *J Mol Biol*, **349**, 861-871.

76 Fornasari, M.S., Parisi, G. and Echave, J. (2007) Quaternary structure constraints on evolutionary sequence divergence. *Mol Biol Evol*, **24**, 349-351.
77 Francino, M.P. (2005) An adaptive radiation model for the origin of new gene functions. *Nat Genet*, **37**, 573-577.
78 Gassner, N.C., Baase, W.A. and Matthews, B.W. (1996) A test of the "jigsaw puzzle" model for protein folding by multiple methionine substitutions within the core of T4 lysozyme. *Proc Natl Acad Sci U S A*, **93**, 12155-12158.
79 Gerstein, M. (1997) A structural census of genomes: comparing bacterial, eukaryotic, and archaeal genomes in terms of protein structure. *J Mol Biol*, **274**, 562-576.
80 Gerstein, M. (1998) How representative are the known structures of the proteins in a complete genome? A comprehensive structural census. *Fold Des*, **3**, 497-512.
81 Gerstein, M. and Levitt, M. (1998) Comprehensive assessment of automatic structural alignment against a manual standard, the scop classification of proteins. *Protein Sci*, **7**, 445-456.
82 Gibrat, J.F., Madej, T. and Bryant, S.H. (1996) Surprising similarities in structure comparison. *Curr Opin Struct Biol*, **6**, 377-385.
83 Goodsell, D.S. (1991) Inside a living cell. *Trends Biochem Sci*, **16**, 203-206.
84 Goodsell, D.S. and Olson, A.J. (1993) Soluble proteins: size, shape and function. *Trends Biochem Sci*, **18**, 65-68.
85 Goodsell, D.S. and Olson, A.J. (2000) Structural symmetry and protein function. *Annu Rev Biophys Biomol Struct*, **29**, 105-153.
86 Gouzy, J., Corpet, F. and Kahn, D. (1999) Whole genome protein domain analysis using a new method for domain clustering. *Comput Chem*, **23**, 333-340.
87 Grant, A., Lee, D. and Orengo, C. (2004) Progress towards mapping the universe of protein folds. *Genome Biol*, **5**, 107.
88 Grimes, J.M., Burroughs, J.N., Gouet, P., Diprose, J.M., Malby, R., Zientara, S., Mertens, P.P. and Stuart, D.I. (1998) The atomic structure of the bluetongue virus core. *Nature*, **395**, 470-478.
89 Gulick, A. (1944) The chemical formulation of gene structure and gene action. *Advances in Enzymology and Related Subjects of Biochemistry*, **4**, 1-39.
90 Hakes, L., Pinney, J.W., Lovell, S.C., Oliver, S.G. and Robertson, D.L. (2007) All duplicates are not equal: the difference between small-scale and genome duplication. *Genome Biol*, **8**, R209.
91 Haldane, J.B.S. (1932) *The Causes of Evolution*. Green and Co.
92 Han, J.H., Batey, S., Nickson, A.A., Teichmann, S.A. and Clarke, J. (2007) The folding and evolution of multidomain proteins. *Nat Rev Mol Cell Biol*, **8**, 319-330.
93 Han, J.H., Kerrison, N., Chothia, C. and Teichmann, S.A. (2006) Divergence of interdomain geometry in two-domain proteins. *Structure*, **14**, 935-945.
94 Hegyi, H. and Gerstein, M. (2001) Annotation transfer for genomics: measuring functional divergence in multi-domain proteins. *Genome Res*, **11**, 1632-1640.
95 Hendrickson, H., Slechta, E.S., Bergthorsson, U., Andersson, D.I. and Roth, J.R. (2002) Amplification-mutagenesis: evidence that "directed" adaptive mutation and general hypermutability result from growth with a selected gene amplification. *Proc Natl Acad Sci U S A*, **99**, 2164-2169.
96 Henikoff, S. and Henikoff, J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, **89**, 10915-10919.
97 Heringa, J. (1998) Detection of internal repeats: how common are they? *Curr Opin Struct Biol*, **8**, 338-345.

98 Hirschberg, D.S. (1975) Linear Space Algorithm for Computing Maximal Common Subsequences. *Communications of the Acm*, **18**, 341-343.
99 Holm, L. and Sander, C. (1993) Protein structure comparison by alignment of distance matrices. *J Mol Biol*, **233**, 123-138.
100 Huang, X. and Miller, W. (1991) A time-Efficient, Linear-Spaced Local Similarity Algorithm. *Adv In Appl Math*, **12**, 337-357.
101 Huang, X.Q., Hardison, R.C. and Miller, W. (1990) A space-efficient algorithm for local similarities. *Comput Appl Biosci*, **6**, 373-381.
102 Iturbe-Ormaetxe, I., Burke, G.R., Riegler, M. and O'Neill, S.L. (2005) Distribution, expression, and motif variability of ankyrin domain genes in Wolbachia pipientis. *J Bacteriol*, **187**, 5136-5145.
103 Janin, J. and Chothia, C. (1990) The structure of protein-protein recognition sites. *J Biol Chem*, **265**, 16027-16030.
104 Jeltsch, A. (1999) Circular permutations in the molecular evolution of DNA methyltransferases. *J Mol Evol*, **49**, 161-164.
105 Jeong, H., Mason, S.P., Barabasi, A.L. and Oltvai, Z.N. (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41-42.
106 Jordan, I.K., Makarova, K.S., Spouge, J.L., Wolf, Y.I. and Koonin, E.V. (2001) Lineage-specific gene expansions in bacterial and archaeal genomes. *Genome Res*, **11**, 555-565.
107 Jordan, I.K., Wolf, Y.I. and Koonin, E.V. (2004) Duplicated genes evolve slower than singletons despite the initial rate increase. *BMC Evol Biol*, **4**, 22.
108 Jung, J. and Lee, B. (2001) Circularly permuted proteins in the protein structure database
10.1110/ps.05801. *Protein Sci*, **10**, 1881-1886.
109 Kajava, A.V. (2001) Review: proteins with repeated sequence--structural prediction and modeling. *J Struct Biol*, **134**, 132-144.
110 Kawabata, T. and Nishikawa, K. (2000) Protein structure comparison using the markov transition model of evolution. *Proteins*, **41**, 108-122.
111 Kelman, Z., Finkelstein, J. and O'Donnell, M. (1995) Protein structure. Why have six-fold symmetry? *Curr Biol*, **5**, 1239-1242.
112 Kimura, M. and Ota, T. (1974) On some principles governing molecular evolution. *Proc Natl Acad Sci U S A*, **71**, 2848-2852.
113 Klotz, I.M., Langerman, N.R. and Darnall, D.W. (1970) Quaternary structure of proteins. *Annu Rev Biochem*, **39**, 25-62.
114 Kobe, B. and Deisenhofer, J. (1994) The leucine-rich repeat: a versatile binding motif. *Trends Biochem Sci*, **19**, 415-421.
115 Kohlstaedt, L.A., Wang, J., Friedman, J.M., Rice, P.A. and Steitz, T.A. (1992) Crystal structure at 3.5 A resolution of HIV-1 reverse transcriptase complexed with an inhibitor. *Science*, **256**, 1783-1790.
116 Kondrashov, F.A., Rogozin, I.B., Wolf, Y.I. and Koonin, E.V. (2002) Selection in the evolution of gene duplications. *Genome Biol*, **3**, RESEARCH0008.
117 Kong, X.P., Onrust, R., O'Donnell, M. and Kuriyan, J. (1992) Three-dimensional structure of the beta subunit of E. coli DNA polymerase III holoenzyme: a sliding DNA clamp. *Cell*, **69**, 425-437.
118 Koonin, E.V., Mushegian, A.R. and Bork, P. (1996) Non-orthologous gene displacement. *Trends Genet*, **12**, 334-336.
119 Koonin, E.V., Wolf, Y.I. and Karev, G.P. (2002) The structure of the protein universe and genome evolution. *Nature*, **420**, 218-223.

120 Koshland, D.E., Jr. (1976) The evolution of function in enzymes. *Fed Proc*, **35**, 2104-2111.
121 Kraus, E., Leung, W.Y. and Haber, J.E. (2001) Break-induced replication: a review and an example in budding yeast. *Proc Natl Acad Sci U S A*, **98**, 8255-8262.
122 Krishna, T.S., Kong, X.P., Gary, S., Burgers, P.M. and Kuriyan, J. (1994) Crystal structure of the eukaryotic DNA polymerase processivity factor PCNA. *Cell*, **79**, 1233-1243.
123 Kulikova, T., Akhtar, R., Aldebert, P., Althorpe, N., Andersson, M., Baldwin, A., Bates, K., Bhattacharyya, S., Bower, L., Browne, P., Castro, M., Cochrane, G., Duggan, K., Eberhardt, R., Faruque, N., Hoad, G., Kanz, C., Lee, C., Leinonen, R., Lin, Q., Lombard, V., Lopez, R., Lorenc, D., McWilliam, H., Mukherjee, G., Nardone, F., Pastor, M.P., Plaister, S., Sobhany, S., Stoehr, P., Vaughan, R., Wu, D., Zhu, W. and Apweiler, R. (2007) EMBL Nucleotide Sequence Database in 2006. *Nucleic Acids Res*, **35**, D16-20.
124 Kummerfeld, S.K. and Teichmann, S.A. (2005) Relative rates of gene fusion and fission in multi-domain proteins. *Trends Genet*, **21**, 25-30.
125 Kurtz, S. and Schleiermacher, C. (1999) REPuter: fast computation of maximal repeats in complete genomes. *Bioinformatics*, **15**, 426-427.
126 Kuwada, Y. (1911) Meiosis in the pollen mother cells of Zea Mays. *L. Bot. Mag.*, **25**, 163.
127 Lackner, P., Koppensteiner, W.A., Sippl, M.J. and Domingues, F.S. (2000) ProSup: a refined tool for protein structure alignment. *Protein Eng*, **13**, 745-752.
128 Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J. and Higgins, D.G. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947-2948.
129 Laskowski, R.A., Chistyakov, V.V. and Thornton, J.M. (2005) PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids. *Nucleic Acids Res*, **33**, D266-268.
130 Lavorgna, G., Patthy, L. and Boncinelli, E. (2001) Were protein internal repeats formed by "bricolage"? *Trends Genet*, **17**, 120-123.
131 Lee, D., Grant, A., Marsden, R.L. and Orengo, C. (2005) Identification and distribution of protein families in 120 completed genomes using Gene3D. *Proteins*, **59**, 603-615.
132 Lespinet, O., Wolf, Y.I., Koonin, E.V. and Aravind, L. (2002) The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Res*, **12**, 1048-1059.
133 Levine, M. and Tjian, R. (2003) Transcription regulation and animal diversity. *Nature*, **424**, 147-151.
134 Levitt, M. (1976) A simplified representation of protein conformations for rapid simulation of protein folding. *J Mol Biol*, **104**, 59-107.
135 Levitt, M. and Warshel, A. (1975) Computer simulation of protein folding. *Nature*, **253**, 694-698.
136 Li, W. and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658-1659.
137 Lindqvist, Y. and Schneider, G. (1997) Circular permutations of natural protein sequences: structural evidence. *Curr Opin Struct Biol*, **7**, 422-427.
138 Liu, J. and Rost, B. (2004) Sequence-based prediction of protein domains. *Nucleic Acids Res*, **32**, 3522-3530.

139 Liu, M. and Grigoriev, A. (2004) Protein domains correlate strongly with exons in multiple eukaryotic genomes--evidence of exon shuffling? *Trends Genet*, **20**, 399-403.
140 Lux, S.E., John, K.M. and Bennett, V. (1990) Analysis of cDNA for human erythrocyte ankyrin indicates a repeated structure with homology to tissue-differentiation and cell-cycle control proteins. *Nature*, **344**, 36-42.
141 Lynch, M. and Conery, J.S. (2000) The evolutionary fate and consequences of duplicate genes. *Science*, **290**, 1151-1155.
142 Lynch, M. and Force, A. (2000) The probability of duplicate gene preservation by subfunctionalization. *Genetics*, **154**, 459-473.
143 Madera, M., Vogel, C., Kummerfeld, S.K., Chothia, C. and Gough, J. (2004) The SUPERFAMILY database in 2004: additions and improvements. *Nucleic Acids Res*, **32**, D235-239.
144 Makarova, K.S., Aravind, L., Galperin, M.Y., Grishin, N.V., Tatusov, R.L., Wolf, Y.I. and Koonin, E.V. (1999) Comparative genomics of the Archaea (Euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell. *Genome Res*, **9**, 608-628.
145 Mande, S.S. and Rao, V.V.R. (2006) Comparative analysis of tandem repeats in the 44 outer membrane proteins of non-pathogenic (K12) and two pathogenic (O157) strains of E-coli. *Current Science*, **90**, 88-94.
146 Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice, D.W., Yeates, T.O. and Eisenberg, D. (1999a) Detecting protein function and protein-protein interactions from genome sequences. *Science*, **285**, 751-753.
147 Marcotte, E.M., Pellegrini, M., Yeates, T.O. and Eisenberg, D. (1999b) A census of protein repeats. *J Mol Biol*, **293**, 151-160.
148 Martin, J. (2005) Prédiction de la structure locale des protéines par des modèles de chaînes de Markov cachées. Université Paris 7, Paris.
149 Matthews, B.W. (1995) Studies on protein stability with T4 lysozyme. *Adv Protein Chem*, **46**, 249-278.
150 Modolell, J. and Campuzano, S. (1998) The achaete-scute complex as an integrating device. *Int J Dev Biol*, **42**, 275-282.
151 Monod, J., Wyman, J. and Changeux, J.P. (1965) On the Nature of Allosteric Transitions: A Plausible Model. *J Mol Biol*, **12**, 88-118.
152 Moxon, E.R., Rainey, P.B., Nowak, M.A. and Lenski, R.E. (1994) Adaptive evolution of highly mutable loci in pathogenic bacteria. *Curr Biol*, **4**, 24-33.
153 Muller, A., MacCallum, R.M. and Sternberg, M.J. (2002) Structural characterization of the human proteome. *Genome Res*, **12**, 1625-1641.
154 Muller, H.J. and Gershenson, S.M. (1935) Inert Regions of Chromosomes as the Temporary Products of Individual Genes. *Proc Natl Acad Sci U S A*, **21**, 69-75.
155 Murzin, A.G., Brenner, S.E., Hubbard, T. and Chothia, C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, **247**, 536-540.
156 Murzin, A.G., Lesk, A.M. and Chothia, C. (1992) beta-Trefoil fold. Patterns of structure and sequence in the Kunitz inhibitors interleukins-1 beta and 1 alpha and fibroblast growth factors. *J Mol Biol*, **223**, 531-543.
157 Myers, E.W. and Miller, W. (1988) Optimal alignments in linear space. *Comput Appl Biosci*, **4**, 11-17.
158 Needleman, S.B. and Wunsch, C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, **48**, 443-453.

159 Neer, E.J., Schmidt, C.J., Nambudripad, R. and Smith, T.F. (1994) The ancient regulatory-protein family of WD-repeat proteins. *Nature*, **371**, 297-300.
160 Nelson, K.E., Clayton, R.A., Gill, S.R., Gwinn, M.L., Dodson, R.J., Haft, D.H., Hickey, E.K., Peterson, J.D., Nelson, W.C., Ketchum, K.A., McDonald, L., Utterback, T.R., Malek, J.A., Linher, K.D., Garrett, M.M., Stewart, A.M., Cotton, M.D., Pratt, M.S., Phillips, C.A., Richardson, D., Heidelberg, J., Sutton, G.G., Fleischmann, R.D., Eisen, J.A., White, O., Salzberg, S.L., Smith, H.O., Venter, J.C. and Fraser, C.M. (1999) Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritima. *Nature*, **399**, 323-329.
161 Novotny, M., Madsen, D. and Kleywegt, G.J. (2004) Evaluation of protein fold comparison servers. *Proteins*, **54**, 260-270.
162 Ohno, S. (1970) *Evolution by Gene Duplication*. Springer-Verlag, New York.
163 Orengo, C.A., Jones, D.T. and Thornton, J.M. (1994) Protein superfamilies and domain superfolds. *Nature*, **372**, 631-634.
164 Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B. and Thornton, J.M. (1997) CATH--a hierarchic classification of protein domain structures. *Structure*, **5**, 1093-1108.
165 Pakula, A.A., Young, V.B. and Sauer, R.T. (1986) Bacteriophage lambda cro mutations: effects on activity and intracellular degradation. *Proc Natl Acad Sci U S A*, **83**, 8829-8833.
166 Papp, B., Pal, C. and Hurst, L.D. (2003) Dosage sensitivity and the evolution of gene families in yeast. *Nature*, **424**, 194-197.
167 Paques, F. and Haber, J.E. (1999) Multiple pathways of recombination induced by double-strand breaks in Saccharomyces cerevisiae. *Microbiol Mol Biol Rev*, **63**, 349-404.
168 Parrini, C., Taddei, N., Ramazzotti, M., Degl'Innocenti, D., Ramponi, G., Dobson, C.M. and Chiti, F. (2005) Glycine residues appear to be evolutionarily conserved for their ability to inhibit aggregation. *Structure*, **13**, 1143-1151.
169 Pascarella, S. and Argos, P. (1992) Analysis of insertions/deletions in protein structures. *J Mol Biol*, **224**, 461-471.
170 Patthy, L. (1999) Genome evolution and the evolution of exon-shuffling--a review. *Gene*, **238**, 103-114.
171 Pauling, L. (1953) Protein Interactions .4. Aggregation of Globular Proteins. *Discussions of the Faraday Society*, 170-176.
172 Pauling, L. and Corey, R.B. (1951) The pleated sheet, a new layer configuration of polypeptide chains. *Proc Natl Acad Sci U S A*, **37**, 251-256.
173 Pauling, L., Corey, R.B. and Branson, H.R. (1951) The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. *Proc Natl Acad Sci U S A*, **37**, 205-211.
174 Payens, T.A.J. (1983) Why are enzymes so large? *Trends in Biochemical Sciences*, **8**, 46.
175 Pearson, C.E., Nichol Edamura, K. and Cleary, J.D. (2005) Repeat instability: mechanisms of dynamic mutations. *Nat Rev Genet*, **6**, 729-742.
176 Pearson, W.R. (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol*, **183**, 63-98.
177 Peifer, M., Berg, S. and Reynolds, A.B. (1994) A repeating amino acid motif shared by proteins with diverse cellular roles. *Cell*, **76**, 789-791.
178 Peng, K., Obradovic, Z. and Vucetic, S. (2004) Exploring bias in the Protein Data Bank using contrast classifiers. *Pac Symp Biocomput*, 435-446.

179 Pereira-Leal, J.B., Levy, E.D. and Teichmann, S.A. (2006) The origins and evolution of functional modules: lessons from protein complexes. *Philos Trans R Soc Lond B Biol Sci*, **361**, 507-517.
180 Pereira-Leal, J.B. and Teichmann, S.A. (2005) Novel specificities emerge by stepwise duplication of functional modules. *Genome Res*, **15**, 552-559.
181 Petit, M.A., Mesas, J.M., Noirot, P., Morel-Deville, F. and Ehrlich, S.D. (1992) Induction of DNA amplification in the Bacillus subtilis chromosome. *Embo J*, **11**, 1317-1326.
182 Ponting, C.P. and Russell, R.B. (2000) Identification of distant homologues of fibroblast growth factors suggests a common ancestor for all beta-trefoil proteins. *J Mol Biol*, **302**, 1041-1047.
183 Qian, J., Luscombe, N.M. and Gerstein, M. (2001) Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. *J Mol Biol*, **313**, 673-681.
184 Reams, A.B. and Neidle, E.L. (2004) Selection for gene clustering by tandem duplication. *Annu Rev Microbiol*, **58**, 119-142.
185 Remm, M., Storm, C.E. and Sonnhammer, E.L. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol*, **314**, 1041-1052.
186 Robson, B. (1999) Beyond proteins. *Trends in Biotechnology*, **17**, 311-315.
187 Rocha, E.P. (2003) DNA repeats lead to the accelerated loss of gene order in bacteria. *Trends Genet*, **19**, 600-603.
188 Rocha, E.P. (2006) The quest for the universals of protein evolution. *Trends Genet*, **22**, 412-416.
189 Rooman, M.J., Rodriguez, J. and Wodak, S.J. (1990) Automatic definition of recurrent local structure motifs in proteins. *J Mol Biol*, **213**, 327-336.
190 Rossmann, M.G., Moras, D. and Olsen, K.W. (1974) Chemical and biological evolution of nucleotide-binding protein. *Nature*, **250**, 194-199.
191 Sasidharan, R. and Chothia, C. (2007) The selection of acceptable protein mutations. *Proc Natl Acad Sci U S A*, **104**, 10080-10085.
192 Saupe, S., Turcq, B. and Begueret, J. (1995) A gene responsible for vegetative incompatibility in the fungus Podospora anserina encodes a protein with a GTP-binding motif and G beta homologous domain. *Gene*, **162**, 135-139.
193 Sawyer, S.A., Kulathinal, R.J., Bustamante, C.D. and Hartl, D.L. (2003) Bayesian analysis suggests that most amino acid replacements in Drosophila are driven by positive selection. *J Mol Evol*, **57 Suppl 1**, S154-164.
194 Schultz, G.E. and Schirmer, R.H. (1979) *Principles of Protein Structure*. Springer-Verlag.
195 Serebrovsky, A.S. (1938) Genes scute and achaete in Drosophila melanogaster and a hypothesis of gene divergency. *C. R. Acad. Sci.URSS*, **19**, 77-81.
196 Shimizu, T., Mitsuke, H., Noto, K. and Arai, M. (2004) Internal gene duplication in the evolution of prokaryotic transmembrane proteins. *J Mol Biol*, **339**, 1-15.
197 Shindyalov, I.N. and Bourne, P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng*, **11**, 739-747.
198 Sikorski, R.S., Boguski, M.S., Goebl, M. and Hieter, P. (1990) A repeating amino acid motif in CDC23 defines a family of proteins and a new relationship among genes required for mitosis and RNA synthesis. *Cell*, **60**, 307-317.

199 Simon, D.M., Clarke, N.A., McNeil, B.A., Johnson, I., Pantuso, D., Dai, L., Chai, D. and Zimmerly, S. (2008) Group II introns in eubacteria and archaea: ORF-less introns and new varieties. *Rna*, **14**, 1704-1713.
200 Skrabanek, L. and Wolfe, K.H. (1998) Eukaryote genome duplication - where's the evidence? *Curr Opin Genet Dev*, **8**, 694-700.
201 Smith, T.F. and Waterman, M.S. (1981) Identification of common molecular subsequences. *J Mol Biol*, **147**, 195-197.
202 Snel, B., Bork, P. and Huynen, M. (2000) Genome evolution. Gene fusion versus gene fission. *Trends Genet*, **16**, 9-11.
203 Somero, G.N. (1995) Proteins and temperature. *Annu Rev Physiol*, **57**, 43-68.
204 Sonnhammer, E.L. and Kahn, D. (1994) Modular arrangement of proteins as inferred from analysis of homology. *Protein Sci*, **3**, 482-492.
205 Srere, P.A. (1984) Why are enzymes so big? *Trends in Biochemical Sciences*, **9**, 387-390.
206 Stadler, L.J. (1929) Chromosome Number and the Mutation Rate in Avena and Triticum. *Proc Natl Acad Sci U S A*, **15**, 876-881.
207 Steinert, P.M., Candi, E., Tarcsa, E., Marekov, L.N., Sette, M., Paci, M., Ciani, B., Guerrieri, P. and Melino, G. (1999) Transglutaminase crosslinking and structural studies of the human small proline rich 3 protein. *Cell Death Differ*, **6**, 916-930.
208 Stephens, S.G. (1951) Possible Significance of Duplication in Evolution. *Advances in Genetics Incorporating Molecular Genetic Medicine*, **4**, 247-265.
209 Steward, A., Adhya, S. and Clarke, J. (2002) Sequence conservation in Ig-like domains: the role of highly conserved proline residues in the fibronectin type III superfamily. *J Mol Biol*, **318**, 935-940.
210 Street, T.O., Rose, G.D. and Barrick, D. (2006) The role of introns in repeat protein gene formation. *J Mol Biol*, **360**, 258-266.
211 Swindells, M.B. (1995) A procedure for detecting structural domains in proteins. *Protein Sci*, **4**, 103-112.
212 Taylor, J.S. and Raes, J. (2004) Duplication and divergence: the evolution of new genes and old ideas. *Annu Rev Genet*, **38**, 615-643.
213 Taylor, W.R. (1986) The classification of amino acid conservation. *J Theor Biol*, **119**, 205-218.
214 Teichmann, S.A. (2002) The constraints protein-protein interactions place on sequence divergence. *J Mol Biol*, **324**, 399-407.
215 Teichmann, S.A. and Chothia, C. (2000) Immunoglobulin superfamily proteins in Caenorhabditis elegans. *J Mol Biol*, **296**, 1367-1383.
216 Teichmann, S.A., Park, J. and Chothia, C. (1998) Structural assignments to the Mycoplasma genitalium proteins show extensive gene duplications and domain rearrangements. *Proc Natl Acad Sci U S A*, **95**, 14658-14663.
217 Teichmann, S.A., Rison, S.C., Thornton, J.M., Riley, M., Gough, J. and Chothia, C. (2001a) The evolution and structural anatomy of the small molecule metabolic pathways in Escherichia coli. *J Mol Biol*, **311**, 693-708.
218 Teichmann, S.A., Rison, S.C., Thornton, J.M., Riley, M., Gough, J. and Chothia, C. (2001b) Small-molecule metabolism: an enzyme mosaic. *Trends Biotechnol*, **19**, 482-486.
219 Tenaillon, O., Denamur, E. and Matic, I. (2004) Evolutionary significance of stress-induced mutagenesis in bacteria. *Trends Microbiol*, **12**, 264-270.
220 Theil, E.C. (1987) Ferritin: structure, gene regulation, and cellular function in animals, plants, and microorganisms. *Annu Rev Biochem*, **56**, 289-315.

221 Thomas, E.E. (2005) Short, local duplications in eukaryotic genomes. *Curr Opin Genet Dev*, **15**, 640-644.
222 Tischler, G. (1915) Chromosomenzahl, Form und Individualitat in Planzenreiche. *Progr. Rei Bot.*, **5**, 164.
223 Todd, A.E., Orengo, C.A. and Thornton, J.M. (2001) Evolution of function in protein superfamilies, from a structural perspective. *J Mol Biol*, **307**, 1113-1143.
224 Tripp, K.W. and Barrick, D. (2004) The tolerance of a modular protein to duplication and deletion of internal repeats. *J Mol Biol*, **344**, 169-178.
225 van Nimwegen, E. (2003) Scaling laws in the functional content of genomes. *Trends Genet*, **19**, 479-484.
226 Veitia, R.A. (2004) Gene dosage balance in cellular pathways: implications for dominance and gene duplicability. *Genetics*, **168**, 569-574.
227 Verstrepen, K.J., Jansen, A., Lewitter, F. and Fink, G.R. (2005) Intragenic tandem repeats generate functional variability. *Nat Genet*, **37**, 986-990.
228 Vogel, C., Bashton, M., Kerrison, N.D., Chothia, C. and Teichmann, S.A. (2004a) Structure, function and evolution of multidomain proteins. *Curr Opin Struct Biol*, **14**, 208-216.
229 Vogel, C., Berzuini, C., Bashton, M., Gough, J. and Teichmann, S.A. (2004b) Supra-domains: evolutionary units larger than single protein domains. *J Mol Biol*, **336**, 809-823.
230 Vogel, C., Teichmann, S.A. and Pereira-Leal, J. (2005) The relationship between domain duplication and recombination. *J Mol Biol*, **346**, 355-365.
231 Wagner, A. (1998) The fate of duplicated genes: loss or new function? *Bioessays*, **20**, 785-788.
232 Wagner, A. (2002) Selection and gene duplication: a view from the genome. *Genome Biol*, **3**, reviews1012.
233 Wang, Z.X. (1998) A re-estimation for the total numbers of protein folds and superfamilies. *Protein Eng*, **11**, 621-626.
234 Waterman, M.S. and Eggert, M. (1987) A new algorithm for best subsequence alignments with application to tRNA-rRNA comparisons. *J Mol Biol*, **197**, 723-728.
235 Waterman, M.S. and Vingron, M. (1994) Rapid and accurate estimates of statistical significance for sequence data base searches. *Proc Natl Acad Sci U S A*, **91**, 4625-4628.
236 Wetlaufe, D.B. (1973) Nucleation, Rapid Folding, and Globular Intrachain Regions in Proteins. *Proceedings of the National Academy of Sciences of the United States of America*, **70**, 697-701.
237 Wilson, C.A., Kreychman, J. and Gerstein, M. (2000) Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J Mol Biol*, **297**, 233-249.
238 Wolf, Y.I., Grishin, N.V. and Koonin, E.V. (2000) Estimating the number of protein folds and families from complete genome data. *J Mol Biol*, **299**, 897-905.
239 Wolffe, A.P. and Guschin, D. (2000) Review: chromatin structural features and targets that regulate transcription. *J Struct Biol*, **129**, 102-122.
240 Wolynes, P.G. (1996) Symmetry and the energy landscapes of biomolecules. *Proc Natl Acad Sci U S A*, **93**, 14249-14255.

241 Wright, C.F., Teichmann, S.A., Clarke, J. and Dobson, C.M. (2005) The importance of sequence diversity in the aggregation and evolution of proteins. *Nature*, **438**, 878-881.
242 Wright, S. (1931) Evolution in Mendelian Populations. *Genetics*, **16**, 97-159.
243 Wuchty, S. (2001) Scale-free behavior in protein domain networks. *Mol Biol Evol*, **18**, 1694-1702.
244 Xu, Z., Horwich, A.L. and Sigler, P.B. (1997) The crystal structure of the asymmetric GroEL-GroES-(ADP)7 chaperonin complex. *Nature*, **388**, 741-750.
245 Ye, Y. and Godzik, A. (2004) Comparative analysis of protein domain organization. *Genome Res*, **14**, 343-353.
246 Young, M. and Cullum, J. (1987) A plausible mechanism for large-scale chromosomal DNA amplification in streptomycetes. *FEBS Lett*, **212**, 10-14.
247 Young, N.M., Williams, R.E., Roy, C. and Yaguchi, M. (1982) Structural comparison of the lectin from sainfoin (Onobrychis viciifolia) with concanavalin A and other D-mannose specific lectins. *Can J Biochem*, **60**, 933-941.
248 Zhang, J., Zhang, Y.P. and Rosenberg, H.F. (2002) Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. *Nat Genet*, **30**, 411-415.
249 Zhang, Y. and Skolnick, J. (2004) Scoring function for automated assessment of protein structure template quality. *Proteins*, **57**, 702-710.
250 Zuker, M. and Somorjai, R.L. (1989) The alignment of protein structures in three dimensions. *Bull Math Biol*, **51**, 55-78.

Résumé :

Les duplications jouent un rôle important dans l'évolution des protéines et sont à l'origine des répétitions intragéniques présentes dans environ 14% des séquences protéiques. Nous avons choisi d'étudier ces répétitions d'un point de vue évolutif. Pour cela, nous avons développé un programme, Swelfe, qui cherche les répétitions à la fois dans les gènes, les séquences d'acides aminés et les structures tridimensionnelles des protéines. Ce programme utilise le même algorithme de programmation dynamique à tous les niveaux et une représentation séquentielle des structures 3D. Les scores et les tests de significativité des répétitions obtenues ont été adaptés pour chaque niveau. Nous avons créé une banque contenant les séquences d'ADN et d'acides aminés correspondant aux structures de la PDB, et comparé Swelfe à DALI pour valider la méthode au niveau des répétitions structurales. Enfin, ce programme est disponible pour la communauté à l'adresse http://bioserv.rpbs.jussieu.fr/swelfe et peut être téléchargé ou utilisé en ligne.

Swelfe a trouvé un nombre important de répétitions dans un ensemble non redondant de séquences nucléiques, séquences protéiques et structures tridimensionnelles, et environ 10% des protéines contiennent des répétitions à au moins un niveau. Cependant, le recouvrement des répétitions aux trois niveaux est assez faible et beaucoup de répétitions ne sont trouvées qu'à un seul niveau, ce qui confirme l'intérêt de cette étude sur les trois niveaux en parallèle Les causes de ce recouvrement faible sont discutées dans la présente thèse. L'étude des répétitions structurales longues montre qu'environ 30% de ces répétitions sont pseudo-symétriques à 180°, comme le sont les deux éléments d'un homo-dimère. L'analyse de ces protéines indique que certaines pourraient effectivement remplacer des dimères.

Abstract :

Duplications play a major role in protein evolution and result in intragenic repeats found in about 14% of protein sequences. We chose to study these repeats from an evolutionary point of view. For this, we developed a program, Swelfe, that looks for repeats in genes, amino acid sequences and three dimensional structures of proteins. This program uses the same dynamic programming algorithm at all levels and a sequential representation of 3D structures. Repeat scores and significance tests are adapted at each level. We created a bank containing DNA sequences and amino acid sequences corresponding to PDB structures, and Swelfe was compared to DALI to validate our method concerning structural repeats. Finally this program is available for the community at http://bioserv.rpbs.jussieu.fr/swelfe and can be downloaded or used online.

Swelfe found an important number of repeats in a non redundant data set of DNA sequences, amino acid sequences and 3D structures and about 10% of proteins contain repeats at last at one level. However the repeat overlap at the three level is weak and most repeats are only found at one level, this confirm the relevance of studying repeats at three levels at the same time. Reasons of the weak overlap are discussed in this thesis. The study of long structural repeats shows that about 30% of these repeats are symmetrical at 180°, as are the two elements of a homodimer. The analysis of these proteins shows that some of them could effectively replace homodimers.