ACADÉMIE DE MONTPELLIER

U N I V E R S I T É D E M O N T P E L L I E R I I

- Sciences et Techniques du Languedoc -

Thèse

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE MONTPELLIER II

Spécialité :	Bioinformatique
Formation Doctorale :	Bioinformatique
École Doctorale :	Sciences Chimiques et Biologiques pour la Santé

Analyse systématique des motifs répétés en tandem dans les séquences protéiques

Par

JULIEN JORDA

Présentée et soutenue publiquement le 15 octobre 2010 devant le Jury composé de :

Dr Maria Anisimova, Institute of Computational Science, ETH Zurich	Rapporteur
Pr Manuel Dauchez, Université de Reims Champagne-Ardenne	Rapporteur
Pr Paul Mangeat, Université de Montpellier II	Président
Dr Andrey Kajava, CNRS, CRBM	Directeur de thèse

« *C'est pour prouver qu'on peut prendre de l'antibiotique dans le vin rouge* »

Dr Lucien Samary, dit Dr Lulu

À mes parents.

À Michel et Christelle.

En souvenir de mon pépé et mi abuela.

Remerciements

Je souhaiterais tout d'abord remercier Maria Anisimova et Manuel Dauchez, les rapporteurs de ce manuscrit, ainsi que Paul Mangeat, le Président, d'avoir accepté d'évaluer et de commenter ce travail.

Je tiens à remercier tout particulièrement Andrey Kajava, qui m'a encadré pendant ces trois années de thèse et bien avant durant mes stages, pour son soutien et sa disponibilité. Merci de m'avoir initié à la recherche en bioinformatique, et de m'avoir laissé un degré de liberté que peu de thésards ont, tout en dispensant de précieux conseils.

Je remercie également le Ministère de l'Education Nationale, de l'Enseignement Supérieur et de la Recherche qui m'a apporté le soutien financier indispensable au déroulement de cette thèse, ainsi que le Professeur Christian Le Peuch pour m'avoir accepté dans l'école doctorale CBS2. Un grand merci à Séverine Bérard pour ses cours, ses conseils et de m'avoir éclairé de sa lanterne dans les profondeurs abyssales d'une thèse en bioinformatique.

Pour ces trois années passées à faire de la recherche, mais aussi à partager de grands moments au « restaurant d'entreprise » (Merci à Daniel des «périphériques» pour cet éclaircissement syntaxique) ainsi qu'aux pots sous les micos, je tiens à remercier l'ensemble du CRBM ainsi que tous ceux qui ont contribué à rendre agréable mon quotidien au labo. Mention spéciale à Pablo Gazon, mon éternel acolyte avec qui j'ai pu faire les 400 coups et partager des moments de virtuosité en termes de malsainité, que ce soit au labo ou en dehors. Un grand merci à Abdullah pour avoir égayé mes derniers moments de thèse, et permis de maintenir la proportion testostéronale suite aux départs des collègues thésards en post-doc.

Une pensée pour la canaille exilée à San Diego, Cyril, ainsi qu'à Binome que je vais bientôt rejoindre en Californie.Je me fourvoierai en remerciant pêle-mêle et sans ordre de préférence: Ben, Krzys, Estelle, Aurélie dit « Dr Le », Sihem, Mathy « Ndeeeeeye », Heiani, Cédric, Pauline, Laetitia « Peach », Aïcha, Sébastien.

Mes pensées vont également aux amis non-CRBMiens mais indispensables à mon bien-être quotidien : Nizam, jeune gourgandin ayant le rôle ingrat de meilleur ami (« nézaam !»); Nano, Franck et David les membres de mon groupe (Suunya rules); Kev, Gillou, Groyom, Bidouille, Fanny, la famille Cenatiempo et certainement tout ceux que j'oublie.

Mes derniers remerciements vont à ma famille, ici et ailleurs, présente ou absente, et tout particulièrement mes parents, pour leur soutien sans faille tout au long de mes études et leurs encouragements; pour leurs valeurs et principes qu'ils m'ont inculqués et que j'essaie d'appliquer au quotidien. Je remercie aussi mon frère (mon frireuh), certainement la personne qui me connaît le mieux et avec qui je prends des fous rires à m'en faire décoller les rétines. Enfin, un énorme merci à ma Titelle avec qui je partage tout de ma vie depuis 5 ans, et m'a toujours soutenu moralement dans mes choix au point de me suivre chez les yankees.

Merci aux Messages à caractère informatif, à Strip-tease, à Bob l'éponge… d'être des sources d'inspiration et de donner un sens à mon humour.

Avant - propos

Depuis que l'homme est homme, il n'a de cesse de vouloir élucider l'énigme entourant les mécanismes régissant son existence. Grâce aux avancées technologiques, il a pu repousser les limites du domaine de l'observable, cherchant à percer le mystère de cette complexe machinerie qu'est un être vivant. L'apparition de la génétique a permis de découvrir que la plupart de nos caractéristiques sont codées à l'intérieur de nos cellules sous forme d'une information. Pendant des millénaires, l'homme a pu manipuler inconsciemment cette information par la domestication des animaux, la greffe de plantes ou le croisement et la sélection des espèces. Autrement connue sous le nom de *génome*, elle a pour support des molécules d'ADN organisées en chromosomes. Ces molécules sont constituées d'un enchainement particulier de *bases* représentées par un alphabet de 4 lettres : A, C, G et T. De cet alphabet, certaines phrases, ou *séquences*, situées à un endroit particulier sur les chromosomes, ou *locus*, vont constituer l'unité élémentaire de l'information génétique que l'on appelle un *gène*. Ce sont ces mêmes gènes qui déterminent la couleur de nos cheveux ou notre taille, et qui lorsqu'ils sont altérés sont responsables de maladies génétiques ou qui, combinés à des facteurs environnementaux, peuvent prédisposer et mener à des pathologies. L'analyse des gènes et l'identification de leurs fonctions sont donc d'importance majeure dans la compréhension du fonctionnement du vivant, dans la prévention et le traitement de ces pathologies.

Cette analyse à un niveau global est possible grâce à une discipline née au cours des dernières décennies : la génomique. Le projet Génome humain, achevé en avril 2003, a permis de séquencer les 3 milliards de bases constituant le patrimoine génétique humain. D'autres génomes, tels que les génomes de bactéries ou de certains eucaryotes ont pu bénéficier des avancées technologiques dans le domaine et être complètement séquencés. Néanmoins, ceci n'est qu'une première étape dans l'analyse des génomes. Chaque individu est unique, cette spécificité est due aux légères variations présentes dans ses gènes, appelées *polymorphismes*.

En 2007, le premier séquençage complet du génome d'un seul et même individu humain a pu être réalisé en seulement deux mois, montrant que la variabilité d'un humain à l'autre est de moins de 1%. Ces polymorphismes constituent des « hotspots » évolutifs et jouent un rôle crucial dans la compréhension des mécanismes d'évolution. Une des sources les plus importantes de variabilité dans les génomes sont les répétitions dans les séquences. Ces répétitions peuvent être de deux types : les répétitions dispersées et les séquences répétées en

tandem. Ces dernières sont composées de motifs répétés successivement dans une séquence, similaires entre eux. Les séquences répétées en tandem peuvent être retrouvées dans les régions intergéniques (ou *ADN non codant*), ou dans les gènes (ou *ADN codant*) et par extension dans les protéines. Du fait de leur grande variabilité entre espèces (ou individus d'une même espèce), elles représentent des candidats de choix pour réaliser un profil génétique d'un individu, utilisé dans les tests de paternité ou l'identification de l'organisme source. Néanmoins, on peut mesurer l'impact direct des répétitions en tandem au niveau des protéines, elles aussi hautement polymorphiques et souvent impliquées dans des fonctions majeures. De même, leur incidence dans de nombreuses maladies neurodégénératives ou des facteurs de virulences d'agents pathogènes en font des acteurs non négligeables dont le rôle est à définir. La découverte et les prédictions structure-fonction des régions répétées peuvent conduire à l'identification de cibles pour de nouveaux antibiotiques et vaccins contre des maladies infectieuses émergeantes et au développement d'inhibiteurs de « l'amyloïdogénèse », souvent observé dans les maladies neurodégénératives. Les objectifs de cette thèse sont d'identifier et analyser à grande échelle les répétitions en tandem dans les banques de données de séquences protéiques. Le volume de ces données augmentant exponentiellement chaque année, une analyse exhaustive de ces banques va nécessiter le recours à l'outil informatique, et plus spécifiquement des techniques du domaine de la *Bioinformatique*.

TABLE DES MATIERES

I - INTRODUCTION

I-1 – Généralités sur les séquences

1- Les séquences nucléotidiques

2- Les séquences protéiques

I-2 - Séquençage des génomes

1- Historique

2- La méthode de Sanger

3- Les techniques de séquençage

4- Séquençage de nouvelle génération (NGS)

5- Observations préliminaires des génomes

I-3 –Les différents types de répétitions dans les génomes

1- Répétitions dispersées

2- Répétitions en tandem

a. les répétitions en tandem dans l'ADN

i. L'ADN satellite

ii. Les minisatellites

iii. Les microsatellites

iv. Autres répétitions nucléotidiques

b. les répétitions en tandem dans les protéines

I-4 - Origine et évolution des répétitions en tandem

I-5 – Structures et fonctions des séquences répétées en tandem

1- Structures des séquences répétées en tandem

2- Fonctions biologiques des séquences répétées en tandem

I-6 - Pathologies associées aux séquences répétées en tandem

1- Généralités

2- Maladies neurodégénératives et développementales

3- Prion
4- Cancers
5- Rôle dans l'évolution et la virulence d'agents pathogènes

I-7– Identification des répétitions en tandem

1- L'identification avant les séquençages
2- Quelques notions algorithmiques
    a. La complexité algorithmique
    b. La programmation dynamique et les algorithmes d'alignement
3- Les programmes d'identification de répétitions en tandem
    a. Les difficultés liées à l'identification
    b. Algorithmes exhaustifs
    c. Heuristiques et améliorations de programmation dynamique
    d. Profils et Modèle de Markov caché (HMM)

II- Objectifs

III- Résultats

Chapitre I : Identification ab initio des répétitions en tandem dans les séquences protéiques

Chapitre II : Construction de PRDB : Protein tandem Repeat DataBase

Chapitre III : Caractérisation des propriétés physicochimiques, structurelles, évolutives et fonctionnelles des homorepeats protéiques

Chapitre IV : Analyse de la perfection des répétitions et leur désordre intrinsèque

IV- Discussion et Perspectives

1- Récapitulatif
2- T-REKS et PRDB : Deux nouveaux outils pour l'analyse exhaustive des répétitions en tandem dans les séquences protéiques
3- Caractérisation des propriétés physicochimiques, structurelles, évolutives et fonctionnelles des homorepeats
4- La perfection des répétitions est inversement corrélée à leur désordre intrinsèque

V- Conclusion

VI- Bibliographie

VII- Annexes

LISTE DES ABREVIATIONS

- ADN : acide désoxyribonucléique.
- ADNc : ADN complémentaire.
- ARN : acide ribonucléique.
- ARNm : ARN messager.
- CNV : Copy Number Variant.
- EST : Expressed Sequence Tag.
- GWAS: Genome-wide Association Studies.
- GO: Gene Ontology.
- HMM : Hidden Markov Model.
- Indel : évènement mutationnel d'insertion ou délétion dans une séquence biologique.
- Îlots CpG : régions riches en dinucléotides CG qui sont des sites de méthylation, souvent impliqués dans la régulation de l'expression des gènes.
- MLVA : Multiple Loci VNTR Analysis.
- NCBI : National Center for Biotechnology Information.
- Pb, Kb, Mb : paire de bases, Kilobases, Mégabases.
- PFAM : Protein Families.
- Psim : coefficient mesurant la similarité d'une répétition protéique.
- RCP : Repeat-Containing Protein. Protéine présentant une ou plusieurs répétitions.
- SNP : Single Nucleotide Polymorphism.
- SSR : Simple Sequence Repeat.
- STR : Short Tandem Repeat.
- Top-IDP : échelle évaluant le désordre intrinsèque dans une séquence protéique.
- TR : Tandem repeat ou répétition en tandem.
- VNTR : Variable Number Tandem Repeat.
- WGS: Whole Genome Shotgun.

1-INTRODUCTION

I-1 - Propriétés inhérentes aux séquences d'acides nucléiques et protéiques

Une séquence peut être définie de manière générale comme une série ordonnée d'unités simple. Dans le contexte biologique elle fait référence aux macromolécules, assimilées à des polymères composés de différents monomères. La suite ordonnée de ces monomères, liés par des liaisons fortes dites « covalentes », constitue une séquence biologique. La nature des monomères et leur ordre vont conférer des propriétés différentes d'une séquence à une autre.

I-1-1 – Propriétés générales des acides nucléiques

Les acides nucléiques sont des polymères dont l'unité simple, ou monomère, est le nucléotide. Il existe deux types d'acides nucléiques : l'acide désoxyribonucléique(ou ADN), support de l'information génétique, et l'acide ribonucléique (ou ARN), qui est une copie de l'ADN. Les nucléotides sont chacun constitués d'acide phosphorique, d'un sucre (le désoxyribose), et d'une base azotée qui permet de les différencier. Dans l'ADN, ils sont au nombre de 4 et sont représentées par des lettres : A pour l'adénine, G pour la guanine, T pour la thymine et C pour la cytosine (Aet G sont appelées bases puriques, T et C bases pyrimidiques) (figure 1). Dans une séquence, l'enchainement des nucléotides se fait via des liaisons covalentes de type phosphodiester. L'ajout d'un nucléotide à une chaine implique la liaison du groupement hydroxyle du 3ème carbone du sucre du dernier nucléotide de la chaine avec le groupement phosphate du 5ème carbone du nucléotide à incorporer, ce mécanisme orientant ainsi la polymérisation d'une chaine d'ADN dans le sens 5' à 3' (figure 1).

Généralement, l'ADN est assemblé en deux chaines elles-mêmes enroulées sous une forme de double hélice.

Les deux brins sont liés ou « « hybridés » entre eux par des liaisons transversales plus faibles appelées « liaisons hydrogène ». Ces liaisons sont établies via une règle de complémentarité canonique du fait de l'encombrement stérique des bases, ainsi :

- C ne peut se lier qu'avec G et inversement (3 liaisons hydrogène)
- A ne peut se lier qu'avec T et inversement (2 liaisons hydrogène)

On dit que ces deux brins d'ADN sont complémentaires et anti-parallèles (ils sont assemblés dans des sens contraires) .Cette structure hélicoïdale étant principalement due aux liaisons hydrogènes, la séparation (ou dénaturation) des brins peut être facilement réalisée lors de mécanismes tels que la réplication ou la transcription.

C’est donc l’ordre, la nature et le nombre de nucléotides qui vont déterminer l’information génétique portée par l’ADN. L’expression de cette information constitue le plus souvent un

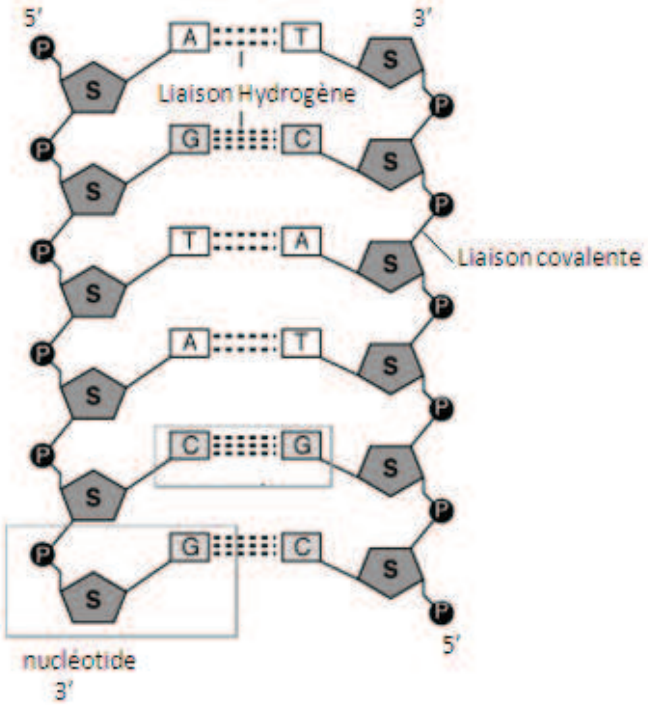


Figure 1 – Représentation schématique du double brin d’ADN. One le modélise généralement comme une « échelle », où les liaisons covalentes constituent les montants tandis que les bases azotées se faisant face, liées par des liaisons hydrogènes forment les barreaux.

« phénotype » (le caractère observable), dont les règles et le codage sont définis par l’unité élémentaire de l’information génétique, appelée gène. Ce phénotype se concrétise la plupart du temps par la présence ou l’absence d’une ou plusieurs protéines. La procédure de l’expression d’un gène aboutissant à la synthèse d’une protéine passe par plusieurs étapes, notamment :

- La transcription, qui copie le code porté par le gène dans une séquence nucléotidique particulière appelée ARNm ou «ARN messager». Il diffère de l’ADN par son organisation en simple brin, sa composition en sucre (impliquant un ribose plutôt que le désoxyribose) et en bases où la thymine(T) est remplacée par une base uracile (U).
- La traduction, qui consiste en l’interprétation de la séquence de l’ARNm en une séquence protéique.
- D’autres mécanismes comme les modifications post-transcriptionnelles vont pouvoir induire une variabilité des propriétés des ARNm, ou même conduire à la synthèse de

séquences protéiques différentes à partir d'un même gène. De même, des modifications post-traductionnelles peuvent avoir lieu sur les protéines, conférant là encore de nouvelles propriétés.

I-1-2 – Les séquences protéiques

Les protéines sont des composants essentiels des organismes et participent à tous les processus existants dans une cellule. L'unité simple pour une protéine est l'acide aminé, l'enchainement de ces unités constitue une séquence protéique ou « structure primaire » de la protéine. Le passage de la séquence du gène, comptant 4 nucléotides, à la séquence protéique constituée de 20 acides aminés différents de par la nature de leur chaine latérale est régi par un système de correspondance appelé code génétique (figure 2).

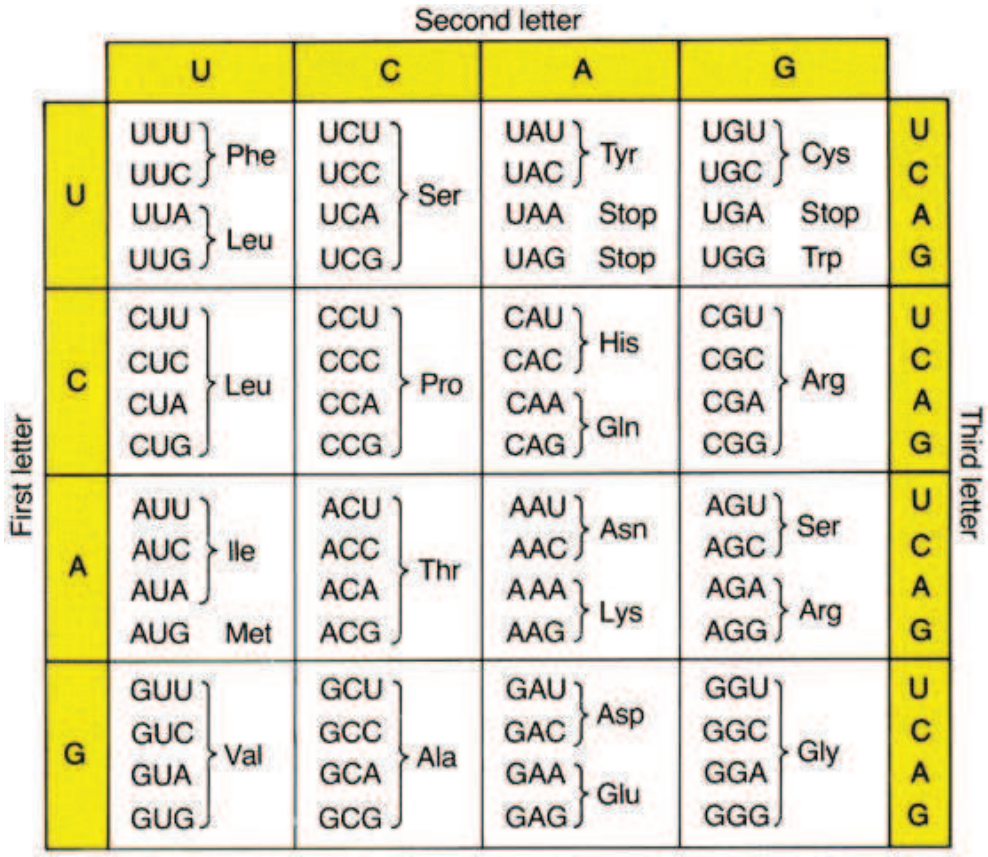


Figure 2 – Le code génétique. Plusieurs triplets peuvent coder pour un même acide aminé, le code génétique est dit redondant. Les 20 acides aminés sont representés dans le tableau par leur notation 3 lettres : l'alanine (Ala ou A), la glycine (Gly ou G), la valine (Val ou V), la leucine (Leu ou L), l'isoleucine (Ile ou I) la Phénylalanine (Phe ou F),le tryptohane (Trp ou W), la sérine(Ser ou S), la méthionine (Met ou M), la thréonine (Thr ou T), la tyrosine (Tyr ou Y), la proline (Pro ou P), l'aspartate (Asp ou D), le glutamate (Glu ou E), l'asparagine (Asn ou N), la glutamine (Gln ou Q), la lysine (Lys ou K), l'arginine (Arg ou R), l'histidine(His ou H) et la cystéine(Cys ou C).

Il repose sur la combinaison des 4 nucléotides en triplets, ou « codons », qui vont permettre de coder pour les 20 acides aminés. On obtient ainsi $4^3 = 64$ possibilités de combinaisons pour un codon. La différence entre ces deux nombres est due à la redondance du code génétique, autorisant plusieurs codons de coder pour un même acide aminé. De même ce code est dit « non chevauchant » car une même base ne peut servir dans la construction d'un autre codon.
Enfin, certains codons sont destinés à l'initiation de la traduction (AUG, codant pour une méthionine) ou la terminaison, appelés codons STOP (UAA, UAG et UGA).
L'enchainement des acides aminés dans une séquence protéique se fait par une liaison amide, appelée aussi liaison peptidique. Elle fait intervenir l'atome de carbone du groupe carboxyle du dernier acide aminé de la séquence avec l'azote de l'acide aminé ajouté. De ce fait, la séquence protéique possède une extrémité porteuse d'un groupe amine libre (N-terminale ou Nt, correspondant par convention au premier acide aminé) et l'autre avec un groupe carboxyle (C-terminale ou Ct). On dit que la séquence est « polarisée ».
Contextuellement, on appelle les acides aminés constitutifs d'une séquence les « résidus » (l'origine de ce mot provient du fait que durant le processus de polymérisation, chaque liaison amide relâche une molécule d'eau, ne laissant qu'un résidu du monomère initial dans la séquence) (figure 3).

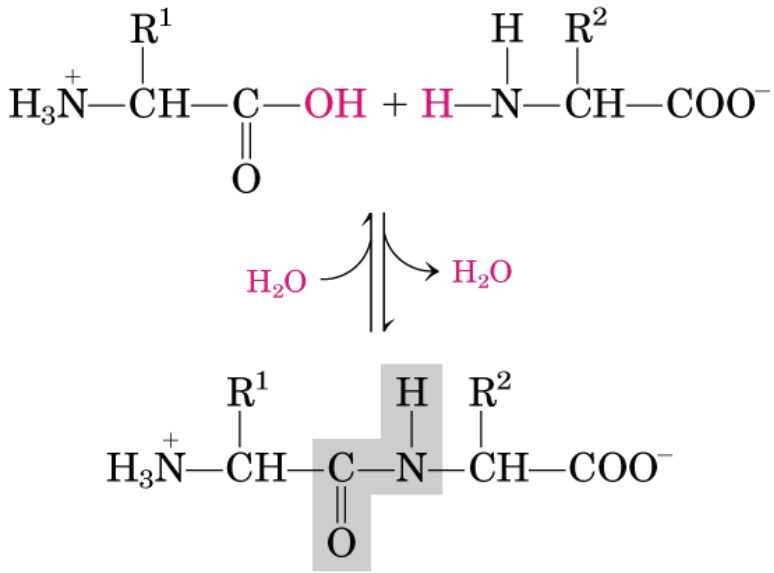


Figure 3 – Liaison peptidique entre deux acides aminés. R1 et R2 sont les chaines latérales respectives des deux acides aminés. La création d'une liaison amide relâche une molécule d'eau.

Les résidus vont pouvoir interagir entre eux, souvent par le biais de liaisons hydrogène entre les groupes amide et carbonyle du squelette peptidique, et constituer des repliements locaux de la chaîne que l'on appelle « structures secondaires ».

Les structures secondaires les plus communes sont l'hélice alpha et le feuillet beta (figure 4).

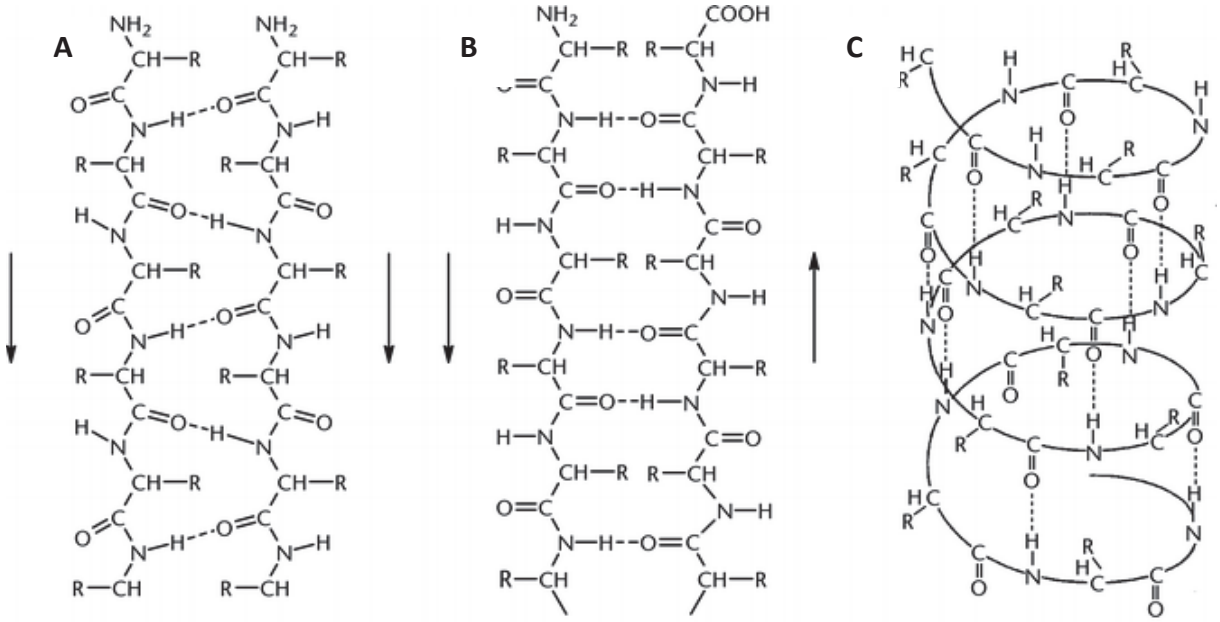


Figure 4 – Les structures secondaires. (A) Feuillet beta parallèle (B) Feuillet Beta antiparallèle, (C) Hélice alpha

D'autres structures secondaires existent, telles que les hélices 3/10, les hélices π, les hélices de type II (polyproline ou polyglycine) bien que retrouvées de manière exceptionnelle dans les protéines naturelles.

Les structures secondaires vont pouvoir acquérir un arrangement dans l'espace grâce aux interactions entre les chaines latérales des acides aminés, on parle de structure tertiaire (figure 5). Ces interactions peuvent être des liaisons faibles ou des liaisons covalentes entre deux résidus cystéine appelées ponts disulfures.

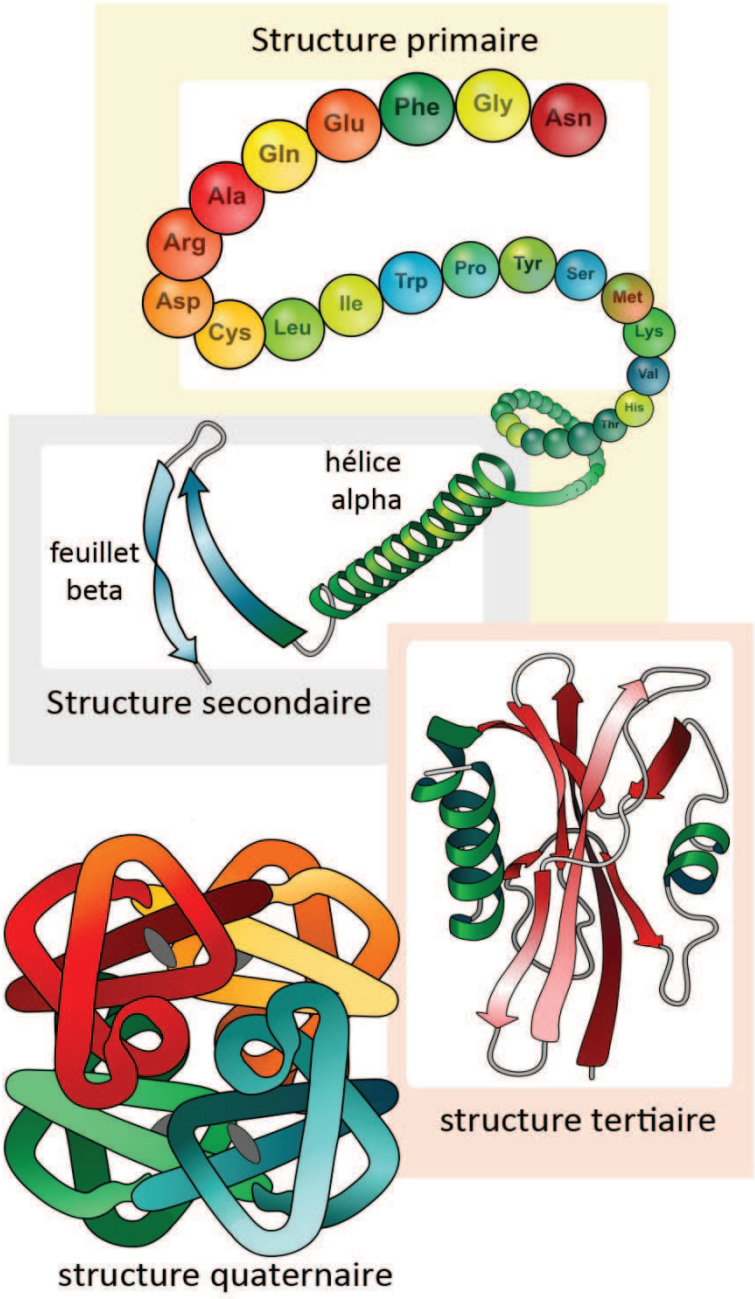
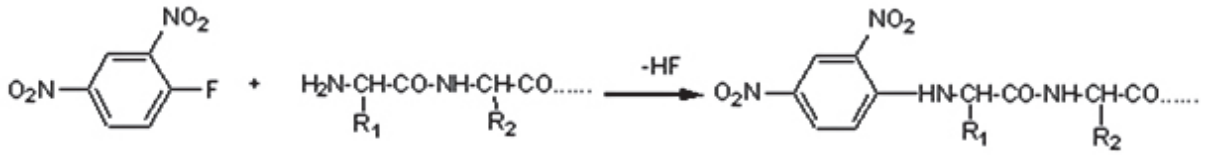


Figure 5 – Différents niveaux de structures dans les protéines.

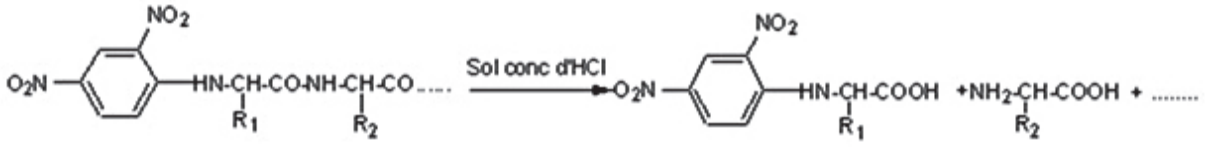
Certaines protéines peuvent être assemblées sous forme de complexes, on dit qu'ils sont multimériques ou encore des oligomères (en opposition à une chaine qui est appelée monomère ou sous-unité). C'est le cas de l'hémoglobine, tétramère protéique chargé de transporter l'oxygène dans le sang des vertébrés. Elle est constituée de deux sous-unités α et deux sous-unités β, chacune possédant en son sein une molécule d'hème.

La structure d'une protéine et son organisation tridimensionnelle est donc intimement corrélée à sa séquence.

Historiquement, la première séquence qui a pu être publié est celle d'une protéine : l'insuline. Elle a été identifiée par Frederick Sanger en 1955, grâce à un réactif qu'il a préalablement mis au point, le 1-fluoro-2,4-dinitrobenzène (DNFB) autrement appelé le r*éactif de Sanger*, et qui découpe les acides aminés à l'extrémité N-terminale de la protéine, le groupement amino agissant comme un réactif nucléophile pour le DNFB et se substitue à l'atome de fluor sur le noyau aromatique:



Le peptide est ensuite hydrolysé :



L'acide peut être par la suite analysé par chromatographie.

Néanmoins, la connaissance seule de la séquence ne permet généralement pas de prédire la structure tertiaire d'une protéine, et nécessite différentes approches bioinformatiques basés sur l'homologie de séquences de protéines dont la structure est connue, ou des algorithmes basés sur des modèles heuristiques.

I-2 - Séquençage des génomes

I-2-1 – Historique

C'est en 1977 que Frederick Sanger propose le premier génome complet d'un organisme biologique, le *bactériophage phi X174*, grâce à une technique de séquençage utilisant la synthèse enzymatique, communément appelée *méthode de Sanger*(Sanger et al., 1977). Il faudra attendre 1995 pour la publication de la première séquence complète d'un génome bactérien, celui d'*Haemophilus influenzae* (Fleischmann et al., 1995). Le Projet Génome Humain (ou Human genome project) entrepris initialement en 1990 par 4 pays (USA, France, Grande-Bretagne, Japon) et rejoint par la suite par d'autres centres de séquençage pour constituer le International Human Genome Sequencing Consortium (IHGSC), a pour objectif

d'étudier le génome humain, mais aussi en parallèle ceux de la souris, de la levure *S. cerevisiae* et du nématode *C. elegans*.

En mars 1999, le séquençage du génome humain à grande échelle a été entamé par le IHGSC et Celera Genomics, société privée fondé par James Craig Venter. S'en est suivi une course au séquençage entre ces deux entités pour aboutir en 2001 au premier « brouillon » de la séquence du génome humain(Lander et al., 2001), dénommée ainsi du fait du réassemblage des fragments d'ADN qui comportait un grand nombre de trous (« gaps ») encore à déterminer. Après une phase de finition de séquençage des gaps, une version finale de la séquence a pu être publiée en 2004 par le consortium international. Depuis, les projets de séquençage se sont vu multiplier et c'est en 2007 que le génome d'un individu est pour la première fois complètement séquencé(Wheeler et al., 2008) (l'individu en question étant James Watson, un des « pères » de la découverte de la structure de l'ADN). Les coûts de moins en moins élevés des séquençages, et la rapidité des techniques de nouvelle génération offrent de nouvelles perspectives, tant dans l'identification des gènes impliqués dans certaines pathologies que dans la compréhension des variations génétiques entre individus. Un exemple est le projet "1000 genome projects" lancé en 2008, qui prévoit le séquençage anonyme de 1000 personnes afin de réaliser un catalogue des variations génétiques humaines que l'on retrouve à une fréquence d'au moins 1%.

I-2-2 – La méthode de Sanger:

Le principe de cette méthode consiste à initier la polymérisation de l'ADN à l'aide d'un petit oligonucléotide, appelé amorce, qui va s'hybrider sur une partie du fragment d'ADN à séquencer. L'élongation de l'amorce est réalisée par des « séquenases» qui sont des ADN polymérases thermostables (on utilisait initialement le fragment de Klenow, une ADN polymérase I dépourvue du domaine exonucléasique 5'→3'). On rajoute dans le milieu les quatre désoxyribonucléotides (dATP, dCTP, dGTP, dTTP) , ainsi qu'en faible quantité l'un des quatre didésoxynucléotides (ddATP, ddCTP, ddGTP ou ddTTP). Les didésoxynucléotides une fois incorporés dans la chaine en cours d'élongation, vont jouer le rôle de terminateurs de chaine. Il en résulte la présence de plusieurs fragments d'ADN de taille variable qui pourront être séparés par la suite par électrophorèse (anciennement sur gel de polyacrylamide, les technologies actuelles utilisant plutôt des capillaires avec un polymère).

I-2-3 – Les techniques de séquençage

Dérivées de la méthode de Sanger, les techniques de séquençage via des séquenceurs permettent de rendre possible le décryptage de génomes entiers. Elles consistent en deux stratégies principales (figure 6) :

- l'approche « WGS » (whole genome shotgun). Cette technique crée, à partir de clones, plusieurs banques de fragments d'ADN génomique séquencés aléatoirement puis réassemblés (Venter et al., 1998). Venter *et al* ont utilisé cette technique lors du séquençage du génome humain.
- L'ordonnancement hiérarchique, consiste à classer les fragments d'ADN clonés dans chromosomes artificiels bactériens(ou BAC) en fonction des chromosomes d'origine avant de les séquencer, constituant ainsi une carte physique des clones(Osoegawa et al., 2001). Lorsque deux clones partagent un nombre significatif de fragments de restriction, on peut supposer qu'ils se chevauchent et forment une région appelée « contig ». Cette méthode a été utilisée avec succès pour le séquençage des génomes de *Caenorhabditis elegans* et de *Mycobacterium tuberculosis*. C'est cette stratégie qui a été mis en œuvre par l'IHGSC lors du Projet Génome Humain(Lander et al., 2001).

Les séquenceurs sont capables de traiter des fragments d'une longueur approximative de 700 nucléotides à chaque cycle de mesure ou « run ». De 1995 à 2005, ces deux méthodes de séquençage ont été les plus utilisées dans l'exploration des génomes.

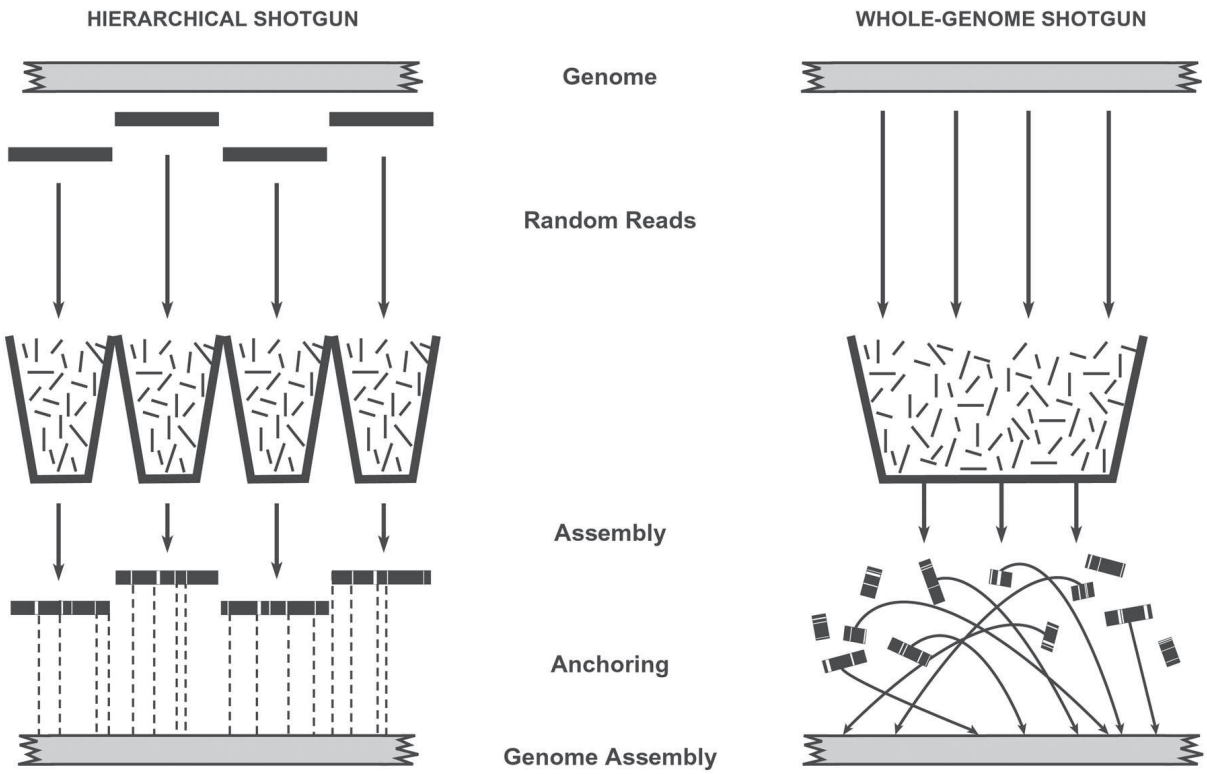


Figure 6 – Comparaison des deux approches de séquençage du génome humain. Figure issue de (Waterston et al., 2002)

I-2-4 – Séquençage de nouvelle génération (NGS)

De nouvelles technologies qui ne sont pas issues de la méthode de Sanger ont succédé aux techniques de séquençage classiques : ce sont les Next-Generation Sequencing ou NGS(Schuster, 2008). Les séquenceurs à très haut débit utilisent des fragments plus petits (25 à 35 nt) que les méthodes précédentes, mais peuvent séquencer plusieurs centaines de milliers de ces séquences en quelques jours pour un coût inférieur.

Une des techniques retrouvées dans ces types de séquenceurs est le pyroséquençage (Ronaghi et al., 1998). Elle est plus rapide et moins onéreuse qu'un séquençage par la méthode de Sanger car elle ne nécessite pas l'utilisation de clones, et la lecture de la séquence après séquençage est directe. Contrairement à la méthode de Sanger, où l'on mélange tous les éléments nécessaires ensemble, l'ajout des nucléotides se fait l'un après l'autre. L'incorporation du bon nucléotide par la DNA polymerase provoque la libération d'un pyrophosphate inorganique, lui-même étant ensuite impliqué dans une chaine de réactions enzymatiques induisant l'émission d'un signal lumineux. Le signal sera alors traduit par un capteur CCD sous forme de pics dans un pyrogramme. La hauteur des pics étant

proportionnelle au nombre de nucléotides incorporés au même moment, on peut déduire immédiatement la séquence à partir du pyrogramme.

Ces méthodes n'ont pas uniquement révolutionné le domaine du séquençage des génomes, mais ont aussi offert de nouvelles perspectives(Marguerat et al., 2008) :

- le reséquençage massif des génomes afin d'identifier les variations tels que les SNPs, les indels, ou même les CNV(copy-number variant) dont le polymorphisme serait aussi important que ceux des SNPs.
- Le séquençage à haut débit RNA-SEQ, qui permet d'évaluer l'expression des gènes par le séquençage des ADNc (Wang et al., 2009)
- Le séquençage Chip-SEQ, pour la cartographie des sites de fixation des protéines sur l'ADN et des modifications épigénétiques de la chromatine,
- Ou encore l'identification des petits ARN, sRNA-SEQ.

Toutes ces applications du séquençage à haut débit sont autant de support pour l'analyse globale des génomes individuels dont les variations en termes de séquence et d'expression, permettraient l'identification de gènes impliqués dans des pathologies. Cette discipline émergente est appelée Genome Wide Association Studies (GWAS). Une première et remarquable GWAS a été conduite lors de la recherche de variants de diabète de type II(Sladek et al., 2007) révélant notamment l'implication des gènes TCF7L2 et SLC30A8 dans cette maladie. Dès lors, l'étude à grande échelle du polymorphisme dans les génomes a permis d'identifier avec succès l'association de gènes avec des maladies telles que les diabètes de type I et II, la maladie de Crohn, ou l'hypertension.

I-2-4 – Observations préliminaires des génomes

L'analyse du génome humain a permis de corroborer plusieurs observations déjà mises en évidence(Lander et al., 2001; Gregory, 2005):

- On peut constater une hétérogénéité du génome à divers niveaux : la répartition des gènes, pourcentage en paires de nucléotides GC/AT, îlots CpG, ou le taux de recombinaison.

- 95% du génome humain, soit une grande majorité, ne code pour aucune protéine.

- De même 50% du génome est constitué de séquences répétées de toutes sortes (figure 7). Ils sont une source importante d'accroissement des génomes, bien que leur distribution soit très hétérogène d'une espèce à une autre.

Les séquences répétées peuvent être retrouvées dans tous les génomes ou types de séquences, nucléotidiques ou protéiques, qui sont ainsi des candidats de choix pour mener des études de génomique comparative.

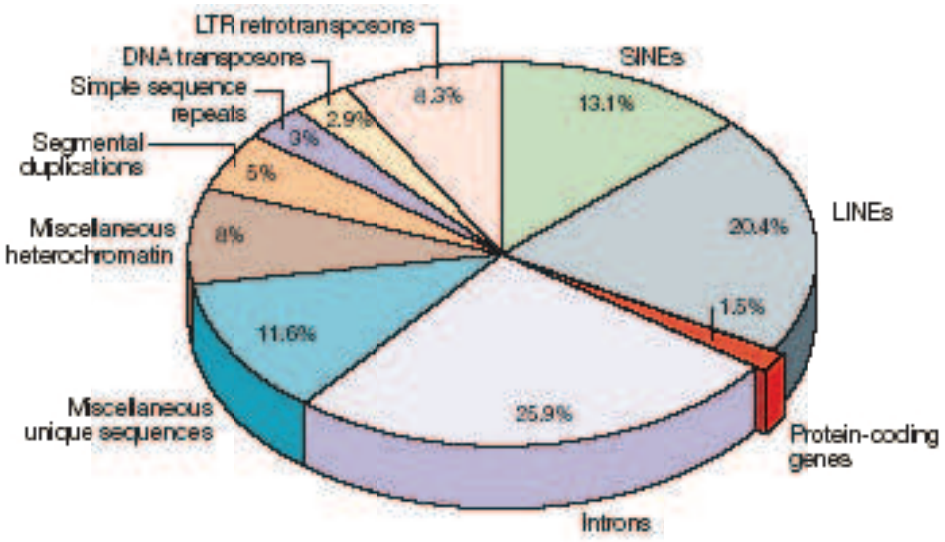


Figure 7 – Distribution générale des différents composants du génome humain. Les 20000- 25000 gènes suspectés représentent moins de 1,5% du génome, tandis qu'une grande majorité est constituée de séquences non-codantes tels que les introns (26%) et les éléments transposables (environ 45%). Figure adaptée de (Gregory, 2005)

I-3 – Les différents types de répétitions dans les génomes

Les séquences répétées peuvent être divisées en deux classes : les répétitions dispersées et les répétitions en tandem.

I-3-1- Répétitions dispersées

Appelées aussi « éléments transposables », elles sont retrouvées dans tous les génomes eucaryotes. Ces séquences sont issues d'un mécanisme appelé transposition où un fragment

du génome va pouvoir se déplacer et se multiplier. Elles ont été décrites la première fois par Barbara McClintock lors de ces études sur le maïs pour lesquelles elle reçut le prix Nobel en 1983. Chez le blé, elles constituent environ 68% du génome, tandis que les séquences répétées en constituent 90%(Li et al., 2004). Leur distribution et les proportions des différents types sont hautement variables d'un génome à un autre(Gregory, 2005). Il existe deux catégories d'éléments transposables : Les transposons et les rétrotransposons (figure 8). Les rétrotransposons représentent environ 42 % du génome humain(Lander et al., 2001).

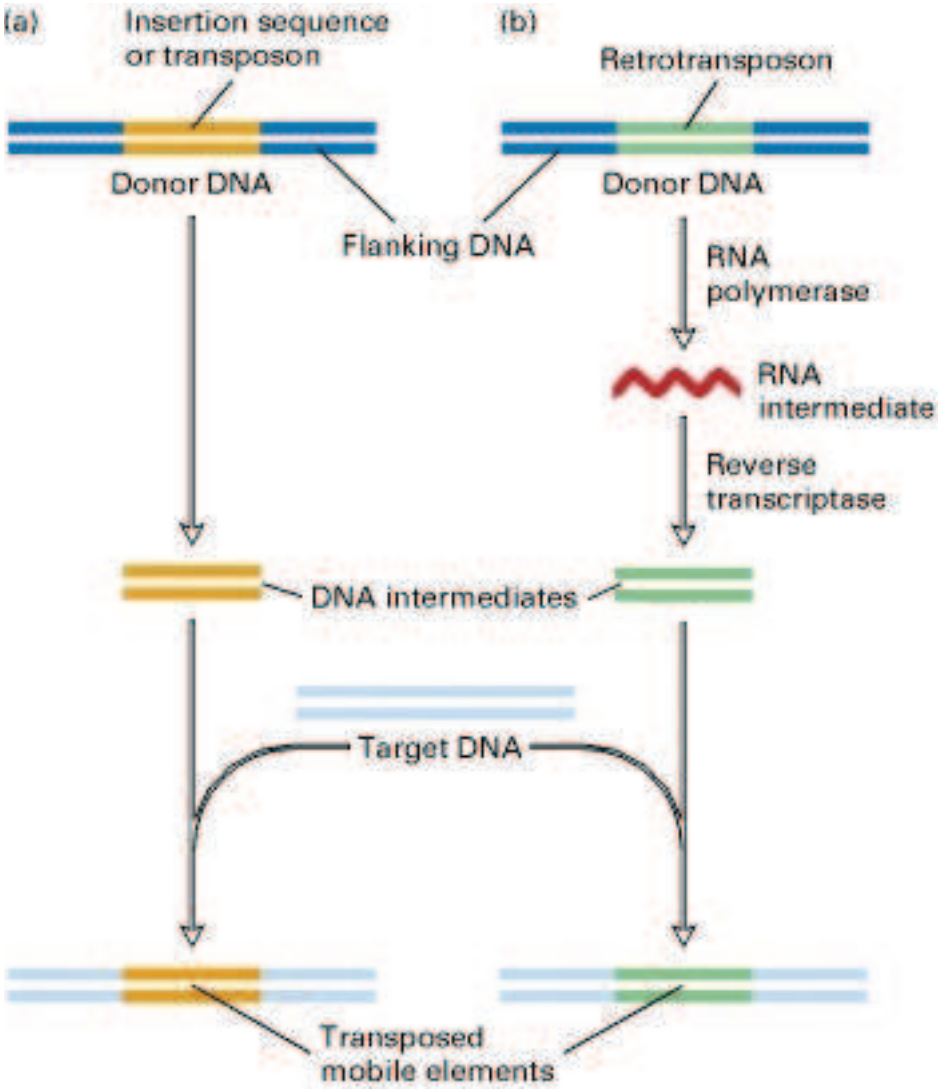


Figure 8 – Les deux classes de répétitions dispersées. (a) Les transposons(en jaune) utilisent des intermédiaires désoxyribonucléiques issus d'excision ou de la réplication de l'ADN donneur lors de la transposition. (b) les rétrotransposons (en vert) sont préalablement transcrits en un ARN qui est lui-même retranscrit en ADN, ce dernier sera réintégré ailleurs dans le génome. Figure adaptée de (Lodish et al., 2000)

- Les transposons :

  Ces éléments ont deux mode de propagation : soit via une excision du motif d'origine (mécanisme de « couper-coller » par le biais d'une enzyme, la transposase, la transposition est conservative) soit de la réplication de ce motif (la transposition est réplicative). Ces éléments transposables sont aussi retrouvés chez les procaryotes.
- Les rétrotransposons : Eléments dont le mode de propagation est proche des rétrovirus, bien qu'ils n'aient pas la capacité infectieuse du fait de l'absence d'enveloppe glycoprotéique. En effet, leur duplication va nécessiter le passage par un ARN intermédiaire qui servira de matrice à la reverse transcriptase et pourra être réintégrée au génome par une intégrase.

  Ils peuvent être classifiés en deux types :

  Les éléments à LTR (« Long Terminal Repeats »), encadrés par de longues répétitions. Les types les plus connus sont les *Ty1-copia* et les *Ty3-gypsy*. Ils encodent eux-mêmes pour leur propre machinerie de transposition.

  Les éléments non LTR

  SINEs (Short interspersed repeated sequences), leur taille varie entre 130 et 500 pb

  LINEs (Long interspersed repeated sequences), leur taille est comprise entre 6000 et 7000pb.

Le rôle des répétitions dispersées est encore peu connu, elles ont longtemps été considérés comme de l'ADN « égoïste », qui ne confère aucun avantage à la cellule(Orgel et al., 1980). Néanmoins, plusieurs études récentes ont pu montrer qu'elles pourraient être impliquées dans l'évolution des génomes (Nowacki et al., 2009).

Lorsqu'on retrouve des séquences dispersées avec un degré de similarité supérieur à 95%, on peut les assimiler à des duplications segmentaires ou encore « Low-Copy Repeats » (LCR)(Cheung et al., 2003). Elles seraient à l'origine des variations du nombre de copies d'un gène entre deux individus d'une même espèce, que l'on appelle aussi « copy number variants » (CNV). Les CNV sont souvent décrits dans des syndromes issus d'anomalies chromosomiques (Wain et al., 2009).

En génétique des populations, elles ont largement été utilisées en tant que marqueurs moléculaires, notamment les séquences ALU chez l'homme, qui sont des SINEs présentant un site de restriction pour l'enzyme Alu I.

I-3-2- Répétitions en tandem

Les répétitions en tandem sont des motifs similaires répétés de façon juxtaposées dans la séquence. Elles sont retrouvées chez toutes les espèces eucaryotes, procaryotes et virus, bien qu'elles soient extrêmement diversifiées et n'aient pas d'apparentes homologies sauf pour les espèces très proches.

I-3-2-a les répétitions en tandem dans l'ADN

Les séquences répétées peuvent être classées en trois grandes familles en fonction de leur longueur : les satellites, les minisatellites et les microsatellites. L'ADN satellite correspond à des structures dont la longueur totale peut atteindre plusieurs Mb. Viennent ensuite les minisatellites, dont le motif unitaire varie de 10 pb à 100pb, formant des séquences répétées longues de 10^2 à 10^5 pb. La dernière classe, appelée microsatellites, regroupe de courts fragments de 1 à 10 pb, pouvant atteindre au total une centaine de pb. Microsatellites et minisatellites appartiennent à la famille des VNTR (« Variable Number of Tandem Repeat »), du fait de leur variabilité en longueur d'une espèce à l'autre. Ils ne sont pas distingués chez les génomes procaryotes ou viraux, on préférera utiliser la définition suivante : SSR pour « simple sequence repeat » et STR pour « short tandem repeat », équivalents à des microsatellites chez les eucaryotes. Dans l'article de Lander *et al*, les microsatellites et les minisatellites couvrent environ 3% du génome humain, tandis que l'ADN satellite atteindrait 5 à 6%.

I-3-2-a-i ADN satellite

Cet ADN hautement répétitif est localisé principalement dans l'hétérochromatine au niveau des régions péricentromériques et des télomères. Initialement considéré comme de l' ADN « poubelle » («junk DNA») (Ohno, 1972), il semble jouer un rôle primordial dans la structuration des génomes : De l'organisation des chromosomes lors de la division cellulaire (Csink and Henikoff, 1998; Henikoff et al., 2001) au processus de spéciation de la cellule (Malik and Henikoff, 2001). Les séquences centromériques, régions les plus tardivement répliquées, permettraient également d'assurer la fin de la réplication. En effet, l'ADN centromérique doit être dépourvu d'origines de réplication afin que le processus de réplication termine complètement son cycle. Sa nature répétitive serait ainsi un moyen de limiter le nombre d'origines de réplication dans cette région (Csink and Henikoff, 1998).

I-3-2-a-ii Minisatellites

Retrouvées dans la plupart des génomes eucaryotes, ces répétitions sont hautement polymorphes. Les unités répétées différentes du motif de base sont appelées les variants, et servent de base pour les analyses génétiques comme le « DNA fingerprinting » (Jeffreys et al., 1988) . Cette technique a pu trouver bon nombre d'applications en tant que marqueurs pour des liaisons génétiques ou encore les tests de paternité. Le haut polymorphisme de ces séquences leur confère une instabilité susceptible de perturber l'expression de gènes voisins. Par exemple, il a été montré que la variabilité des minisatellites avaient une influence dans la prédisposition à certains cancers et le diabète de type 1(IDDM2) (Bennett et al., 1995; Ninomiya et al., 2006).

I-3-2-a-iii Microsatellites

Communément appelés SSR ou STR, ils abondent dans les génomes eucaryotes et procaryotes, principalement dans les régions non codantes(Ellegren, 2000)et sont en majorité di-nucléotidiques dans le génome humain(environ 0,5% du génome, représentés à 50 % de motifs (CA)n et 35 % de (AT)n)(figure 9). Le niveau de mutations des microsatellites est de plusieurs ordres de magnitude supérieur au niveau de mutations global des séquences d' ADN(10^3 – 10^6 par locus et par génération)(Weber and Wong, 1993; Ellegren, 2000). On peut les retrouver également dans les régions codantes, que ce soit dans les gènes ou les «expressed sequence tags» (EST) (Morgante et al., 2002). A la différence des régions intergéniques où ils sont le plus souvent di ou tetra-nucléotidiques, les microsatellites à l'intérieur des régions codantes sont majoritairement à base de trinucléotides. Ils sembleraient subir une pression sélective afin de ne pas décaler le cadre de lecture lors de la transcription (Metzgar et al., 2000). Dans les régions géniques codant pour une protéine, ils pourront être assimilés comme source de répétitions en tandem dans les séquences protéiques.

Table 14 SSR content of the human genome

Length of repeat unit	Average bases per Mb	Average number of SSR elements per Mb
1	1,660	36.7
2	5,046	43.1
3	1,013	11.8
4	3,383	32.5
5	2,686	17.6
6	1,376	15.2
7	906	8.4
8	1,139	11.1
9	900	8.6
10	1,576	8.6
11	770	8.7

SSRs were identified by using the computer program Tandem Repeat Finder with the following parameters: match score 2, mismatch score 3, indel 5, minimum alignment 50, maximum repeat length 500, minimum repeat length 1.

Figure 9 – Distribution des longueurs du motif unitaire dans les microsatellites. Figure extraite de Lander et al(Lander et al., 2001)

I-3-2-a-iv Autres répétitions nucléotidiques

Parmi les modifications post-transcriptionnelles que subit un ARN lors de sa phase de maturation, il en existe une, la polyadénylation, qui consiste en l'adjonction d'environ 200 adénosines à son extrémité 3'. Cette queue poly-A est importante dans la stabilité de l'ARN et dans son transport du noyau vers le cytoplasme, ou encore dans le recrutement par le ribosome pour la traduction, bien qu'elle-même ne soit jamais traduite.

I-3-2-b les répétitions en tandem dans les séquences protéiques

Au niveau protéique, elles sont retrouvées dans plus de 14% des séquences connues, et sont retrouvées dans 1/3 des protéines humaines ayant des fonctions fondamentales(Marcotte et al., 1999). Comme chez les microsatellites, le niveau de mutation des répétitions protéiques est nettement supérieur à celui des séquences non répétées, leur dynamique évolutive étant intimement liée aux mécanismes évolutifs proposés précédemment(Ellegren, 2000).
Une séquence protéique répétée en tandem est définie comme la succession d'au moins deux motifs d'acides aminés similaires ou « copies». La longueur du motif peut varier de 1 résidu (ces répétitions sont appelées « homorepeats » , « homopeptides », « single amino acid

repeats », ou « amino acid runs ») à des domaines homologues de plus d'une centaine de résidus(Heringa, 1998; Andrade et al., 2000). Leur nombre est tout aussi variable, pouvant atteindre des centaines de copies comme c'est le cas pour le collagène.

Elles sont majoritairement retrouvées chez les eucaryotes et beaucoup plus rares chez les procaryotes, une des hypothèses expliquant cette affinité serait que les répétitions sont favorablement localisées dans les protéines dont les fonctions seraient propres aux eucaryotes (Karlin and Burge, 1996; Marcotte et al., 1999; Faux et al., 2005). De même, il a été montré que les répétitions en tandem dans les séquences protéiques étaient souvent riches en petits résidus, et/ou polaires et/ou chargés(Karlin and Burge, 1996; Marcotte et al., 1999). Cependant, il existe dans un même règne une forte variabilité de leur distribution et leur composition d'une espèce à l'autre. Par exemple, un organisme eucaryote tel que *Plasmodium falciparum,* responsable du paludisme, peut voir jusqu'à 40% de ses séquences protéiques contenir des répétitions en tandem, alors que d'autres organismes tels que *Drosophila melanogaster* ou *Caenorhabditis elegans* ne présenteront qu'environ 27% et 13% respectivement(Alba et al., 2007).

Il est à noter qu'on retrouve des homorepeats rajoutés sur les chaines latérales de certains résidus lors des modifications post- traductionnelles. Par exemple, la polyglutamylation (ajout de plusieurs résidus E), la polyglycylation (ajout de plusieurs résidus G) affectent la stabilité, la localisation et l'interaction protéine-protéine de la tubuline, constituant élémentaire des microtubules (Janke et al., 2005).

I-4 – Origine et évolution des répétitions en tandem

Lors de la réplication, les séquences d'ADN sont recopiées à l'identique. Néanmoins, l'ADN polymérase, enzyme responsable de la synthèse des nouveaux brins, peut faire des erreurs qui sont autant d'événements susceptibles de modifier la séquence d'origine ; ce sont les mutations. Des mutations peuvent aussi apparaître suite à des agressions physiques ou chimiques, telles que l'exposition aux ultraviolets ou aux rayons X, certains solvants et dérivés de benzène. Comme toutes les séquences dans l'ADN, les séquences répétées en tandem sont sujettes à plusieurs types d'événements mutationnels :

- Substitution : remplace un nucléotide par un autre, elle peut être de transition (le type de base est conservé) ou de transversion (remplace une purine en pyrimidine ou inversement).
- Insertion : insère dans la séquence un ou plusieurs nucléotides,
- Délétion : supprime un ou plusieurs nucléotides,
- Inversion : change le sens d'un fragment d'ADN de 5' ->3' vers 3'->5' et inversement.

Si ces mutations surviennent dans les régions codantes, elles pourront avoir un impact direct sur la séquence des protéines qui en sont issues. Dans les cas des substitutions, on dit de la mutation qu'elle est « silencieuse » lorsqu'elle n'affecte pas la séquence protéique, du fait de la redondance du code génétique. Si le nouveau nucléotide entraine une modification de l'acide aminé codé, on parlera de mutation « faux sens ». Lorsqu'elle change le codon d'origine en un codon stop, elle est une mutation « non sens », aboutissant à la production d'une protéine tronquée et pouvant affecter significativement sa fonction. De même, les mutations d'insertion et de délétion de nucléotides dont le nombre n'est pas un multiple de trois vont provoquer un décalage du cadre de lecture. Ce décalage peut là aussi avoir des conséquences dramatiques sur la fonction de la protéine produite.

Les mutations dues à des erreurs de réplication apparaissent à un ratio de 10^{-9} nucléotides, correspondant sensiblement au taux d'erreur lors de l'incorporation d'un nucléotide par l'ADN polymérase (ce ratio, initialement plus élevé, est abaissé grâce au mécanisme de correction de la polymérase).

Outre ces mutations, il existe d'autres événements qui sont propres aux séquences répétées en tandem, et sont à l'origine de leur évolution :

- L'expansion qui est un gain de un ou plusieurs motifs dans la séquence répétée,
- La contraction correspondant à une perte de motif.

Durant les dernières décennies, plusieurs modèles ont été construits afin d'expliquer les mécanismes responsables de ces événements mutationnels.

Le modèle le plus populaire a été établi par Levinson et Gutman (Levinson and Gutman, 1987), il soutient l'hypothèse que ces phénomènes de contraction - expansion sont dus à des glissements de l'ADN polymérase(« replication slippage ») suite à un mauvais appariement des brins d' ADN lors de la réplication (« slipped strand mispairing », figure 10). Si ce modèle peut expliquer l'expansion d'un même motif à l'identique, il trouve ces limites quand

à l'explication des différentes mutations portées par chacun de ces motifs. Le modèle de Levinson prévoit que ces variations de motifs sont dues à des événements mutationnels aléatoires. Cette hypothèse peut être plausible dans le cas de répétitions courtes de quelques pb (Messier et al., 1996), mais elle parait peu applicable pour des répétitions plus longues.

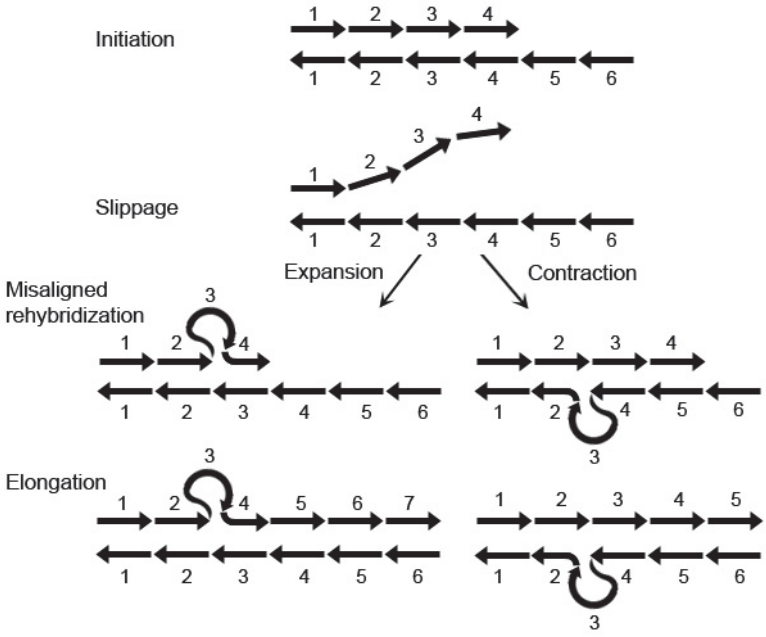


Figure 10 – Glissement de l'ADN polymérase et mésappariement des brins d'ADN.

Un second modèle propose que leur évolution serait aussi dépendante d'un mécanisme de recombinaison : le crossing-over inégal (Haber and Louis, 1998; Richard and Paques, 2000). Ce phénomène survenant lors de la méiose chez les eucaryotes (et d'une manière plus hypothétique lors de la réparation de l'ADN chez les bactéries) consiste en un échange non réciproque de fragments d'ADN entre deux chromosomes homologues ou deux chromatides sœurs (figure 11). La présence de motifs répétés en tandem augmentent les chances de mésalignement lors du brassage chromosomique et favoriserait ce mécanisme(Graur and Li, 2000). Le crossing-over inégal permet la recombinaison de grands fragments d'ADN de plusieurs centaines de pb. Cependant, un bon nombre de biais mutationnels liés à la nature des répétitions en tandem ont pu être mis à jour, tels que l'augmentation du niveau de mutations en fonction de la taille des allèles ainsi qu'une activité accrue de phénomènes de contractions pour les allèles les plus longs(Xu et al., 2000; Huang et al., 2002). Ces observations laissent supposer que les deux mécanismes peuvent rentrer en jeu mais que leur action dépend aussi d'autres facteurs, comme la structure même des séquences répétées.

De manière simplifiée, les variations de répétitions courtes telles qu'on les trouve dans les SSR seraient engendrées par des erreurs de réplication tandis que les répétitions plus longues comme les minisatellites dépendraient de la machinerie de recombinaison(Richard and Paques, 2000).

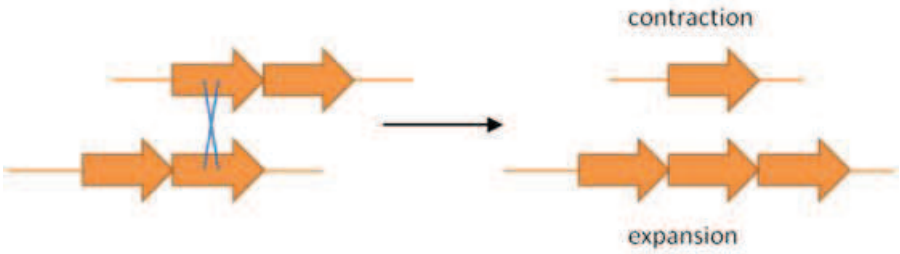


Figure 11 – Crossing-over inégal. On obtient ici une contraction de la séquence répétée sur le premier brin et une expansion dans le second brin.

Les polymorphismes générés par ces mécanismes, les événements mutationnels auxquels les répétitions sont soumises ainsi que les disparités entre les génomes font que leur identification dans les séquences protéiques est d'autant plus complexe.

I-5 – Structures et Fonctions des séquences répétées en tandem

I-5-1 Structures des séquences répétées en tandem

Les méthodes de détermination de structure, telles que la cristallographie à rayons X ou la RMN, ont pu mettre en évidence plusieurs structures tridimensionnelles. Les premières études de biologie structurale ont été menées sur les protéines fibreuses porteuses de répétitions de 2 à 7 résidus telles que le collagène, la kératine, la tropomyosine ou la fibroïne issue de la soie)(Fraser and MacRae, 1973).

Dans les années 90, sur les protéines avec des répétitions plus longues (de 5 à 50 résidus) ont suscité un gain d'intérêt suite à l'identification de structures particulières. Des structures telles que les β-hélices(Yoder et al., 1993), les β-rolls(Baumann et al., 1993), les structures en « fer à cheval » des protéines à répétitions riches en leucines dites LRR (ou « Leucine Rich Repeats »)(Kobe and Kajava, 2001), β-propellers(Fulop and Jones, 1999), et les superhélices α-hélicales(Groves and Barford, 1999) ont pu être associées aux protéines portant des répétitions en tandem(figure 12).

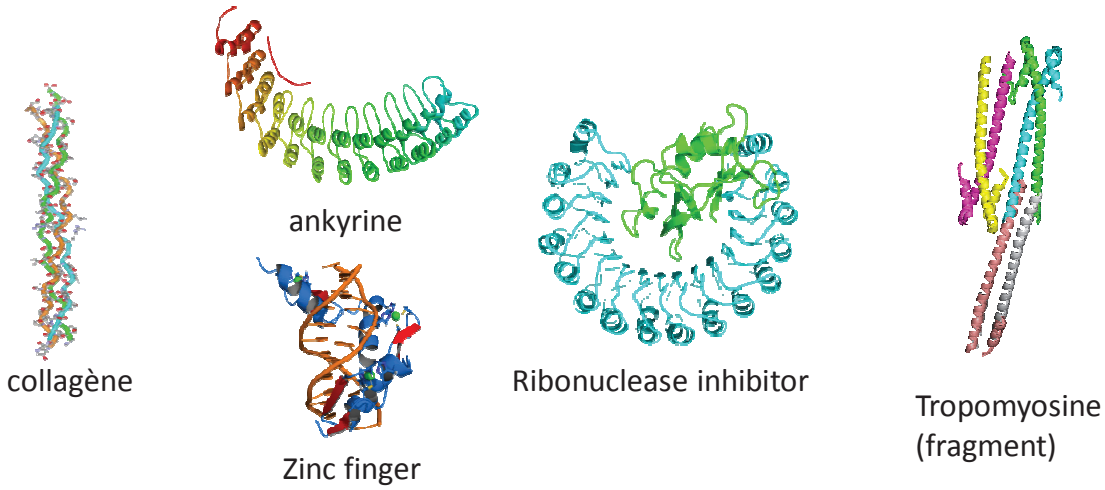


Figure 12. Exemples de structures tridimensionnelles de séquences répétées. Le collagène possède une structure fibreuse en polyproline de type II. L'ankyrine et la « ribonuclease inhibitor » adoptent une configuration solénoïdale tandis que la tropomoyosine a une conformation « α - helical coiled coil ». Les domaines « Zinc finger » appartiennent à la classe IV de protéines globulaires à domaines répétés décrites dans (Kajava, 2001).

Certaines de ces structures ont été classifiées en fonction de la longueur de leur motif unitaire (Kajava, 2001)(figure 13) :

Classe I : de 1 à deux résidus, structures cristallines. Les structures connues sont pour la plupart issues de peptides synthétiques(Kajava, 2001; Jorda et al., 2010), les peptides naturels ayant plutôt une tendance à s'autoassembler pour former des agrégats insolubles(Fraser and MacRae, 1973) toxiques pour la cellule. Les homorepeats ou homopeptides, qui ont des caractéristiques physicochimiques inhabituelles du fait de la répétition d'un même résidu, appartiennent à cette classe. Ces homopeptides peuvent adopter des structures différentes selon la nature du résidu répété. Par exemple, les peptides poly-A ou poly-K peuvent s'auto-assembler pour former des structures cristallines à base d'hélices α ou de structures β(Blout, 1962; Auer and Doty, 1966) tandis que d'autres peptides seront constitués exclusivement de structures α-hélicales ou β. Les homopeptides polyglutaminiques (polyQ) sont impliqués dans bon nombre de maladies neurodégénératives du fait de leur capacité à s'agréger sous forme de fibrilles amyloïdes(Orr and Zoghbi, 2007). Les structures de cette classe trouvent aussi leur utilité en science des biomatériaux(Tormo et al., 1992; Altman et al., 2003).

Classe II : structures fibreuses avec motifs de 3 à 4 résidus où chaque motif correspond à un tour de la structure secondaire. Ces types de structures sont principalement sous forme multimérique, la structure étant stabilisée par les interactions interchaines. Un exemple de ce type de structure est retrouvé dans le collagène, protéine la plus abondante du corps humain et composant de la matrice extracellulaire. Le motif est du type G-X-Y où X et Y peuvent être n'importe quel résidu, le plus souvent des prolines ou des hydroxyprolines. La conformation finale, appelée polyproline de type II, est un assemblage de trois chaines en hélice. Cette même classe regroupe les structures du type « α - helical coiled coil », des super enroulements de 2 à 7 hélices α que l'on retrouve dans plusieurs protéines de fonction biologiques primordiales (par exemple, les facteurs de transcription c-fos et jun)(Glover and Harrison, 1995).

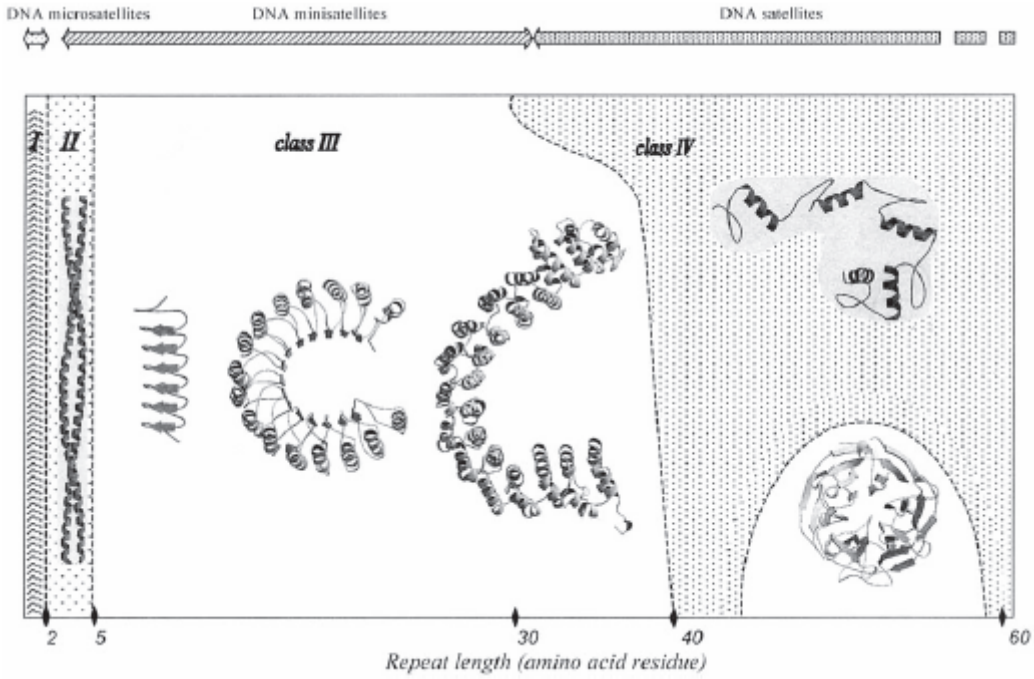


Figure 13. Classification de structures associées aux séquences répétées en tandem en fonction de la longueur du motif. Figure extraite de (Kajava, 2001).

Classe III : On y distingue des structures appelées solénoïdes qui sont constituées d'unités structurelles contenant 2 à 4 segments de structure secondaire. Les segments peuvent être d'un même type ou mixtes (hélice α, brin β ou hélice 3/10)(Groves and Barford, 1999; Kobe and Kajava, 2000; Kobe and Kajava, 2001). Ces unités structurelles ont une conformation telle que la chaine polypeptidique va former une super-hélice. Ce type de structure peut présenter une unité structurelle courte (4 à une dizaine de résidus, par exemple dans l' UDP-N-acetylglucosamine acyltransferase chez *Escherichia coli* (Raetz and

Roderick, 1995)) comme plus longue (de 15 à 40 résidus), où elle correspond le plus souvent à un tour de la super-hélice. Elles incluent des structures en fer à cheval retrouvées chez les LRR (structures α/ β, figure 13)(Kobe and Kajava, 2001), des structures droites constituées généralement de segments β solénoïdaux, et des structures droites ou semi incurvées α – solénoïdales comme l'ankyrine (figure 12) ou la clathrine(Ybe et al., 1999) .

Classe IV : Elle regroupe les répétitions dont le motif unitaire est supérieur à 30-40 résidus. L'unité structurelle est suffisamment grande pour se replier dans une conformation stable et constituer un domaine à elle seule. C'est le cas des domaines « Zinc finger »(Lee et al., 1989), dont le rôle est de se lier aux acides nucléiques et que l'on retrouve dans bon nombre de protéines impliquées dans la régulation d'expression des gènes. En marge de ce type de structure souvent globulaire, des exceptions existent, comme par exemple les conformations dites en « β-propeller »(Fulop and Jones, 1999). Parmi elles on trouve les « WD repeats » (souvent terminés par un dipeptide tryptophane – aspartate) longs d'environ 40 résidus, qui ne sont ni solénoïdaux, ni globulaire. Chaque motif structurel est un feuillet β comprenant 4 brins antiparallèles légèrement torsadés, de sorte que le premier brin et le quatrième sont chacun dans un plan perpendiculaire à l'autre. Les WD repeats dont le nombre peut varier de 4 à 8 sont organisés de façon circulaire(Neer et al., 1994; Li and Roberts, 2001).

I-5-2 Fonctions biologiques des protéines portant des séquences répétées en tandem

Les protéines porteuses de répétitions en tandem sont impliquées dans une grande variété de fonctions cellulaires fondamentales telles que la régulation de l'expression des gènes, des fonctions structurelles, la signalisation, la régulation de voies de signalisation ou encore le transport. Par exemple, les WD repeats sont retrouvés dans une multitude de fonctions cellulaires fondamentales, du contrôle du cycle cellulaire à l'apoptose (notamment dans les protéines régulatrices du cycle cellulaire telles que *CDC4*, *CDC20*, *CDC40*, *p55*, *Mad2)*(Li and Roberts, 2001). Elles tendent, notamment chez les eucaryotes, à être retrouvées dans les facteurs de transcriptions et les grands complexes protéiques comme les ribosomes (Karlin and Burge, 1996; Alba and Guigo, 2004; Faux et al., 2005). Elles peuvent jouer un rôle structurel : l'exemple le plus significatif étant le collagène, qui représente près de 30% de la masse protéique chez l'humain, et participe à l'organisation de la matrice extracellulaire où sa structure fibreuse permet aux tissus de résister aux contraintes mécaniques. L'élastine, une autre protéine de la matrice extracellulaire porteuse de répétitions, confère aux tissus des

propriétés élastiques. Bien que sa structure à ce jour n'ait pu être déterminée, il a pu être mis en évidence une structure particulière dans la protéine: la desmosine. Celle-ci résulte de l'association covalente des chaines latérales de 4 résidus lysine, elle est responsable des vertus élastiques de l'élastine. De récentes études ont pu attribuer de nouvelles implications, comme dans les interactions protéines-protéines(Shimohata et al., 2000; Buchanan et al., 2004) où les répétitions joueraient un rôle de modulateur. De même, certaines protéines porteuse de motifs structurels répétés ont une fonction de reconnaissance protéine-protéine : les LRR se lient à bon nombre de protéines globulaires, la plus connue étant l'inhibiteur de la ribonucléase(RI)(Kobe and Deisenhofer, 1993; Kobe and Kajava, 2001). Un autre exemple sont les « α - helical coiled coils » qui ont un rôle structurel et médient l'oligomérisation de sous-unités protéiques (Parry et al., 2008). Il a aussi été montré que les variations d'une séquence répétée dans le facteur de transcription *Runx-2* (constituée d'une répétition polyQ suivie d'une répétition polyA) avait une influence sur la forme du crâne des chiens dont les lignées sont issues de races pures, notamment sur la longueur et la clinorhynchie (la courbure du museau) (Fondon and Garner, 2004)(figure 14). Ces subtiles évolutions morphologiques sont ainsi dépendantes de l'évolution de la séquence répétée et sont très rapides.

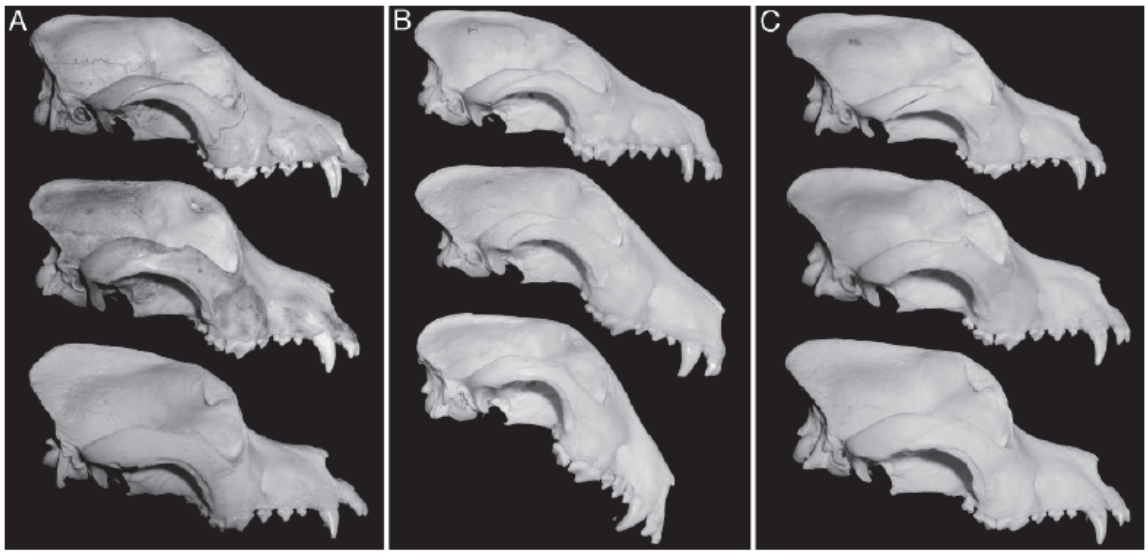


Figure 14. Evolution morphologique de cranes canins. (A)Cranes de Saint-bernard de 1850, 1921 et 1967. (B) Cranes de Bull terrier de 1931, 1950 et 1976. (C) Cranes de Terreneuve de 1926,1964 et 1971. L'analyse des répétitions de Runx-2 chez le bullterier de 1931 montre un allèle intermédiaire($Q_{19}A_{14}$) plus présent que le bull terrier moderne ($Q_{19}A_{13}$). Figure extraite de (Fondon and Garner, 2004).

D'autres exemples ont pu montrer que l'expansion de répétitions poly-A (Albrecht et al., 2004) ou la suppression de ces répétitions(Anan et al., 2007) pouvaient conduire à des changements phénotypiques plus drastiques, démontrant que ces régions répétées avaient bien une fonction déterminante dans l'évolution des espèces.

Cependant, le nombre de fonctions associées aux répétitions reste relativement faible en regard de leur abondance dans les séquences. Beaucoup de répétitions dont le rôle est méconnu sont souvent considérées comme neutres, ou des « linkers » entre des domaines fonctionnels (Huntley and Golding, 2000).

I-6 – Pathologies associées aux séquences répétées en tandem

I-6-1 – Généralités

Outre leurs propriétés exceptionnelles en tant que marqueurs génotypiques, les séquences répétées en tandem et leurs polymorphismes ont pu être associé durant les dernières décennies à plusieurs pathologies. Pendant 20 ans, la mise en évidence de ces corrélations ont été menées via des techniques s'appuyant sur la Bioinformatique. On retrouve parmi elles le déséquilibre de liaison (« Linkage Disequilibrium» ou LD) ou le clonage positionnel (« positional cloning »). La première permet de mesurer l'association de deux ou plusieurs allèles situés sur des loci différents et qui apparaissent dans une population à des fréquences élevées, mais qui ne sont pas nécessairement sur le même chromosome (Horikawa et al., 2000; Pritchard and Przeworski, 2001). On utilise généralement cette mesure lors de comparaisons entre deux types de populations : une affectée par une pathologie, et une autre saine, et ainsi mettre en évidence des fréquences anormales de certains allèles susceptibles d'être impliqués dans la pathologie. La seconde consiste en la recherche de la position d'une séquence impliquée dans un phénotype (dans notre cas, une pathologie).

Les « Genome Wide Association Studies », discipline récente, permettent de réaliser ces tests de corrélations à l'échelle de génomes et d'analyser les variations de génomes d'une population malade comparée à celle d'une population saine. Ces variations sont généralement basées sur l'étude des variations des SNPs et plus récemment sur les CNVs, tout deux sources de polymorphismes, mais aussi sur les VNTR lors de MLVA (« Multiple Loci VNTR Analysis »). Un des tests les plus connus est celui de Cochran-Armitage (Cochran, 1954;

Armitage, 1955), qui permet de mesurer la corrélation entre une variable dichotomique (typiquement « malade » et « contrôle ») et une variable à plusieurs caractères. Ce test est implémenté dans un programme du nom de PLINK (Purcell et al., 2007), la référence actuelle en termes d'outil pour les GWAS.

Toutes ces méthodes sont capables de mettre en exergue les polymorphismes issus des répétitions en tandem et d'identifier de nouvelles implications dans des pathologies ; A ce jour, on compte une grande variété de maladies qui leur sont associées, exogènes ou endogènes.

I-6-2 – Maladies neurodégénératives et développementales

Elles sont souvent liées à des expansions anormales de répétitions trinucléotidiques. La première pathologie dont l'association ait été mise en évidence est le syndrome de l'X fragile (FRAXA). Cette maladie est une affection liée à l' X se manifestant par un retard mental plus ou moins sévère, et la deuxième maladie d'origine génétique après le syndrome de Down (trisomie 21). Elle est due à une expansion excessive de triplets CGG portés par le premier intron du gène FMR1 (Knight et al., 1992; Jin and Warren, 2000). Cette région est hyperméthylée chez les individus porteurs de la maladie, induisant une répression de l'expression du gène FMR1 et par conséquence un retard mental. Comme l'indique la Table 1, bon nombre de répétitions dans l'ADN non codant sont associées à ces types de pathologies. Une dérégulation des protéines qui se fixent habituellement sur ces régions (« RNA binding proteins ») et qui contrôlent la maturation de l'ARNm vont dans la plupart des cas mener à un dysfonctionnement dans l'expression des protéines incriminées. D'autres répétitions seront retrouvées dans l'ADN codant et par conséquence dans les séquences protéiques sous forme de répétitions d'acides aminés.

Pathologie	Type de répétition	Gène impliqué
Syndrome de l'X fragile	(CGG)n	FMR1
Retard mental lié à FRAXE	(CGG)n	FRAXE
Syndrome de Jacobsen	(CGG)n	CBL2
Ataxie de Friedreich	(GAA)n	FRAXA
Dystrophie myotonique de Steinert (DM1)	(CTG)n	DMPK

Table 1 – Pathologies associées à des répétitions de triplets dans les régions non codantes de gènes.

Chez l'homme, les deux types les plus fréquents dans les pathologies neurodégénératives sont les expansions de polyglutamines (polyQ) et les expansions de polyalanines (polyA). Les répétitions pathogéniques de triplets CAG sont hautement polymorphiques, en particulier chez l'homme (Andres et al, 2003), et sont retrouvées sous forme d' homorepeats poly Q dans les séquences protéiques. Souvent impliqués dans les maladies neurodégénératives, ils jouent un rôle primordial dans l'amyloïdogénèse lorsque la longueur de la répétition dépasse un seuil critique (Table 2). L'amyloïdogénèse consiste en la transformation de protéines solubles en des agrégats insolubles de forme fibrillaire. Les fibrilles amyloïdes sont des structures quaternaires qui adoptent une conformation « cross-β », similaire à celle retrouvée dans les β-solénoïdes, où les structures β sont empilées perpendiculairement au sens de l'élongation du fibrille.

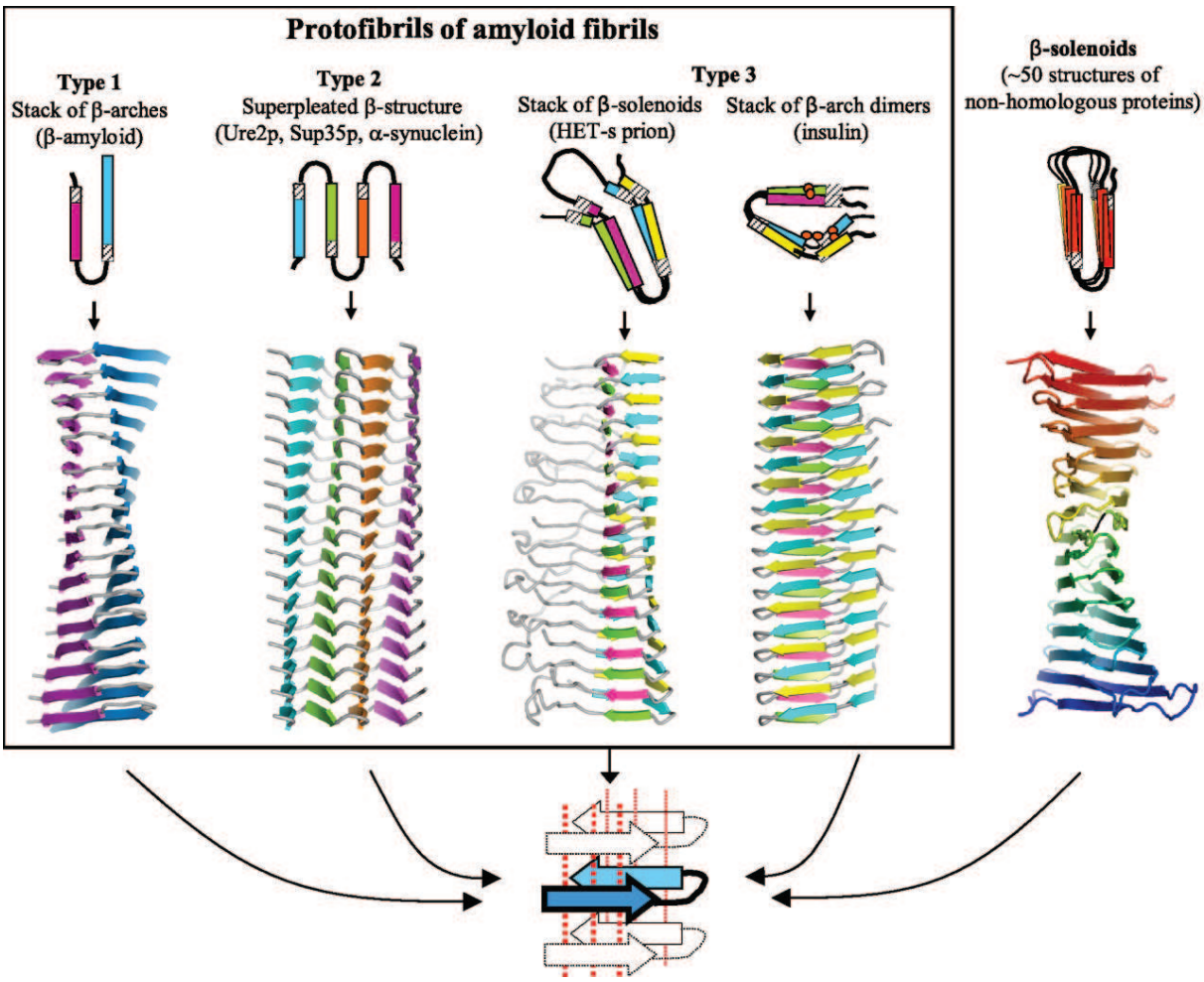


Figure 15. Différents modèles structuraux de fibrilles amyloïdes. Tous dérivent d'un seul et même motif structurel unitaire : l'arche β. Figure extraite de (Kajava et al., 2010)

Plusieurs modèles structuraux ont pu être proposés(Perutz et al., 1994; Thakur and Wetzel, 2002; Kajava et al., 2004; Baxa et al., 2006) pour les différents types de fibrilles existants, il a été démontré récemment qu'ils dérivaient tous d'un seul et même motif structurel : l'arche β (Kajava et al., 2010) (figure 15). Chacun des brins de l'arche est impliqué dans des feuillets différents, ainsi les interactions dans une même arche se font via les chaines latérales des résidus qui la composent. Les arches interagissent entre elles via des liaisons hydrogènes entre leurs chaines principales pour former une structure fibrillaire stable. Les agrégats vont, en général, induire non pas une perte de fonction mais un gain de toxicité pour la cellule, responsable de la maladie(Bauer and Nukina, 2009).

Généralement, la sévérité de la maladie est corrélée avec la longueur des répétitions, comme dans le cas des maladies à polyglutamines (polyQ). Cependant d'autres répétitions, telles que les polyA, ont un comportement différent (Brown and Brown, 2004). Les répétitions polyQ tendent à être plus polymorphiques et atteindre des longueurs plus grandes que les polyA. De même le seuil critique des polyA est nettement inférieur à celui des polyQ. Enfin, les polyA conduisent plutôt à une perte de fonction partielle ou totale de la protéine tandis que les pathologies associées aux polyQ sont déclenchées par une accumulation en agrégats toxiques dans la cellule(Gatchel and Zoghbi, 2005). On compte plus d'une dizaine de pathologies neurodégénératives issues d'expansion aberrante de polyQ (Table 2), comme l'amyotrophie spinobulbaire (SBMA ou « Kennedy disease »), la chorée de Huntington (HD), l'atrophie dentatorubrale et pallidoluysienne (DRPLA), 6 types d'ataxie spinocérebelleuse (SCA1, SCA2, SCA3 ou « « Machado-Joseph disease », SCA6, SCA7, SCA17) ou encore l'épilepsie.

Les répétitions polyA seraient quant à elle corrélées à des pathologies d'ordre développemental, tels que la synpolydactylie ou la dystrophie musculaire oculopharyngée (OPMD) (Table 3).

Pathologie	Type de répétition	Gène incriminé (protéine produite)	Fonction	Longueur normale	Longueur pathologique
SCA1	(CAG)n	SCA1(ataxine 1)	Transcription	6–39	40–82
SCA2	(CAG)n	SCA2 (ataxine 2)	métaboisme de l'ARN	15–24	32–200
SCA3(MJD)	(CAG)n	SCA3(ataxine 3)	Activité de dé-ubiquitinylation	13–36	61–84
SCA6	(CAG)n	CACNA1A	sous unité de canal à calcium	4–20	20–29
SCA7	(CAG)n	SCA7(ataxine 7)	Transcription	4–35	37–306
SCA17	(CAG)n	SCA17(TBP)		25–42	47–63
DRPLA	(CAG)n	DRPLA(atrophine 1)	Transcription	7–34	49–88
SBMA	(CAG)nc	AR(récepteur androgène)	récepteur d'hormones stéroïdes	9–36	38–62
HD	(CAG)n	HD(huntigtine)	transport, transcription	11–34	40–121

Table 2. Caractéristiques des maladies à polyglutamine (polyQ). Le tableau décrit les intervalles de longueur des répétitions dans une population normale et dans une population atteinte par la pathologie (Shimohata et al., 2000; Bauer and Nukina, 2009)

Pathologie	Type de repetition	Gène incriminé	fonction	Longueur normale	Longueur pathologique
synpolydactylie	GCG	HOXD13	Facteur de transcription	15	22-29
Dysplasia cleidocraniale	GCG	RUNX2	Facteur de transcription	17	27
OPMD	GCG	PABPN1	Facteur de transcription	10	11-17
Holoprosencéphalie	GCG	ZIC2	Facteur de transcription	15	25
Syndrome génital pieds-mains	GCG	HOXA13	Facteur de transcription	18	24-26
Blepharophimosis	GCG	FOXL2	Facteur de transcription	14	22-24
Syndrome infantile de spasme	GCG	ARX	Facteur de transcription	16 12	18 ou 23 20
Retard mental lié à l' X	GCG	SOX 3	Facteur de transcription	15	26
Syndrome d'hypoventilation centrale congénitale	GCG	PMX2B	Facteur de transcription	20	25-29

Table 3. Caractéristiques des maladies à polyalanine (polyA).

D'autres répétitions, aux motifs plus longs que les homorepeats, ont pu être mises en évidence et être associées à des maladies neurodégénératives majeures. Par exemple, il existe des répétitions d'un motif de 11 résidus du type (XKTKEGVXXXX) dans l'α-synucléine, dont l'agrégation en fibrilles est en partie responsable de la maladie de Parkinson(El-Agnaf and Irvine, 2000). La fonction des répétitions dans l'alpha synucléine reste encore incomprise, mais il a été proposé que le nombre de ces répétitions influerait sur les propriétés amyloïdogéniques de la protéine. De même, on peut retrouver des répétitions au niveau de la protéine tau, incriminée dans la maladie d'Alzheimer.

I-6-3– Maladies neurodégénératives infectieuses

Les prions sont des protéines infectieuses qui via un changement de conformation sont capables de transformer la version normale de la protéine dans la cellule en version infectieuse. Une fois la forme infectieuse établie, elle pourra être transmise à toutes les cellules filles et à ses descendantes, et ce sans intervention d'acides nucléiques, rendant ce mode de transmission unique. Le prion partage des propriétés similaires à celles des fibrilles amyloïdes (Balbirnie et al., 2001), formant des plaques amyloïdes comme dans les maladies neurodégénératives précédemment décrites. Le principal agent causal chez les mammifères des maladies à prion, telles que l'encéphalopathie spongiforme bovine et la maladie de Creutzfeldt-Jakob chez l'homme(Belay, 1999) est la PrP (« proteinase resistant protein »). L'isoforme normal de PrP, appelée PrPc, est impliqué dans le développement du système nerveux chez l'embryon, et dans la survie et l'adhésion cellulaire chez l'adulte. L'isoforme pathologique PrPsc, plus communément appelé prion, serait capable de changer la conformation de PrPc, basée sur des hélices α, en structure majoritairement β(Pan et al., 1993). Leur agrégation aboutit à la formation de structures en fibrilles amyloïdes stables, détruisant la structure du tissu dont l'aspect est semblable à une éponge.

PrP est issu de l'expression du gène PRNP(Oesch et al., 1985) présent dans le génome de la cellule hôte. Un des facteurs génétiques de l'origine des maladies à prion est la présence de répétitions en tandem (octarepeat) dans la région N-terminale de PrP. Une expansion anormale de ces répétitions provoquerait le changement de l'isoforme normal en l'isoforme infectieux, où PrPsc serait non-structuré (Kushnirov et al., 2006; Goldmann, 2008; Malaise et al., 2008; Wang et al., 2009). La délétion de cette région résulterait en une baisse de l'oligomérisation des PrP tandis qu'une expansion l'augmenterait. Ces observations

démontrent que la région répétée n'est pas nécessaire à la progression de la maladie, mais qu'elle peut faciliter la pathogénèse et la propagation du prion(Yam et al.).

D'autres formes de prions ont pu être identifiées, notamment chez les levures, comme chez *Saccharomyces cerevisiae* qui en présente au moins 4 : Sup35, Ure2p, Rnq1p, and Swi1p (Crapeau et al., 2009). Les prions de levures ont aussi la capacité de s'agréger en fibrilles amyloïdes (Kajava et al., 2004) et contiennent des répétitions en tandem. Les répétitions sembleraient être primordiales dans la stabilité structurelle des fibrilles (Kajava et al., 2004; Shkundina et al., 2006).

I-6-4- Cancer

Bien que leur rôle dans les cancers soit encore peu connu et controversé, il a été montré que le polymorphisme des séquences répétées de certaines protéines oncogènes pouvait influer sur la carcinogénèse. Plusieurs études ont mis en évidence l'implication de la protéine récepteur des androgènes (AR) dans des cancers comme le cancer de la prostate ou le cancer du sein(Hao et al., 2010). Cette protéine, définie comme un facteur transcriptionnel activé via les hormones stéroïdiennes, est l'agent causal de l' amyotrophie spinobulbaire (Table 2). Elle présente une séquence répétée polyQ hautement polymorphe dans un domaine nécessaire à son activité transcriptionnelle. Néanmoins, les conséquences du polymorphisme de cette séquence sont variables selon le type de cancer et leurs analyses ont pu aboutir à des conclusions conflictuelles(Rebbeck et al., 1999; Given et al., 2000; Yu et al., 2000; Giguere et al., 2001). Les dernières en date supportent l'hypothèse qu' une expansion du nombre de répétitions aurait un effet oncoprotecteur dans le cancer du sein (Hao et al., 2010) et dans le cancer de la prostate(Nelson and Witte, 2002) lorsqu'elle dépasse 22 résidus. Par analogie, des répétitions courtes augmenteraient le risque de cancer et influeraient sur sa sévérité.

Un autre type de répétition rencontré dans le gène muc-1, codant pour la mucine, serait associé aux risques de tumeurs éptihéliales(Fontenot et al., 1993). Cette répétition riche en proline adopterait une conformation en hélice polyproline et présenterait des propriétés antigéniques lorsqu'elle serait exposée à la surface de la cellule.

Enfin, l'oncogène du sarcome d'Ewing (EWS) présente en N-terminal un domaine d'activation transcriptionelle (EWSactivation domain, EAD) composé majoritairement de répétitions d'un hexapeptide. La translocation de ce domaine est susceptible de produire

plusieurs protéines de fusion (les EFPs) qui ont une activité oncogénique régie par l'activation transcriptionnelle anormale de gènes cibles(Lee, 2007; Ng et al., 2007).

I-6-5- Rôle dans l'évolution et la virulence d'agents pathogènes

Il est de notoriété publique que les micro-organismes infectieux ont une capacité accrue d'évolution et développent des résistances très rapidement nécessaires à leur survie dans un organisme hôte(van Belkum, 1999). Ce phénomène est appelé variation de phase(van der Woude and Baumler, 2004). Cette capacité de variation est entre autres due aux cycles courts de leur multiplication, mais aussi à leur propension d'accumuler plus rapidement des mutations. La haute fréquence des événements mutationnels survenant au niveau des répétitions en tandem est un facteur prépondérant dans cette variabilité(van der Woude and Baumler, 2004), permettant au parasite d'échapper au système immunitaire ou d'augmenter sa pathogénicité. Des répétitions hautement polymorphiques ont été identifiées dans plusieurs parasites eucaryotes ou protozoaires. Par exemple, chez *Plasmodium falciparum* (agent pathogène de la malaria), environ 35% de ses protéines contiennent des répétitions en tandem riches en asparagine(Singh et al., 2004). L'étude des gènes antigéniques Csp, Msp-1 et Msp-2 (Rich and Ayala, 2000; Singh et al., 2004), a pu montrer que ce polymorphisme était à l'origine d'une grande variété de malarias(Snewin et al., 1991; McCutchan et al., 1992). Chez *Staphylococcus aureus*, des protéines de surface nécessaires à la reconnaissance des molécules d'adhésion de la matrice extracellulaire de l'hôte présentent de nombreuses répétitions en tandem, dont le nombre de répétitions influe sur l'accessibilité du domaine actif(McDevitt and Foster, 1995; Shopsin et al., 1999). *Trypanosoma cruzi* et *bruzi* (responsables de trypanosomiases) *Leishmania spp* (leishmaniose) ou encore *Bacillus anthracis* (anthrax) sont d'autres agents infectieux bactériens recensant une grande proportion de séquences répétées en tandem dans les protéines(Goto et al., 2008). Un autre rôle hypothétique de ces répétitions serait la stabilisation de l'activité enzymatique dans le sang. Il a été observé chez *Trypanosoma cruzi* que la demi-vie dans le sang de la trans-sialidase (secrétée par le parasite) était augmentée de 7à 35h(Buscaglia et al., 1999).

I-7 – identification des répétitions en tandem.

I-7-1 – L'identification avant les séquençages

Avant la généralisation du séquençage des génomes complets, l'identification des répétitions en tandem était réalisée via la technique de « Southern blot ». Les répétitions étaient ciblées au hasard de la recherche de séquences polymorphes du génome humain, ou en utilisant une séquence synthétique répétée en tandem comme sonde(Vergnaud, 1989).
Le séquençage d'un nombre croissant de génomes et la naissance de la bioinformatique vont faciliter grandement l'identification de ce type de séquences et permettre leur étude exhaustive.

I-7-2 – Quelques notions algorithmiques

Les dynamiques évolutives des répétitions dans les séquences génèrent plusieurs obstacles lors du développement d'outils bioinformatiques visant l'identification de ces répétitions. En effet, les différents évènements mutationnels qu'elles subissent au cours de l'évolution suscitent des questions à prendre en considération:

- Comment différencier les répétitions en tandem des répétitions dispersées dans les séquences ?
- Comment évaluer la similarité (ou la dégénérescence) des répétitions ?
- La définition de répétitions en tandem est assez subjective, quelles sont les valeurs à prendre en compte en termes de longueur des répétitions, de similarité ou encore de composition ?
- L'utilisation de méthodes exactes est souvent très couteuse en temps et en espace, ce coût d'autant plus grand si l'on conduit des explorations à grande échelle (comme les génomes ou les grandes banques de données). Comment réduire ce coût ?

Durant les dernières décennies, cette problématique a suscité la naissance de plusieurs algorithmes, recensés dans cette section. Pour une meilleure compréhension, quelques rappels traitants de l'algorithmique et des principales méthodes employées sont évoqués ci-après.

I-7-2-a – Complexité algorithmique

Un algorithme est une suite d'instructions élémentaires nécessaire à la résolution d'un problème. Cette résolution a un coût dans l'espace mais aussi dans le temps, estimer le coût

maximum d'un algorithme revient à définir sa complexité. La complexité en espace s'entend en termes de mémoire utilisée lors de l'exécution de l'algorithme, tandis que la complexité en temps mesure le nombre d'instructions nécessaire pour atteindre la solution. Par abus de langage, lorsqu'on parle de complexité de manière générale, cela sous-tend la complexité en temps. Généralement, on s'intéresse au temps maximum que pourrait prendre l'algorithme, correspondant au pire cas de figure qu'il aurait à résoudre. Ainsi, si l'on considère que chaque instruction élémentaire prend le même temps, la complexité ne mesure que l'ordre de grandeur du temps d'exécution (Table 4).

Par exemple, si le temps d'exécution d'un algorithme pour résoudre un problème de taille de données n est donné par $t(n)=xn^2 + yn + z$, on ne considèrera que le terme de plus grand ordre, soit xn^2. On note dans ce cas que l'algorithme a une complexité en $O(n^2)$ (on dit aussi que la complexité est quadratique).

Ainsi, tout algorithme peut voir son efficacité mesurée grâce à cette mesure. Lorsque la complexité atteint des ordres de grandeurs exponentiels pour lesquels la résolution d'un algorithme serait impossible à l échelle humaine, on a le plus souvent recours à des approximations de cet algorithme qui permettraient, dans un cas précis, de réduire l'ordre de grandeur de la complexité. C'est ce que l'on appelle des « heuristiques ».

Complexité	Type	n=10	n=30	n=60
$O(\log n)$	logarithmique	1µs	1,5 µs	1,8µs
$O(n)$	Linéaire	10 µs	30 µs	60 µs
$O(n\log n)$	Quasi-linéaire	10 µs	44 µs	107 µs
$O(n^p)$ ex : p=2	polynomiale	100 µs	900 µs	3,5ms
$O(C^n)$ ex : C=2	exponentielle	1ms	18m	366 siècles

Table 4. Différents types de complexités et leurs temps de résolution moyenne. Soit n le nombre d'opérations élémentaires exécutées par l'algorithme et 1µs son temps d'exécution.

I-7-2-b- La programmation dynamique et les algorithmes d'alignement

Une des bases dans l'analyse de séquences est la comparaison d'une séquence avec une autre afin d'estimer leur niveau de « similarité » (on parle d'homologie lorsqu'elles dérivent d'une

même séquence ancestrale et qu'elles ont une histoire évolutive commune). Ce niveau de similarité, ou au contraire cette « distance », peut être mesuré à l'aide de métriques, définie comme une fonction d sur un ensemble X tel que $d : X \times X \rightarrow R$ vérifie les propriétés suivantes :

• $d(x, y) \geq 0$, non négativité,

• $d(x, y) = 0 \Leftrightarrow x = y$, identité des indiscernables,

• $d(x, y) = d(y, x)$, symétrie,

• $d(x, z) \leq d(x, y) + d(y, z)$, inégalité triangulaire.

Les distances mesurées entre deux séquences sont basées sur les différences de caractères à un autre. Le passage d'une séquence à l'autre va nécessiter des opérations de transformation semblables à des évènements mutationnels ponctuels, que l'on appelle « opérations d'édition » :

- Des substitutions, correspondant au remplacement d'un caractère par un autre,
- Des délétions et/ou insertions appelés également indels, où l'on rajoute ou supprime un caractère

Ces opérations servent de base au calcul de la distance entre les deux séquences considérées. On retrouve parmi les métriques les plus utilisées dans la comparaison de chaines de caractères et de séquences la distance de Hamming (Hamming, 1950) ou la distance de Levenshtein (appelée aussi distance d'édition).

Soit deux séquences S1 et S2 toutes deux constituées de n caractères pris dans un alphabet $\sum$ tels que $|\sum|=26$.



Deux opérations de substitution sont nécessaires pour transformer S1 en S2. La distance de Hamming est dH=2

La distance de Hamming ne s'applique que lorsque les séquences ont une même longueur et ne fait intervenir que des substitutions. Lorsque les séquences ont une longueur différente, leur comparaison va faire appel à l'opération d'indel, occasionnant des trous ou des « gaps », représentés par le caractère «-». On va pouvoir ainsi créer un alignement entre S1 et S2. Un

alignement de deux séquences est modélisé en une matrice de deux lignes, la première étant appelée séquence source et la seconde la séquence cible. Soit *a* un caractère de S1 et *b* un caractère de S2, une colonne de l'alignement est représentée sous la forme:

$\begin{pmatrix} a \\ b \end{pmatrix}$

si a=b c'est un appariement exact (ou « match »), sinon c'est une substitution de a par b(ou « mismatch »)

notation: *Sub* (S1[a],S2[b])

$\begin{pmatrix} a \\ - \end{pmatrix}$

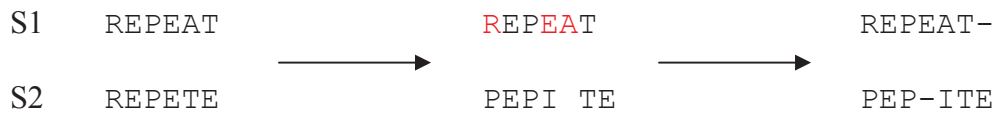
Correspond à la délétion de a

notation: *Del*(S1[a])

$\begin{pmatrix} - \\ b \end{pmatrix}$

Correspond à l'insertion de b

notation: *Ins*(S1[a])



Dans ce cas de figure, la séquence S1 va subir 2 opérations de substitutions (R en P, et E en I), mais aussi une opération de délétion (délétion de A) et une opération d'insertion (insertion de E). La distance de Levenshtein prend en compte ces deux types d'événements, définie comme le nombre d'opérations à appliquer à une chaine de caractères pour obtenir la seconde. Une généralisation de cette distance introduit une notion de « coût » ou score en fonction de l'opération appliquée sur la séquence, indépendamment de sa position. Ainsi, quand pour la distance de Hamming et la distance d'édition simple on comptabilisait le coût d'une opération comme égal à 1, dans la version généralisée les coûts sont dépendant du type d'opération. Si l'on applique cette métrique sur un alignement et que l'on considère le «-» comme un caractère constitutif de la séquence, on peut définir des paires de caractères issus de chacune des séquences ainsi que leurs coûts. La somme des coûts associés à chaque colonne va représenter la distance d'édition de l'alignement.

De nombreux algorithmes calculant un alignement de deux séquences ont pu être publiés et sont pour la plupart basés sur la programmation dynamique (Needleman and Wunsch, 1970; Smith and Waterman, 1981).

Soient deux séquences S et T à aligner via la programmation dynamique de longueurs n et p. On définit la matrice M de dimension (n+1)X(p+1) comme une matrice de coûts minimums (ou de « score »). Le remplissage de la matrice comporte deux phases , $\forall\, i \in [1; n]$ et $\forall j \in [1;p]$:

- L'initialisation :

  $M(0 , 0) = 0$;

  $M(i , 0) = M(i\text{-}1 , 0) + Del(S[i])$;

  $M(0 , j) = M(0 , j\text{-}1) + Ins(T[j])$;

- La formule de récurrence :

$$M(i , j) = \min \begin{cases} M(i\text{-}1 , j\text{-}1) + Sub(S[i] , T[j]) \\ M(i\text{-}1 , j) + Del(S[i]) \\ M(i , j\text{-}1) + Ins(T[j]) \end{cases}$$

Par conséquence, le calcul du coût à une position (i,j) dépend directement des valeurs retrouvés aux positions (i-1,j-1), (i-1,j) et (i,j-1) (figure 16). La matrice une fois remplie nous renseigne sur le coût de transformation de S en T. Pour recréer le ou les alignements optimaux des deux séquences à partir de cette matrice, la méthode de « backtracking » va être employée.

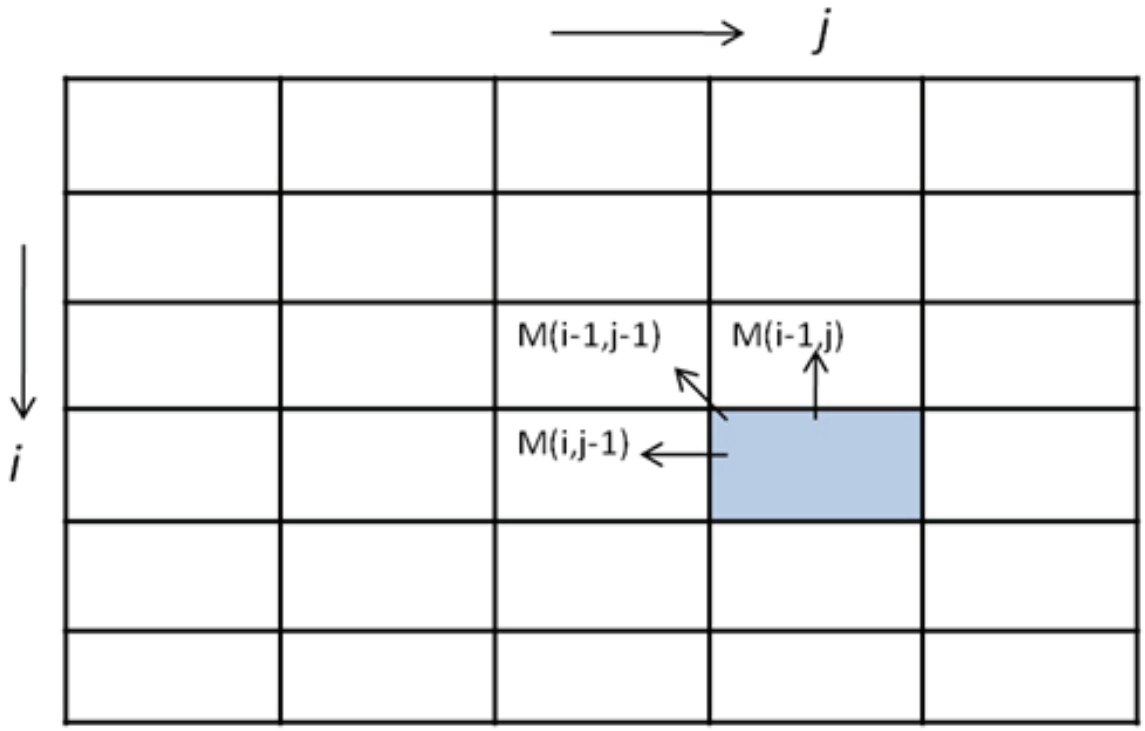


Figure 16. Remplissage de la matrice de coût par programmation dynamique. La valeur de M(i , j) ne dépend que des valeurs voisines en M(i-1,j-1), M(i,j-1) et M(i-1,j).

Procédure de backtracking (figure 17) :

a- On part de la cellule M(n,p) pour remonter vers la cellule M(0,0)

b- On passe de la cellule M(i,j) à une des trois positions voisines susceptible d'avoir généré la valeur de M(i,j). Chaque possibilité est représentée par une flèche.

c- On reproduit ainsi un chemin de M(n,p) à M(0,0) permettant de reconstituer un alignement optimal, le passage d'une cellule à une autre pouvant être interprété selon la règle suivante:

- Une flèche diagonale de la cellule (i,j) vers la cellule (i-1,j-1) est un appariement exact entre S[i] et T[j] si S[i]=T[j] ou une substitution si S[i]≠T[j],
- Une flèche verticale de (i,j) vers (i-1,j) est une délétion du caractère S[i] de S (on le représente dans l'alignement en insérant un «-» dans la ligne de T)
- Une flèche horizontale de (i,j) vers (i,j-1) est une insertion du caractère T[j] de T (on le représente dans l'alignement en insérant un «-» dans la ligne de S)

Plusieurs chemins sont possibles, aboutissant à autant d'alignements optimaux. Un exemple d'alignement optimal obtenu à partir de la matrice de la figure 17 est le suivant :

```
PEPI-TE
REPEAT-
```

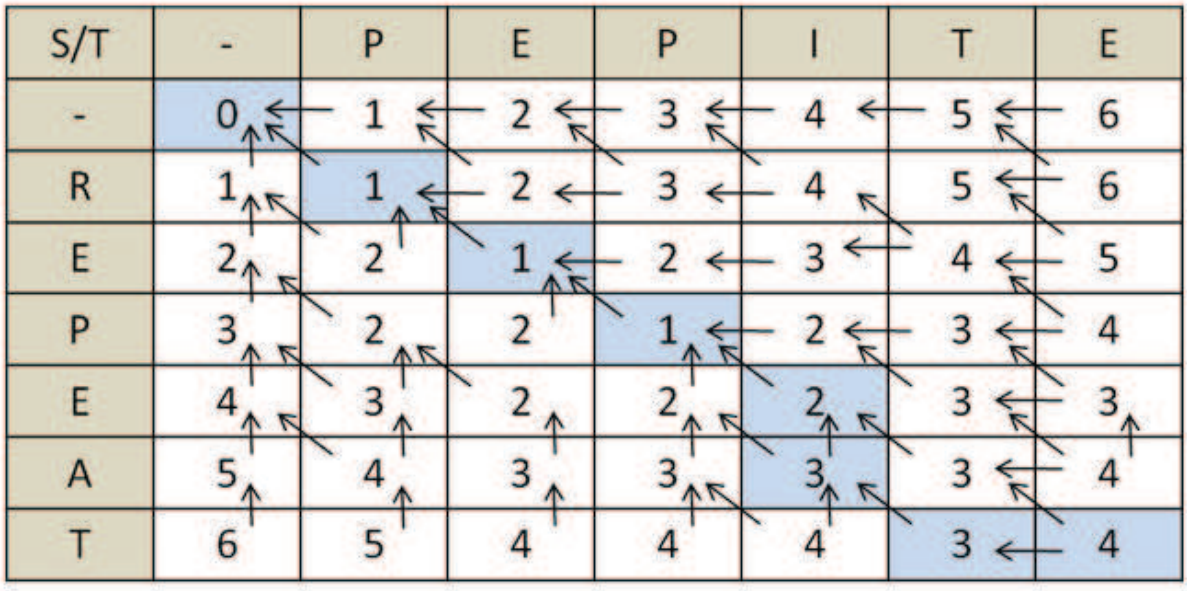


Figure 17. Exemple de backtracking sur une matrice de coûts issue de deux séquences à aligner : S= (REPEAT) et T=(PEPITE). Les coûts associés aux différentes opérations sont Sub(a,b)=0 si a=b, Sub(a,b)=1 si a≠b et Ins(a)=Del(b)=1.

Cette méthode d'alignement est appelée alignement global(Needleman and Wunsch, 1970), du fait qu'elle prend en compte l'intégralité des deux séquences à aligner. La complexité de cet algorithme est de l'ordre O(np), la matrice étant constituée de (n+1) lignes et (p+1) colonnes.

On peut distinguer d'autres types d'alignement, eux aussi basés sur la programmation dynamique mais pour lesquels l'initialisation, le remplissage ou la méthode de backtracking diffèrent.

L'alignement local, basé sur l'algorithme de Smith et Waterman(Smith and Waterman, 1981) a pour but d' aligner uniquement des parties des séquences d'intérêt plutôt que l'ensemble d'elles. Son application sur les séquences protéiques permet de mettre à jour des similarités de motifs entre les deux séquences alors qu'elles sont assez différentes dans leur globalité. La complexité de cet algorithme est la même que celle de l'alignement global. Un autre type d'alignement est l'alignement semi-global, semblable au global mais avec des caractéristiques de remplissage de matrice proches de celles de Smith et Waterman.

Tous ces algorithmes sont dits exacts, car ils donnent des solutions optimales aux problèmes posés. Néanmoins, leur complexité font que leur application est extrêmement coûteuse en temps, notamment lorsqu'il faut produire une grande quantité d'alignements avec des banques de séquences. A l'opposé des algorithmes exacts, il existe des algorithmes heuristiques approximant les alignements optimaux avec un ordre de complexité moindre. C'est le cas des algorithmes implémentés dans BLAST (Altschul et al., 1990) et FASTA(Lipman and Pearson, 1985), qui sont des programmes couramment utilisés pour l'alignement local d'une séquence avec les séquences issues d'une banque. Pour ces programmes, les coûts ou scores liés à des opérations de substitutions sont définis par des matrices dites de substitution. Les plus usuelles sont les matrices PAM (« Point Accepted Mutation ») et les matrices BLOSUM(« BLOcks SUbstitution Matrix »)(Henikoff and Henikoff, 1992; Henikoff and Henikoff, 2000). Les matrices BLOSUM sont construites à partir de BLOCKS, portions de séquences hautement conservées ne contenant pas d'indels, leur alignement permet de déduire une probabilité de substitution d'un résidu par un autre et ainsi créer une matrice de substitution. Le nom d'une matrice BLOSUM est suivi d'un indice caractérisant le pourcentage d'identité des séquences utilisées pour construire la matrice. Ainsi, une matrice BLOSUM62 a été générée à partir de séquences ayant 62% d'identité (figure 18).

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Figure 18. Matrice de substitution BLOSUM62. La substitution d'un résidu par un autre est associée à un score, utilisé dans le calcul de score d'un alignement.

Une fois tous les alignements locaux crées, ils vont être classés selon leurs scores, seules les paires dont le score est supérieur à un seuil défini sont prises en considération. Le score S d'un alignement renseigne sur sa qualité, tandis que d'autres mesures évaluent sa significativité en termes statistiques, comme la P-value (probabilité de trouver un tel alignement dans des séquences aléatoires de même composition et taille), ou l'E-value (nombre d'alignements avec un score égal ou supérieur à S susceptibles d'être retrouvés au hasard dans une banque de séquences).

Une généralisation de l'alignement de deux séquences à n séquences de longueur l est appelée l'alignement multiple. L'approche est identique à celles qui ont été évoquées précédemment, impliquant une matrice de programmation dynamique à n dimensions. Si l'approche pour deux séquences était de complexité polynomiale, l'alignement multiple atteint une complexité exponentielle en $O((2l)^n)$, les rendant inutilisables en pratique lorsque n est grand. De nombreuses heuristiques ont pu être proposées pour l'alignement multiple sous optimal, elles peuvent être classées en plusieurs catégories :

- Algorithmes progressifs :

  Basés sur l'alignement préalable des paires de séquences possibles, puis en les combinant de la paire la plus similaire à celle la plus distante. La fonction de score utilisé pour l'alignement des paires de séquences peut être appliquée à l'ensemble de l'alignement multiple appelée la fonction des sommes des paires ou «sum of pairs» (SP). L'ordre dans lequel les séquences vont être rajoutées à l'alignement est défini par un arbre, appelé aussi arbre guide, déterminé par des algorithmes de clustering, comme le «Neighbour Joining»(Saitou and Nei, 1987).

  On retrouve dans cette classe l'algorithme d'alignement en étoile (center-star)(Gusfield, 1993), ou ceux des programmes CLUSTALW(Thompson et al., 1994) et T-COFFEE(Notredame et al., 2000).

- Algorithmes itératifs :

  Plus précis que les algorithmes progressifs mais plus lents car à la différence des algorithmes progressifs, ils alignent les séquences de façon simultanée. Le programme DIALIGN(Subramanian et al., 2005) est basé sur ce type d'algorithme.

Des programmes tels que MUSCLE(Edgar, 2004), ou MAFFT(Katoh et al., 2005) emploient des méthodes empruntant à la fois aux algorithmes progressifs et itératifs, ou d'autres sont fondés sur des algorithmes génétiques (SAGA(Notredame and Higgins, 1996)). Tous ces algorithmes sont différents de par leur approche et ne donnent pas de résultat exacts, il n'existe à l'heure actuelle aucune méthode universelle primant sur toutes les autres, leur utilisation dépendra du domaine d'application. Les applications de l'alignement multiple sont diverses, telles que la caractérisation de régions conservées entre plusieurs séquences ou la construction d'arbres phylogénétiques. De même plusieurs programmes d'identification de répétitions en tandem s'appuient sur ces algorithmes, les motifs constitutifs d'une même région répétée pouvant être alignés entre eux afin de déterminer leurs similarités.

I-7-3– Les programmes d'identification de répétitions en tandem

I-7-3-a- Les difficultés liées à l'identification

Nous avons pu montrer dans les sections précédentes que les répétitions en tandem subissent un grand nombre d'événements mutationnels responsables de leur polymorphisme. Ces multiples mutations vont « dégénérer » la répétition d'origine, rendant l'identification des

unités répétées et de ses limites dans une séquence beaucoup plus floues(Benson and Dong, 1999). La recherche des répétitions en tandem dans les séquences en vue de leur analyse systématique a suscité le développement de plusieurs générations de programmes, pouvant être classés selon plusieurs approches.

I-7-3-b- Algorithmes exhaustifs

Les approches simplistes :

Lorsqu'on n'autorise aucune mutation ponctuelle dans la séquence, il est possible de retrouver toutes les répétitions maximales exactes dans une séquence de longueur n avec une complexité linéaire O(n) (Gusfield and Stoye, 2004).

Les approches SSA :

Les premières approches d'identification des répétitions se sont appuyées sur les méthodes d'alignement local décrites précédemment, où la séquence d'intérêt est comparée à elle-même, d'où leur nom : « Self Sequence Alignment » (SSA). Chaque alignement sous-optimal se voit attribuer un score permettant d'évaluer sa significativité. La complexité étant de l'ordre $O(n^2)$, leur utilisation contre des banques de données conséquentes les rend inadaptés. Ils vont détecter de manière aspécifique tous les types de répétitions dans les séquences (aussi bien en tandem que dispersées), et se révèlent bien moins efficaces lorsque la longueur du motif est inférieure à 10 résidus. Elles sont appliquées dans les programmes comme INTREP(Marcotte et al., 1999), RADAR(Heger and Holm, 2000) ou TRUST(Szklarczyk and Heringa, 2004).

Les approches basées sur des métriques :

Elles autorisent l'intervention d'événements mutationnels, les différences de deux répétitions consécutives ne devant pas dépasser une distance k . Cette distance peut être assimilée à la distance de Hamming, comme dans l'algorithme de (Landau et al., 2001) qui a une complexité en O(nk log(n/k)) ; ou la distance d'édition dans le programme TRED (Sokol et al., 2007) dont la complexité est en O(nklog klog(n/k)) (Par exemple la séquence R = caagct|cagct|ccgct est une répétition où k=2 si la métrique est la distance d'édition). Une amélioration de l'algorithme basé sur la distance de Hamming par (Kolpakov and Kucherov, 2003) existe, portant la complexité à O(nk log(k)+S) où S est le nombre de séquences

répétées identifiées. Une version de cet algorithme a été implémentée dans le programme mreps (Kolpakov et al., 2003) avec des étapes heuristiques de filtrage.

I-7-3-c- Heuristiques et améliorations de la programmation dynamique

Afin de diminuer la complexité en temps lors de l'identification des répétitions, certains programmes vont avoir recours à des heuristiques. Plusieurs méthodes heuristiques sont utilisées dans une première phase de « filtrage », visant à recenser les candidats pour des répétitions en tandem hypothétiques dans une séquence. Via l'analyse de critères statistiques, les régions candidates sont sélectionnées, sur lesquelles d'autres algorithmes de détection plus coûteux seront appliqués durant la phase dite de « validation ». Le programme le plus connu est Tandem Repeats Finder ou TRF (Benson, 1999), développé principalement pour la recherche de répétitions en tandem dans les séquences d'ADN. Il présente deux paramètres principaux : la probabilité de match(ou d'identité entre deux résidus alignés) p_m, et la probabilité d'indel p_i. La phase initiale repose sur la méthode d'extension de «graines» ou « seeds», dont la complexité est linéaire.

L'extension de graines ou « seed extension » (figure 19):

Les graines sont de courtes sous-séquences issues de la séquence analysée, appelées aussi « k-tuples », générées à partir d'une fenêtre de *k* résidus consécutifs. Chaque occurrence issue d'une même graine va être répertoriée dans un historique ainsi que sa position. Pour chacune des occurrences, la distance *d* entre sa position et celles des occurrences précédentes dans la séquence va être calculée, elle-même stockée dans une liste de distances D_d. Pour chacun des motifs candidats de longueur *d*, des tests statistiques sont appliqués en fonction de la distribution de la longueur du motif candidat *d*, la probabilité de match p_m, et la probabilité d'indel p_i et la taille de la fenêtre *k.* Les deux critères statistiques principaux sont la somme des faces (« sum of heads ») et la taille apparente.

La somme des faces :

Lorsque deux motifs candidats de longueur *d* se succèdent, leur alignement va être modélisé par une séquence de tirages de Bernoulli indépendants, assimilable à un jet de pièce. On associe un match à face (« head » ou H) et une substitution ou un indel à pile (« tail » ou T) (figure 20). Soit Rd,k,p_m la variable aléatoire correspondant au nombre total de faces dans les suites consécutives de k faces (ou « runs » de k faces) ou plus dans une séquence iid (indépendante et identiquement distribuée) de Bernoulli de longueur d avec une probabilité de

face égale à pM. La distribution de cette probabilité peut être approximée à une loi normale, utilisée pour déterminer le plus grand nombre x tel que dans 95% des cas Rd,k,p_m>x. Par exemple, si pm=0.75, k=5 et d=100, la somme des faces sera de 26. Autrement dit, pour un motif de 100 résidus et si les copies doivent matcher au moins sur 75 positions, si l'on compte les matchs qui remplissent une fenêtre de longueur 5, on s'attend dans 95% du temps à compter au moins 26 matchs.

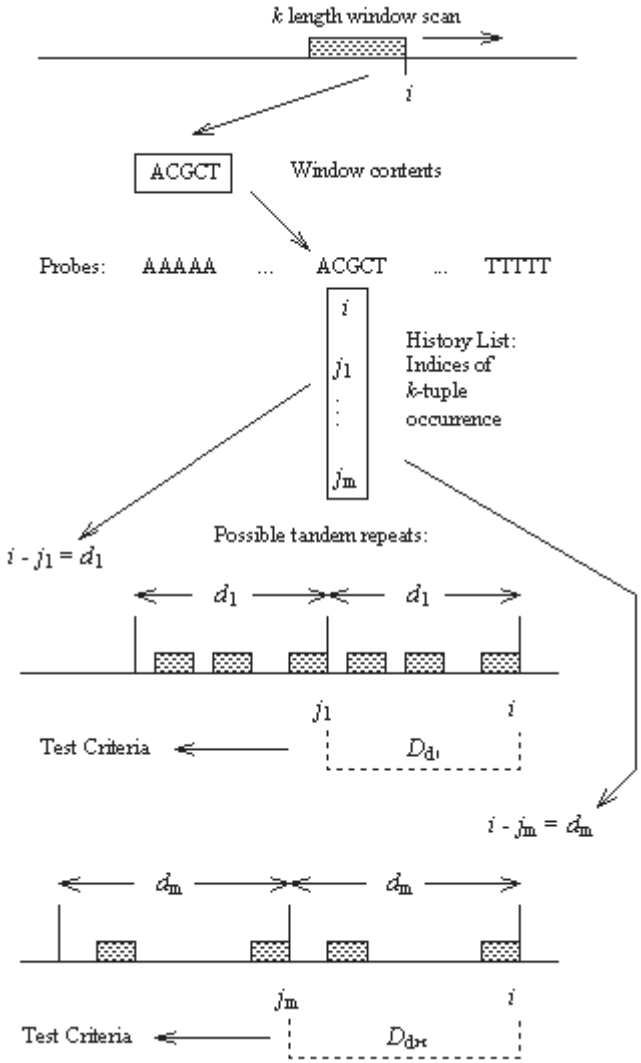


Figure 19. Plateforme de détection des répétitions par TRF. La séquence est scannée par une fenêtre de *k* résidus, permettant de répertorier tous les motifs candidats possibles. Après détermination des distances entre chacune des occurrences, les motifs candidats subissent deux tests statistiques.

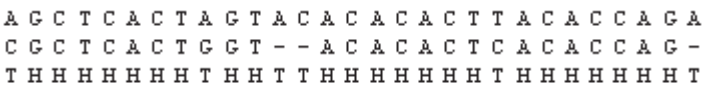


Figure 20. Modélisation d'un alignement de deux motifs adjacents. H représente un match, tandis que H indique une substitution, une insertion ou une délétion.

La taille apparente :

Utilisé pour distinguer les répétitions en tandem des répétitions en dispersées. On associe la variable aléatoire Sd,k,p_m à la distance entre le premier et le dernier run de k faces. La distrubution de la variable est déterminée par simulation car elle est dépendante de la du premier test statistique. On génère des séquences aléatoires à partir de $p_{m,}$ pour lesquelles on calcule la distance entre le premier et dernier run si elles valident le premier test. On détermine ainsi le critère de taille apparente qui est le nombre maximum y de sorte que Sd,k,p_m soit supérieur à y dans 95% des cas.

Lorsque les motifs candidats passent avec succès les deux tests, ils sont traités dans la seconde phase. Celle-ci fait appel à une adaptation de la programmation dynamique utilisée pour l'alignement de séquences circulaires, notamment les séquences répétées en tandem : le WDP ou «Wraparound Dynamic Programming»(Fischetti et al., 1993). Elle permet d'aligner un motif plusieurs fois avec une séquence donnée tout en gardant la complexité d'algorithmes de programmation dynamique classiques. Si au moins deux copies du motif ont pu être alignées, alors n considère être en présence d'une répétition en tandem.

A l'instar de TRF, d'autres programmes utilisent ce genre de pipeline, ou sont basés sur le WDP.

I-7-3-d- Profils et Modèle de Markov caché(HMM)

Lorsque les séquences répétées en tandem sont vraiment dégénérées, elles sont d'autant plus susceptibles de passer outre les critères de sélection définis par les programmes décrits précédemment. Une approche puissante consiste en l'utilisation de « profils », qui contrairement aux méthodes conventionnelles de comparaison de séquences, n'utilise non pas

une séquence comme objet de requête mais une famille de séquences similaires(Gribskov et al., 1987). Cette comparaison est possible à partir d'un alignement de plusieurs séquences sélectionnées *a priori*. A partir de l'alignement multiple, des scores sont associés aux différents événements mutationnels à chacune des positions de la séquence et sont représentés sous forme d'une matrice de score, ou profil(Bucher and Bairoch, 1994). Une approche de modélisation du profil est le modèle de Markov caché (Hidden Markov Model, ou HMM). Initialement utilisé pour des problèmes de reconnaissance vocale, puis appliqué à la biologie(Krogh et al., 1994), il est apparenté aux automates probabilistes. C'est un modèle statistique représentant un processus de Markov dont l'état n'est pas observable, mais qui peut être prédictible. Il est typiquement composé :

- D'états
- De transitions entre les états : A chaque transition est associé une lettre de l'alphabet considéré (appliqué aux séquences protéiques l'alphabet sera de 20, 4 pour les nucléotidiques)
- De probabilités de transitions

On l'appelle « caché » car on accepte que le modèle génère une séquence d'évènements ou observations consécutives dont on ne connaît pas la séquence d'états qui les génère. Le modèle est capable d'évoluer en étoffant son jeu de données de référence (phase d'apprentissage) et permet d'estimer la séquence la plus probable d'états via des algorithmes tels que l'*algorithme de Viterbi*(Selin, 1967).

L'application de ce modèle aux alignements de séquences biologiques comprend plusieurs contraintes (figure 21) :

- Trois états sont considérés : un état match/mismatch, un état insertion et un état délétion.
- L'automate nécessite une phase d'apprentissage à partir d'un alignement de séquences initial, permettant de calculer les probabilités de transition (la probabilité de passer d'un résidu à un autre).
- Des probabilités d'émission de résidu pour chaque état match/mismatch, c'est la partie cachée du modèle.

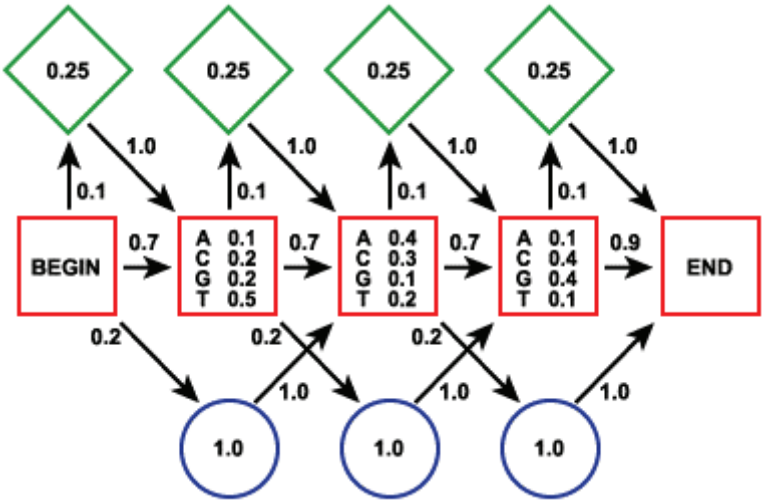


Figure 21. Exemple de modèle de Markov pour une séquence nucléotidique. Les carrées rouges correspondent aux états match/mismatch, les losanges verts aux états d'insertion, les cercles bleus aux états de délétion. Chaque état possède des probabilités d'émission de résidus cachées.

Le système va être ensuite appliqué sur une banque de données de séquences afin de déterminer si d'autres séquences sont susceptibles d'appartenir au groupe de séquences initial, appelés « matches ». L'exécution d'un profil renvoie une liste ordonnée des matches potentiels triés par leur score. La significativité des résultats ne pouvant être calculée via des méthodes directes, elle est estimée statistiquement par des méthodes empiriques (passage du profil contre des séquences aléatoires de même composition et longueur).

Les HMM ont pu être implémentés dans le package HMMER(Sonnhammer et al., 1998), composés de plusieurs de plusieurs outils tous basés sur les HMM. On les retrouve également dans le package *pftools* proposé sur le site de PROSITE(Bucher et al., 1996; Falquet et al., 2002), il utilise une version généralisé des profils classiques et permet la conversion de plusieurs types de formats en profils. Ce package a largement été utilisé par l'équipe pour mettre en évidence des motifs dans des séquences et prédire leur structure(Kobe and Kajava, 2001).

Cette approche est plus sensible que les méthodes classiques de comparaison de séquence, cependant elle trouve ses limites pour la recherche globale de répétitions en tandem dans les génomes, la phase d'apprentissage nécessitant la connaissance préalable des séquences générant les profils.

2-OBJECTIFS

Les objectifs de cette thèse peuvent être résumés en plusieurs points :

- Identification *ab initio* et à grande échelle des répétitions en tandem dans les séquences protéiques

- Stockage des répétitions identifiées dans une base de données ergonomique et accessible au public

- Analyse exhaustive des répétitions visant à :

    - Caractériser les propriétés physicochimiques, structurelles, évolutives et fonctionnelles communes aux répétitions dans les séquences protéiques,

    - Mettre en évidence les éventuelles corrélations entre les différents critères quantitatifs et qualitatifs des répétitions.

3-RESULTATS

CHAPITRE 1

Identification *ab initio* des répétitions en tandem dans les séquences protéiques

Résumé

Ce chapitre expose la procédure employée pour l'exploration exhaustive des répétitions en tandem dans les plus grandes banques de données de séquences protéiques. Au préalable, un inventaire des programmes d'identification existants a été mené, lesquels ont été évalué par rapport à nos besoins. Les besoins sont fondés sur plusieurs critères : la possibilité d'explorer les séquences protéiques et d'identifier les répétitions sans connaissance préalable (*ab initio*), la rapidité et l'efficacité du programme. Ce chapitre met le doigt sur l'inadéquation des programmes actuels vis-à-vis de ces attentes: la plupart ont été conçus spécifiquement pour les séquences nucléotidiques, ou proposent des algorithmes de complexité trop grande pour être exploitable contre de grandes banques de séquences protéiques. Dans cette démarche, nous avons développé un nouveau programme d'identification des répétitions en tandem du nom de T-REKS. Il repose sur le clustering des longueurs de courtes sous chaines (ou « short strings ») via l'algorithme des K-Means. Afin de tester ses performances, nous nous sommes livrés à des bancs d'essais avec d'autres programmes de référence.

Dans une dernière partie, des tests préliminaires d'exploration de séquences nucléotidiques par T-REKS ont été effectués.

Le fonctionnement du programme est détaillé dans l'article ci-dessous:

Jorda J, Kajava AV.: **T-REKS: identification of Tandem REpeats in sequences with a K-meanS based algorithm.** Bioinformatics. 2009 Oct 15;25(20):2632-8.

BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 20 2009, pages 2632–2638
doi:10.1093/bioinformatics/btp482

Sequence analysis

T-REKS: identification of Tandem REpeats in sequences with a K-meanS based algorithm

Julien Jorda* and Andrey V. Kajava*

Centre de Recherches de Biochimie Macromoléculaire UMR 5237, CNRS, University of Montpellier 1 and 2, Montpellier, France

Received on March 25, 2009; revised on August 5, 2009; accepted on August 7, 2009

Advance Access publication August 11, 2009

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: Over the last years a number of evidences have been accumulated about high incidence of tandem repeats in proteins carrying fundamental biological functions and being related to a number of human diseases. At the same time, frequently, protein repeats are strongly degenerated during evolution and, therefore, cannot be easily identified. To solve this problem, several computer programs which were based on different algorithms have been developed. Nevertheless, our tests showed that there is still room for improvement of methods for accurate and rapid detection of tandem repeats in proteins.

Results: We developed a new program called T-REKS for *ab initio* identification of the tandem repeats. It is based on clustering of lengths between identical short strings by using a K-means algorithm. Benchmark of the existing programs and T-REKS on several sequence datasets is presented. Our program being linked to the Protein Repeat DataBase opens the way for large-scale analysis of protein tandem repeats. T-REKS can also be applied to the nucleotide sequences.

Availability: The algorithm has been implemented in JAVA, the program is available upon request at http://bioinfo.montp.cnrs.fr/?r=t-reks. Protein Repeat DataBase generated by using T-REKS is accessible at http://bioinfo.montp.cnrs.fr/?r=repeatDB.

Contact: julien.jorda@crbm.cnrs.fr; andrey.kajava@crbm.cnrs.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

An increasing number of complete genome sequences are being generated and deposited into the databases. The next great challenge is to understand the genome data. A significant portion of the proteins carrying fundamental functions contain arrays of tandem repeats (Marcotte *et al*., 1999). Over the last years a number of evidences has been accumulated about the high incidence of tandem repeats in the sequences of virulence factors of pathogenic agents, toxins and allergens (Kajava *et al*., 2006). Furthermore, the tandem repeats frequently occur in amyloidogenic, prion and other disease-related sequences (Baxa *et al*., 2006; Nelson and Eisenberg, 2006). This implies that this class of sequences may have a broader role in human diseases than was previously recognized. Along this line, the discovery of these domains and their structure–function study promise to be a fertile direction for research leading to the identification of targets for new medicaments and vaccines.

A systematic bioinformatics analysis of protein repeats in genomes can provide a global view on these motifs, on their structures, functions and evolution and, in its turn, may result in a significant improvement of our understanding of the biological meanings of the genome sequences. The 'biological' tandem repeats are usually not perfect containing a number of mutations (substitutions, insertions, deletions) accumulated during evolution and some of them cannot be easily identified. Over the last years, several algorithms, software and approaches have been developed (Andrade *et al*., 2000; George and Heringa, 2000; Heger and Holm, 2000; Landau, Schmidt et al., 2001; Kajava and Steven, 2006; Newman and Cooper, 2007) for identification of repeats in biological sequences. Programs such as INTREP (Marcotte *et al*., 1999), RADAR (Heger and Holm, 2000) and TRUST (Szklarczyk and Heringa, 2004) are based on sequence self-alignment (SSA) algorithms. These programs are especially efficient for detection of long repeats (more than ~10 residues long), however, they frequently fail to identify short repeats and do not distinguish between tandem and interspersed repeats. In addition, the SSA algorithms with their time complexity of $O(n^2)$ (where n is the length of sequence), are relatively slow, and, therefore, do not suit well for the large scale analysis. Other type of programs, such as Tandem Repeats Finder (Benson, 1999), XSTREAM (Newman and Cooper, 2007) or MREPS (Kolpakov *et al*., 2003) rely on short string extension algorithm or as STAR (Delgrange and Rivals, 2004) and TRED (Sokol *et al*., 2007) use improved dynamic programming algorithms. They have time complexity lower than SSA algorithms and therefore, are more rapid. Most of these programs are predestined for DNA sequences. XSTREAM is well adapted for a large-scale search of protein repeats, however, it fails to identify some tandem repeats. It is worth also mentioning, approaches that apply sequence profile methods and HMMs (Gribskov *et al*., 1987). These approaches are the best for detection of long imperfect repeats (Kajava *et al*., 1995; Lupas *et al*., 1997). However, they require *a priori* generated alignments of putative repeats and, therefore, are not suitable for automated *ab initio* large scale analysis.

Thus, despite the existence of a number of methods for determination of tandem repeats there is still room for their improvement and for development of an accurate and rapid program

*To whom correspondence should be addressed.

 © The Author 2009. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

dedicated for the systematic analysis of the repeats in genomes. In this work, we describe a new program T-REKS for protein tandem repeat identification which is based on the analysis of distribution of short strings within the sequence by using *K*-means algorithm. Furthermore, we used output of our program to collect a database of protein tandem repeats. The content of this database is publicly accessible via a web-site. Our objective is to make the results of the systematic bioinformatics analysis available via a regularly updated web-server, which will be a useful tool for scientists interested in structure, function, evolution and application in medicine and technology of proteins with tandem repeats.

2 METHODS

A flowchart of the algorithm is shown on Figure 1.

2.1 Short string probes and *K*-means clustering

To probe an analyzed sequence for the presence of tandem repeats we use short strings (SS) no longer than the repeat length (see example in Fig. 2). For proteins, the size of SS was chosen equal to two because the longer SS were less efficient for detection of two-residue repeats and some degenerous repeats. The search with one-residue probe turned to be less selective. Prior to the application of our algorithm, all homorepeats (tandem repeats of a single residue) were excluded from the analyzed sequences and were registered in our Protein Repeat DataBase for further studies (Fig. 1).

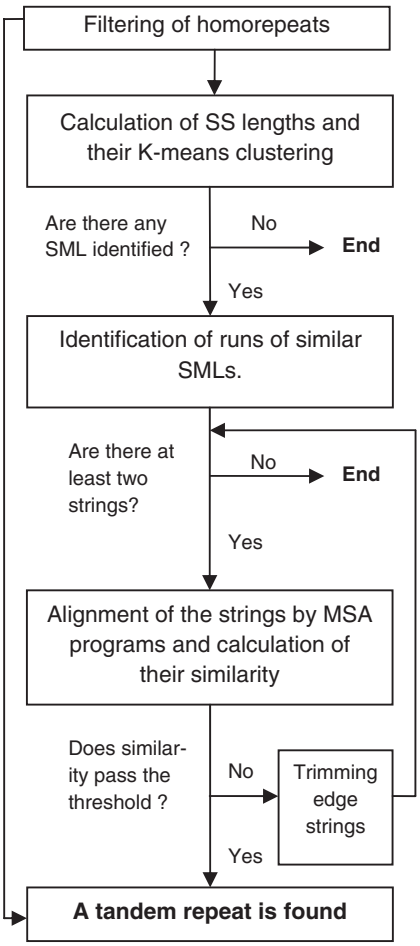


Fig. 1. Flowchart of the T-REKS program applied to a given sequence.

In a tandem repeat region, the most frequently occurred length between identical SSs should be equal to the repeat length. Therefore, detection of regions of an analyzed sequence where certain lengths between identical SSs have anomalously high occurrence may lead to the localization of the tandem repeats. Tandem repeats in the sequences of biological macromolecules (proteins and DNA) have two properties that may hamper the application of this approach. First, the degenerous character of the 'biological' repeats diffuses peaks of the SS length distribution. Second, a given protein or DNA sequence can have several different tandem repeat regions. In this case the same type of SS can be involved in different repeats and calculation of a simple mean of the length occurrences allows to identify only repeats of one length and leaves the remaining repeats unrecognized. To overcome these problems, we use a well-known algorithm for unsupervised classification called *K-means algorithm* (MacQueen, 1967). In our program, the lengths between identical neighbouring SSs were used as datapoints of the K-means algorithm to find potential lengths of the tandem repeats. For example, in Figure 2 the datapoints are lengths 5, 10 and 11 of a short string **EL**. This method partitions all datapoints into K clusters for user-defined K. For each partition a central datapoint (*centroid)* may be defined. K initial centroids are selected from the dataset either randomly, by applying hierarchical clustering or other techniques. Then euclidian distances are calculated between each datapoint and the centroids to assign the datapoint to the cluster which has the nearest centroid.

After this, the positions of the centroids are changed within each of the clusters and this procedure iteratively repeated until the consistency of clusters does not change anymore.

Usually, the K starting datapoints are selected randomly. However, it has been demonstrated that different choice of the initial centroids can lead to different results. To solve this problem, we implemented an algorithm which determines the initial centroids based on divisive hierarchical clustering (Johnson, 1967). This algorithm starts from one single cluster which includes all datapoints. This cluster is then iteratively subdivided into smaller clusters based on a rule to maximize the distance between the clusters. This process stops once the number of clusters becomes equal to user-defined K. The K centroids of these clusters constitute starting points in subsequent clustering by the K-means algorithm. Application of this algorithm to protein sequences required adjustment of several parameters. For example, K can not be less than the number of different types of tandem repeats in the analyzed sequence.

On the other hand, time complexity of this algorithm is $O(n \cdot K)$ and this favors smaller K. Our tests suggested that $K = 10$ gives the most accurate and rapid results for the identification of the protein repeats. The accuracy of this approach is also a function of the length of the analyzed sequence. Statistically, the longer is the sequence the higher is the number of occurences of a given short string. The increase of the occurrences will amplify a background noise and decrease the quality of detection at the clustering steps. Our tests of the algorithm with $K = 10$ shows that this problem appears when protein length is longer than 1500 residues. In this case, our program splits the sequences before and concatenates after the analysis.

2.2 Establishment of tandem repeat lengths

The procedure consists of three steps:

(1) The first step is separately applied to each type of SS found in the analyzed sequence. For example, Figure 2 demonstrates this procedure for a short string **EL**. All lengths between neighbouring **EL**s (5, 10, 11) are considered as datapoints for K-means algorithm. Within each cluster generated by this algorithm, we select the most frequent length and call it a Short string Main Length (SML). If a cluster has several most frequent lengths that occur the same number of times, the shortest length is chosen. As a result of this step, K

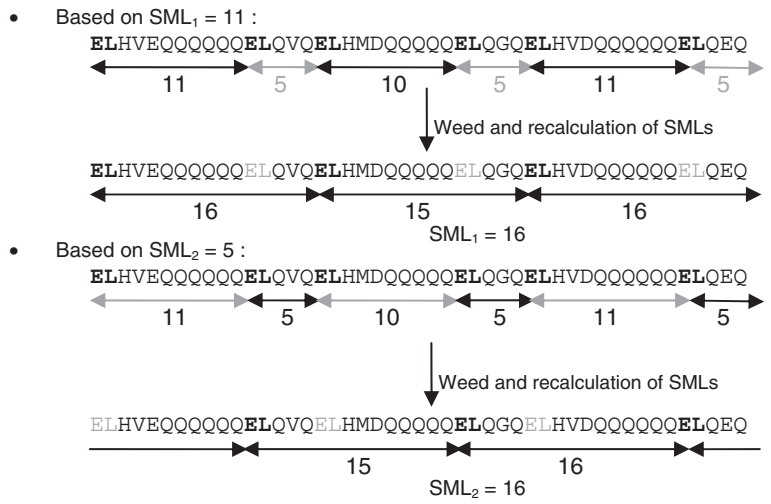


Fig. 2. Example of establishment of SMLs for short string **EL** and $K = 2$.

different SMLs of **EL** are obtained. In Figure 2, $K = 2$ and $SML_1 = 11$; $SML_2 = 5$.

(2) Not all SMLs may correspond to the tandem repeat lengths, because a given short string may occur more than one time within a repeat (Fig. 2). To unmask the real repeat length, we use a procedure of SS 'weed'. First, we consider only **EL**s which are separated by lengths that are equal or close to the SMLs. The threshold of closeness of the length to the SML (Δl) is defined by users as a value which is proportional to the length. This is to take into account the variability of the lengths in biological tandem repeats. Second, we scan the analyzed sequence and do not consider a downstream **EL** of the neighbouring SS except for those which length correspond to one chosen SML (weeded ELs are in grey in Fig. 2). After this, we recalculate lengths and store them. The scanning is repeated for each of the SMLs and leads to K new sets of recalculated lengths. If the most frequent recalculated length differs from the current SML of a given set, it will be considered as a new SML ($SML_1 = SML_2 = 16$ in Fig. 2) and the weed is applied to the next SML. This operation ends when all K SMLs are tested. The original SML remains unchanged when it occurs more frequently than the recalculated lengths or if SMLs and the recalculation lengths are equal to each other. Steps 1 and 2 are performed one by one for all types of SSs of the analyzed sequence.

(3) Finally, K-means algorithm is simultaneous applied to all SMLs of all type of short strings. This operation provides K most frequent SMLs that can be considered as candidates for the real repeat lengths of the analysed sequence.

2.3 Contiguity filtering

An array of tandem repeats is defined as at least two adjacent copies having similar lengths. To take into account the contiguity of the repeats, we select SSs whose *SMLs* are equal or close (within user-defined Δl) to the most frequent SML of a given cluster created after the last step of K-means algorithm. Then, we scan the sequence by considering only these SSs and by looking for sequential repetitions of approximately equal lengths that we define as 'runs' of SMLs. Regions containing these 'runs' may represent tandem repeats. This procedure is redone one by one for all K clusters of SMLs. According to the locations and values of SML found in runs, hypothetical repeats are identified.

2.4 Extension and bridging of runs

Runs with the same SML can be interrupted by a region that is 2 or more times larger than this SML. To clarify the question, whether this region together with two flunking runs belongs to one tandem repeat or not we apply the following procedure. First, we divide the sequence downstream of the run into strings of length equals to the SMLs identified for the current cluster. Then each string joins the preceding run one by one in their order of appearance in the sequence if it contains half or more occurrences of short strings whose SMLs are present in the current cluster. In a special case of two-residue repeats, when the repeat length is equal to the SS length, we consider a string as a part of the run if it has at least one common residue with the SS. We add strings to the preceding runs until we arrive to the next run and by doing so we 'bridge' two runs with the same SMLs. This process starts after each run and stops at any string that has unacceptable SS composition.

2.5 Similarity filtering

The final step of the program is to evaluate the level of sequence similarity between the putative repeats of each run by using Multiple Sequence Alignement (MSA) approaches. For this purpose we used a combination of three MSA programs, a build-in program based on 'center-star algorithm', and two external programs CLUSTALW (Thompson *et al*., 1994) and MUSCLE (Edgar, 2004).

Based on the obtained MSA of the repeats constituting the runs, we deduce a consensus sequence and subsequently use it as a reference for similarity calculation. Let us consider an alignment made by m repeats of length l. In this alignment an indel is considered as an additional 21st type of amino acid residues. We calculate D_i that is a *Hamming distance* (Hamming, 1950) between the consensus sequence and a repeat R_i with $1 \le i \le m$. Then, we define a similarity coefficient for the whole alignment as $P_{\text{sim}} = (N - \sum_{i=1}^{m} D_i)/N$ with $N = m \times l$ and $0 \le P_{sim} \le 1$. Our program allows to select the runs which have the repeat similarities higher than a user-defined P_{sim} threshold (elsewhere noted as $P*_{\text{sim}}$).

Sometimes, strings located at the extremities of the alignment can diminish the total level of the alignment similarity. To solve this problem, we apply an additional trimming to alignments with similarities not exceeding the $P*_{\text{sim}}$. The operation eliminates the extreme strings one by one until it passes the threshold.

If a run meets all the criteria defined above, it will be considered as a tandem repeat.

2.6 Sequence databases for tests of T-REKS and the other programs

During debugging and tests of T-REKS we used two tandem repeat databases. First, we downloaded from TRIPS website (Katti *et al*., 2000) a database of tandem repeats detected by a sliding window technique and empirical mismatch levels from Swissprot Release of July 1999. The tandem repeats of TRIPS contain only residue substitutions and not indels. We extracted from this database all the data except homorepeats that brought the number of sequences to 890. Second, we generated artificial databanks of 1000 amino acid sequences each of them 1000 residue long. These sequences contained tandem repeats with *a priori* known features. For this, we used a built-in Java linear congruential generator to produce random sequences from an alphabet of 20 amino acids. Then, we inserted one array of perfect tandem repeats in each of these random sequences. The inserted tandem repeats were different between each other having variable repeat lengths (from 2 to 21), number of copies (from 2 to 20). Then we randomly mutated the perfect repeats by substitutions of amino acid residues or by introduction of indels (either insertion or deletion). One original sequence with a perfect tandem repeat yielded a set of sequences with the similarity levels between a user defined P_{sim} value and 1. The generated repeats were then aligned, their similarity level was calculated and if it was over $P*_{\mathrm{sim}}$ the tandem repeat was removed. Following the described procedure, we generated nine databanks with different similarity levels ($P_{\mathrm{sim}} \geq 0.50$; $P_{\mathrm{sim}} \geq 0.55$; $P_{\mathrm{sim}} \geq 0.60$; $P_{\mathrm{sim}} \geq 0.65$; $P_{\mathrm{sim}} \geq 0.70$; $P_{\mathrm{sim}} \geq 0.75$; $P_{\mathrm{sim}} \geq 0.80$; $P_{\mathrm{sim}} \geq 0.85$ and $P_{\mathrm{sim}} \geq 0.90$) and stored them in Generated Repeat Databanks (GRD) from GRD50 to GRD90 correspondingly. Each sequence in the GRDs contains one tandem repeat which is flanked by the randomized sequences. None of the tested programs (see Section 3.2) found tandem repeats within the random sequences of the GRDs. Positive hits were always located within the inserted tandem repeats. Therefore, during the tests, number of the hits was counted as number of sequences with the identified repeats.

To control the level of false positive results, we also generated a databank of 890 random sequences (the same number of sequences and characters as in TRIPS). The first random protein sequence was obtained by using the RandSeq tool from Expasy (Gasteiger *et al*., 2003) with the average amino acid composition of SwissProt. Then, this random sequence has been shuffled by the ShuffleSeq tool from EMBOSS (Rice *et al*., 2000) to obtain the other random sequences. To draw a line between true and false positive results we used the following procedure. Frequently, the random sequences also contain short runs of tandem repeats. We assume that the number of perfect tandem repeats X found by chance in a random sequence database follows a binomial distribution $X \sim B(n,p)$ approximated by a Poisson Distribution with parameter $\lambda = np$ where $p = (1/20)^{l*(m-1)}$, $n = 1.5 \times 10^8$ residues (size of the SwissProt database) and $l \times m$ is the total length of the tandem repeat region. Based on this approximation, occurrences of tandem repeats with $l \times m$ equal or longer than 14 residues is $\lambda = 0.1$ and the probability not to find such repeats $P(X = 0)$ is 0.896 (Fig. 3). Therefore, we considered 14 residues as the minimal length of the true repeat run with potential biological meaning. Within T-REKS, homorepeats are treated differently, so we found it reasonable to fix their minimal length separately. In accordance with our calculations [$\lambda = 0.005$ and $P(X = 0) = 0.994$], it is equal to 9 residues.

3 RESULTS AND DISCUSSION

3.1 Performance of T-REKS depending on its parameters and options

T-REKS was tested against the GRDs having different levels of similarity P_{sim}. Figure 4a shows decrease of the number of the undetected repeats with the increase of the repeat similarity. During this test the allowed similarity threshold $P*_{\mathrm{sim}}$ of T-REKS was set to 0.7. Another test of T-REKS against the random sequence databank by varying its $P*_{\mathrm{sim}}$ threshold showed that the number of false positive results (that are hits of total length more than 14 residues) drastically drops to 0 when $P*_{\mathrm{sim}}$ becomes equal or more than 0.7 (Fig. 4b). We selected $P*_{\mathrm{sim}} \geq 0.7$ as the default value for the web version of T-REKS.

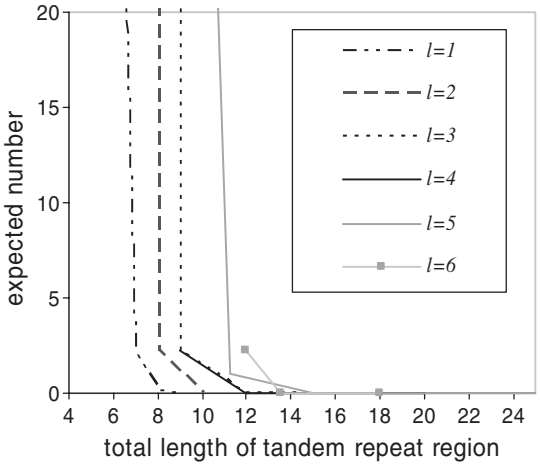


Fig. 3. Distribution of expected number (λ) of tandem repeats with different l found in a random protein sequence of 1.5×10^8 residues depending on total length of the repetitive region.

The test also revealed that at a given $P*_{\mathrm{sim}}$ value of the program a MSA mode which uses one by one external programs CLUSTALW and MUSCLE find the biggest number of tandem repeats.

3.2 T-REKS performance compared with the other programs

Tests of the other existing programs for identification of protein repeats, such as TRED, INTREP and XSTREAM demonstrated that they, similarly to T-REKS (Fig. 4) face difficulties in correct determination of repeats with the decrease of the repeat similarity level. At the same time, all four programs passed successfully the test against the random sequence databank, since no false positives repeats could be detected.

To benchmark T-REKS and these programs we used databank of repeats TRIPS and SwissProt (Release of January, 2009) taken from the repository of NCBI. It is worth mentioning that the comparison of the programs is complicated by differences in the tandem repeat definitions. T-REKS and XSTREAM have the closest definitions. To match the definitions given by these two programs we set the similarity filtering parameters $P*_{\mathrm{sim}}$ of T-REKS and I of XSTREAM to 0.7. The minimal total length of tandem repeat has been set to 14 residues for both T-REKS and XSTREAM. Our tests show that T-REKS finds more tandem repeats in protein sequences than other tested programs (Table 1). (Supplemental Data show examples of the tandem repeats of TRIPS databank that were found only by T-REKS and not by XSTREAM which is the most rapid program with similar definition of the repeats). At the same time, the performance of T-REKS was one of the most rapid. Although XSTREAM is faster than T-REKS, speed is becoming an uncritical issue when both programs need only minutes to analyse one genome. For example, T-REKS needs 17 min to analyse a medium size

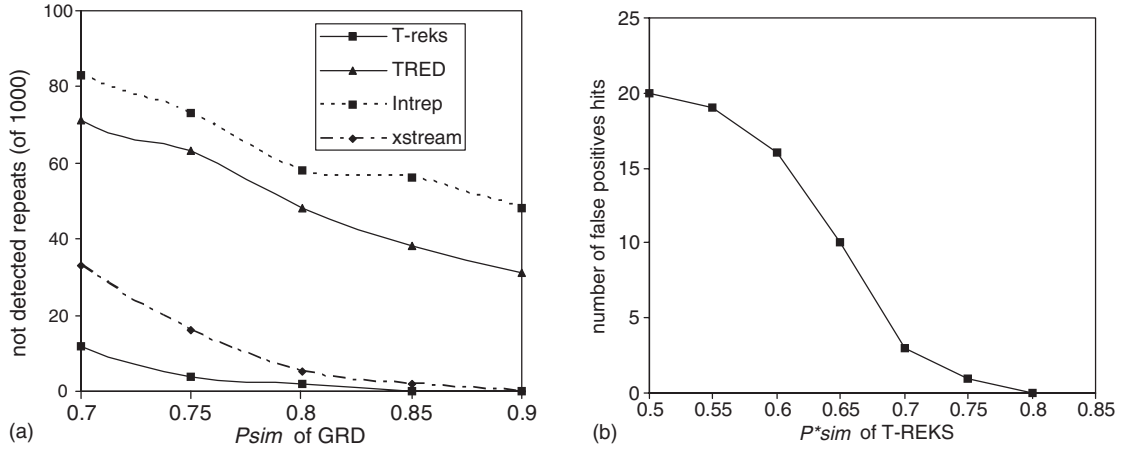


Fig. 4. (**a**) Number of sequences not detected by programs among the 1000 ones of each GRD. $P*_{sim}$ = 0.7 for T-REKS and I = 0.7 for XSTREAM. (**b**) Number of false positive sequences found in the Random sequence databank by T-REKS depending on its $P*_{sim}$ treshold.

Table 1. Benchmark of T-REKS, INTREP, TRED and XSTREAM programs executed on two databanks of protein sequences

	TRIPS (890 sequences with tandem repeats)		SWISSPROT (356 232 sequences)	
	Sequences identified[a]	Execution time	Sequences identified[a]	Execution time
T-REKS[b]	888	4 min	33 780	4 h 30 min
INTREP[c]	863	25 min	20 607[d]	23 h
TRED[e]	865	4 min	15 274	13 h 25 min
XSTREAM[f]	872	50 s	15 204	25 min

Benchmark has been performed with a Personal Computer Pentium 4 3.00 GHz and 2 Gb of RAM.

[a] Sometimes, the number of identified tandem repeats exceeds the number of sequences due to ability of programs to find several tandem repeats in the same sequence.

[b] This work, tandem repeats are defined as having the minimal length of 14 residues (nine residues for homorepeats) and $P_{sim} \geq 0.70$.

[c] Marcotte *et al.* (1999) repeats are defined as having P-value $< 10^{-3}$ (default value).

[d] INTREP results include both tandem and interspersed repeats.

[e] Sokol and Benson (2007) to match with T-REKS, 14 residue minimal length of tandem repeat region was selected, other parameters were set to their default values.

[f] Newman and Cooper (2007) to match the definition of tandem repeats with T-REKS the following parameters were used: $I = i = 0.7$, $minP = 1$, $minC = 2$, $mind = 14$, sort seed length = 2.

genome of *Drosophila melanogaster* (by using a Personal Computer Pentium 4 3.00 GHz and 2 Gb of RAM), while XSTREAM uses 2 min.

3.3 Implementation

T-REKS has been implemented in Java language and has a built-in GUI to allow the user to set parameters needed for the identification of tandem repeats. Although, T-REKS is tuned to explore protein sequences, the same version of the program can also be used for nucleotidic sequences or ones based on another alphabet. A standalone version of the program can be downloaded from our web page (http://bioinfo.montp.cnrs.fr/?r=t-reks). The program can also be used through a web interface at the same webpage. The web version, in addition to the basic features, is adjusted to treat large-scale protein sequence databanks such as Swissprot, NR or PDB taken from the NCBI repository.

Several parameters and options of the program can be defined by users. Among them are: Δl—allowed percentage of length variability (default value, which is fixed in the web version, is equal to 20% of l. It was chosen based on the analysis of the known repeats of biological importance.), $P*_{sim}$—similarity threshold (default value is equal to 0.7) and an option to allow/disallow overlaps of different tandem repeats that can be detected in the same region of a given sequence. In the case of the overlapping tandem repeat regions, priority was given to the longer one with the higher P_{sim}. In the standalone version, it is possible to choose between three MSA modes:

(1) a built-in mode which uses a 'center-star' algorithm,

(2) an external mode that uses two external programs CLUSTALW and MUSCLE one after another or

(3) a hybrid mode which uses in sequential order the 'center-star', CLUSTALW and MUSCLE programs.

The build-in MSA mode 1 was developed to make the standalone version of T-REKS more convenient for downloading. The combination of CLUSTALW and MUCLE yields the best alignment results and is the only mode available in the web version to favour accuracy at the expense of rapidity. This MSA mode 2 is applied to obtain the most reliable results for our protein repeat database called PRDB (see next section). The hybrid mode 3 represents the most optimal version in terms of accuracy and rapidity.

3.4 Protein Repeats DataBase (PRDB)

T-REKS output is used to fill our PRDB. A pilot version of PRDB can be found on our website http://bioinfo.montp.cnrs.fr/?r=repeatDB. At this moment, the PRDB contains about 1105 entries of tandem repeats from PDB, 50 789 from Swissprot. The information about the identified repeats is displayed in a table with characteristics such as organism, repeat length, number of copies, level of the repeat similarity, consensus sequence, position in the sequence, subcellular

Table 2. Comparison of repeats found by our program and Tandem Repeats Finder in the Human Frataxin gene intron 1

T-REKS[a]/TRF[b]			
Start	End	Copy length	Copy number
163/–	**188/–**	**12/ –**	**2/ –**
822/822	856/854	7/14	5/2.4
1786/1787	1912/1874	44/44	3/2
2167/–	**2184/–**	**1/–**	**18/–**
2185/2183	2210/2211	3/3	9/9.7
2387/–	**2410/–**	**6/–**	**3/–**

Additional repeats identified by T-REKS are indicated in bold.
[a]This work.
[b]Benson (1999).

localization and complete alignment of the copies. It allows users to choose repeats based on their organism, the consensus pattern, amino acid composition, tendency to be unstructured and the other parameters. We plan to improve, regularly update and maintain this database.

3.5 Application of T-REKS to the identification of tandem repeats in DNA sequences

T-REKS can, in principle, be also applied to nucleotide sequences because its algorithm is not linked to any particular alphabet. However, depending on the number of the characters in the alphabet, T-REKS may require optimization of certain parameters. The probability to find tandem repeats by chance in DNA sequences is higher than in protein sequences with correspondingly 4- and 20-letter alphabet. For example, comparing random 'Bernoulli' DNA and protein sequences of the same size, the expected (mean) number of repeats of l-residue length, copied m-times in DNA is $(20/4)*l*(m-1)$ times higher than in proteins. Thus, DNA sequences have a higher background noise in the distribution of the SS lengths and this can hamper the tandem repeats detection. To improve performance of T-REKS on DNA sequences we increase the length of SS to four residues. This modification reduced the gap between the expected values of tandem repeats found by chance in proteins and DNA. In addition, the number of clusters, K, was increased from 10 (for proteins) to 20 (for DNA). Similarly to protein sequences, T-REKS splits the DNA sequences that are longer than 1500 residues before and concatenates after the analysis. After these modifications, T-REKS yields better results on DNA sequences than its version with protein-specific parameters. For example, our tests of T-REKS on the Human Frataxin Gene (Friedreich's ataxia) intron 1 sequence, showed that it detects not only tandem repeats previously obtained by Tandem Repeats Finder (TRF) (Benson, 1999) but also some additional repeats (Table 2). Thus, T-REKS can also be used for detection of tandem repeats in DNA sequence databases. We are aware that DNA sequences of 1500 nucleotides may, in principle, contain more than 20 different lengths of tandem repeats (maximum number of clusters at $K = 20$) and that the splitting step may lead to the failure to detect some repeats. More accurate tests and optimization of the T-REKS parameters for DNA is a subject of our further studies.

4 CONCLUSIONS

In this article, we described a new program for *ab initio* identification of tandem repeats in protein sequences called T-REKS. It is based on K-means clustering of putative lengths of tandem repeats. T-REKS finds more tandem repeats in protein sequences than other tested programs. At the same time, it demonstrates one of the most rapid performances. Thus, this approach is well-suited for large scale analysis of tandem repeats.

T-REKS has been developped in a dual mode: a standalone mode with a user-friendly graphical interface for local use, and a Web-interface version. The latter version of the program is connected to our webserver and is able to store the results in a database of protein tandem repeats called PRDB. Both versions of T-REKS and the database PRDB are available to public via our webpage at http://bioinfo.montp.cnrs.fr/?r=t-reks. We plan to use this database for systematic large scale analysis of protein tandem repeats in genomes in order to obtain a global view on the structure, function and evolution of these motifs. T-REKS can be also used for detection of tandem repeats in DNA. Its further optimization for DNA sequences will be a subject of our future studies.

ACKNOWLEDGEMENTS

The authors thank Dr M. Anisimova, Dr J. Arunachalam and Dr S.A. Kondratov for critical reading of the manuscript and suggestions.

Funding: Ministère de l'Education Nationale, de la Recherche et de la Technologie (MENRT) grant to J.J.

Conflict of Interest: none declared.

REFERENCES

Andrade,M.A. *et al*. (2000) Homology-based method for identification of protein repeats using statistical significance estimates. *J. Mol. Biol.,* **298**, 521–537.

Baxa,U., *et al*. (2006) Structure, function, and amyloidogenesis of fungal prions: filament polymorphism and prion variants. *Adv. Protein Chem.*, **73**, 125–180.

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

Delgrange,O. and Rivals,E. (2004) STAR: an algorithm to search for Tandem Approximate Repeats. *Bioinformatics*, **20**, 2812–2820.

Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.,* **32**, 1792–1797.

Gasteiger,E. *et al*. (2003) ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res*., **31**, 3784–3788.

George,R.A. and Heringa,J. (2000) The REPRO server: finding protein internal sequence repeats through the Web. *Trends Biochem Sci*., **25**, 515–517.

Gribskov,M. *et al*. (1987) Profile analysis: detection of distantly related proteins. *Proc. Natl Acad. Sci. USA*, **84**, 4355–4358.

Hamming,R. (1950) Error detecting and error correcting codes. *Bell System Technical J.*, **29**, 147–160.

Heger,A. and Holm,L. (2000) Rapid automatic detection and alignment of repeats in protein sequences. *Proteins*, **41**, 224–237.

Johnson,S. (1967) Hierarchical clustering schemes. *Psychometrika*, **32**, 241–241.

Kajava,A.V. *et al*. (2006) Beta-structures in fibrous proteins. *Adv. Protein Chem.*, **73**, 1–15.

Kajava,A.V. and Steven,A.C. (2006) The turn of the screw: variations of the abundant beta-solenoid motif in passenger domains of Type V secretory proteins. *J. Struct. Biol.*, **155**, 306–315.

Kajava,A.V. *et al*. (1995) Modeling of the three-dimensional structure of proteins with the typical leucine-rich repeats. *Structure*, **3**, 867–877.

Katti,M.V. *et al*. (2000) Amino acid repeat patterns in protein sequences: their diversity and structural-functional implications. *Protein Sci*., **9**, 1203–1209.

Kolpakov,R. *et al*. (2003) mreps: efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Res*., **31**, 3672–3678.

Landau,G.M. *et al*. (2001) An algorithm for approximate tandem repeats. *J. Comput. Biol.*, **8**, 1–18.

Lupas,A. *et al*. (1997) A repetitive sequence in subunits of the 26S proteasome and 20S cyclosome (anaphase-promoting complex). *Trends Biochem Sci.*, **22**, 195–196.

MacQueen (1967) Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, Berkeley.

Marcotte,E.M. *et al*. (1999) A census of protein repeats. *J. Mol. Biol.*, **293**, 151–160.

Nelson,R. and Eisenberg,D. (2006) Structural models of amyloid-like fibrils. *Adv. Protein Chem.*, **73**, 235–282.

Newman,A.M. and Cooper,J.B. (2007) XSTREAM: a practical algorithm for identification and architecture modeling of tandem repeats in protein sequences. *BMC Bioinformatics*, **8**, 382.

Rice,P. *et al*. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.*, **16**, 276–277.

Sokol,D. *et al*. (2007) Tandem repeats over the edit distance. *Bioinformatics*, **23**, e30–e35.

Szklarczyk,R. and Heringa,J. (2004) Tracking repeats using significance and transitivity. *Bioinformatics*, **20**(Suppl. 1), i311–i317.

Thompson,J.D. *et al*. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

CHAPITRE 2

Construction de PRDB : Protein tandem Repeat DataBase

Résumé

Les nouvelles disciplines telles que la génomique comparative ont apporté des approches puissantes d'analyse des génomes, permettant de manipuler de grands volumes de données et d'obtenir une vision globale de leurs caractéristiques. Afin d'étudier à grande échelle les répétitions en tandem dans les protéines, nous avons développé une base de données du nom de PRDB, regroupant les résultats issus de l'exploration des banques de données par T-REKS. Cette base comporte aussi plusieurs outils facilitant l'analyse par l'utilisateur et délivre les informations à travers une interface web ergonomique. Ce chapitre propose une explication détaillée du contenu et du fonctionnement de PRDB. En outre, quelques exemples des applications intégrées à cette base sont explicités, démontrant que PRDB n'est pas simplement une base de données mais une plate-forme de choix quant à l'analyse de répétitions en tandem dans les séquences protéiques.

PRDB est ouvert au public et disponible à l'adresse suivante :

http://bioinfo.montp.cnrs.fr/?r=repeatDB

INTRODUCTION

Jusqu'à présent, la plupart des analyses de répétitions en tandem n'ont couvert que certains types de répétitions et de protéines, ou un nombre limité d'organismes et d'espèces (Katti et al., 2001; Huntley and Golding, 2002; Karlin et al., 2002). Le volume croissant des données issues des projets de séquençage, et les divers exemples d'implication des répétitions en tandem dans des fonctions importantes nous ont poussés à les identifier exhaustivement dans les plus grandes banques de données de séquences. T-REKS identifie plus efficacement les répétitions en tandem que les autres programmes et se positionne parmi les meilleures performances en temps d'exécution. Toutefois, lorsque des millions de séquences doivent être traitées, il est essentiel d'organiser les résultats sous une forme plus exploitable que celle proposée par T-REKS. En outre, il existe en général une carence en ressources traitant des répétitions en tandem dans les protéines, la plupart des bases de données existantes étant construites à partir de données d'ADN génomique (Le Fleche et al., 2001; Boby et al., 2005; Gelfand et al., 2007)et/ou peu adaptées à l'analyse spécifique des répétitions en tandem (Kalita et al., 2006). Pour faire face à ce besoin, nous avons crée la Protein Repeats Database (PRDB), en vue d'étudier les répétitions protéiques à travers les génomes.

Nous décrivons ici la base de données PRDB qui est régulièrement mise à jour et disponible au public sur http://bioinfo.montp.cnrs.fr/?r=repeatDB

IMPLEMENTATION ET RESULTATS

Construction de la base de données

PRDB est une base de données relationnelle construite sur le système de gestion de base de données PostgreSQL, elle regroupe les informations sur les répétitions en tandem (TR) identifiées par T-REKS et les séquences protéiques qui les contiennent (notées RCP).

Trois banques de données de référence ont été intégrées dans PRDB à partir du serveur ftp du NCBI:

(i) Non-redundant de NCBI (NR), qui est l'une des plus complète, constituée des traductions non redondante CDS Genbank, Swiss-Prot et APB,

(ii) Swiss-Prot, une banque de données hautement annotée disponible sur le site Web d' Expasy (Bairoch and Apweiler, 2000). Elle présente une redondance minimale et un haut niveau d'intégration avec les autres bases de données.

(iii) la Protein DataBank (PDB), une base de données de structures tridimensionnelles de macromolécules biologiques déterminées expérimentalement (Berman et al., 2000).

Lors de l'exploration de ces trois banques par T-REKS(Jorda and Kajava, 2009), plusieurs paramètres ont été définis :

Le premier paramètre est le coefficient de similarité *Psim* définit précédemment(Jorda and Kajava, 2009) pour évaluer le niveau de perfection avec lequel les TRs sont répétés. Ce coefficient varie avec le nombre d'événements mutationnels intervenus sur la séquence : un TR parfait a un *Psim* égal à 1. Après plusieurs bancs d'essais et approches statistiques estimant la significativité des résultats, nous avons défini le seuil minimum du *Psim* à 0.7, éliminant les résultats pour lesquels le *Psim* sera inférieur à cette valeur.

Dans un deuxième temps, le pourcentage de variabilité de longueur des répétitions due aux événements d'« indels » a été fixé à 20%.

La méthode d'alignement multiple utilisée pour aligner les motifs unitaires a été définie en « external mode ». Il combine deux programmes externes : MUSCLE(Edgar, 2004)et CLUSTALW(Thompson et al., 1994), leur utilisation concomitante donnant des résultats plus fiables.

Enfin, les chevauchements de TRs identifiés dans la même région d'une séquence sont filtrés par une option disponible dans T-REKS.

Caractéristiques des séquences des TRs et RCPs

T-REKS peut accéder à toutes les séquences des banques de données importées dans PRDB, puis y stocker les résultats après leur exploration. En plus de leur identification, les caractéristiques des TRs et RCPs sont introduites dans PRDB afin de les utiliser comme critères de filtrage (figure 22). Une observation des TRs permet de déduire les premiers critères évidents: le nombre de copies, la longueur du motif unitaire, la longueur totale de la région du TR, la position des premiers et derniers résidus des TRs, l'alignement multiple des copies et le motif consensus dérivé de cet alignement multiple.

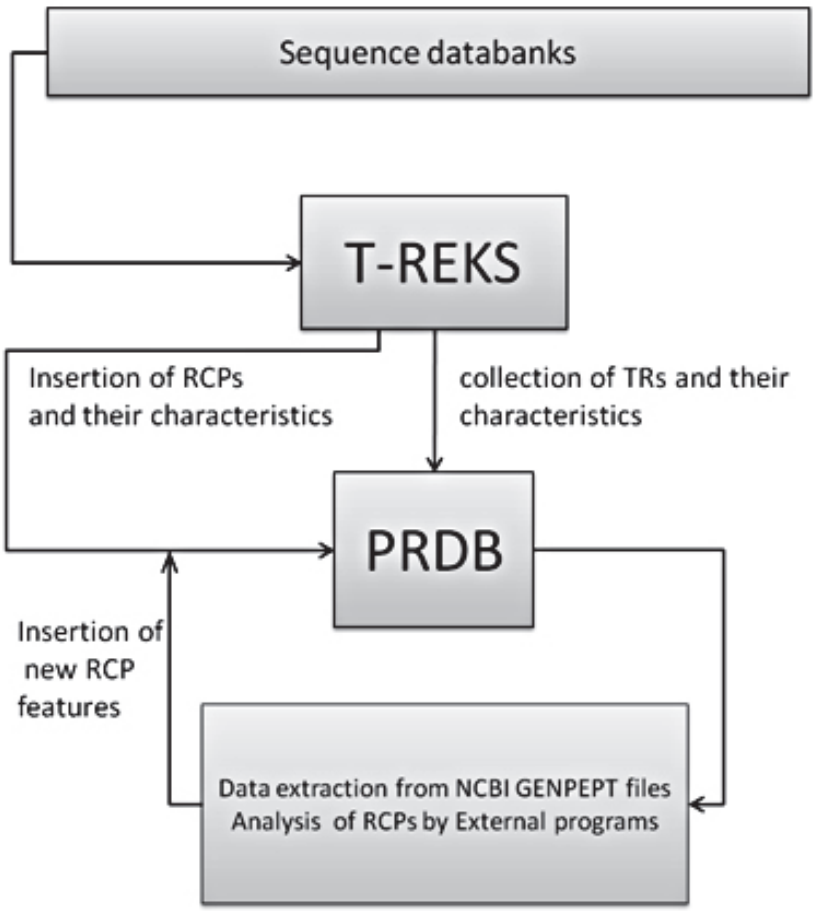


Figure 22. Organigramme de la procédure de remplissage de PRDB.

Par la suite, les caractéristiques liées aux RCPs ont été rajoutées à PRDB. Elles se composent de la longueur de la séquence, la banque source dont elles sont issues, et l'identifiant GI de NCBI.

Caractéristiques provenant d'autres annotations

Ces informations proviennent des données annotées dans la base de données GenPept disponible sur la plate-forme web du NCBI, et sont obtenus par l'intermédiaire de l'utilitaire Entrez "Efetch". Efetch permet de récupérer l'ensemble des données associées à une liste d'identifiants de séquence définis par le NCBI, annotés « GI number ». La première caractéristique est le nom des organismes, ceux-ci ont été classés par règnes après leur insertion dans PRDB. Le règne eucaryote a été divisé en plusieurs sous-règnes prédéfinis, organisant les organismes dans un arbre taxonomique simple.

Les données dans GenPept sont très variables et dépendent de la qualité des annotations de la base de données source. En effet, certaines séquences comportent beaucoup plus

d'informations que d'autres. En conséquence, les informations complémentaires ont été récupérées pour les RCPs de Swiss-Prot, estimant que cette banque de données est plus annotée que PDB ou NR. Elles comprennent les caractéristiques des protéines telles que les mots-clés (utilisés lors d'une recherche rapide sur NCBI), leurs localisations subcellulaires et leurs fonctions moléculaires. Les mots-clés et les localisations ont été directement insérés dans PRDB tandis que les fonctions moléculaires, qui sont annotés avec leur identifiant Gene Ontology(Ashburner et al., 2000) ont nécessité une étape préalable de récupération des données complètes. La structure et les données concernant les fonctions moléculaires de la base Gene Ontology ont été intégrées à PRDB et liées aux identifiants GO annotés dans les fichiers GenPept. Les localisations subcellulaires et les fonctions moléculaires ont été classées dans des familles, afin de hiérarchiser l'ensemble des données et de simplifier leur accès via une interface.

Caractéristiques issues de l'analyse des séquences par des programmes externes

Des caractéristiques supplémentaires ont été générées à partir de l'analyse des TRs et des RCPs par d'autres programmes. Par exemple, le potentiel de formation de structure des régions répétées(« structure forming potential ») a pu être évalué en se basant sur l'échelle de *TOP-IDP*(Campen et al., 2008) qui mesure la propension d'une séquence à être intrinsèquement désordonnée. Ainsi, un TR avec une valeur de *TOP-IDP* négative est considéré comme déstructuré, une valeur positive indique quant à elle que le TR est structuré. Une autre caractéristique des séquences protéiques provient de l'identification de leurs domaines fonctionnels PFAM(Finn et al.; Finn et al., 2006). A cet effet, le programme HMMPFAM du package HMMER (Eddy, 2009) a été appliqué sur les RCPs de Swiss-Prot. Il compare un profil HMM avec la séquence étudiée, les résultats (ou matchs) ayant chacun un score et une E-value: plus le score est élevé et la E-value est faible, plus le match est fiable. Nous n'avons établi aucun seuil de score ou d'E-value limitant le nombre de matchs, ceci afin de maximiser les domaines PFAM potentiels associés aux RCPs. Toutefois, par souci de clarté des résultats, une requête classique sur PRDB ne renverra que des résultats dont la E-value est supérieure à 0,1. Les résultats ont été stockés dans PRDB et classés par groupes syntaxiques ou par la collection des « clans » PFAM(Finn et al., 2006).

INTERFACE ET VISUALISATION DES DONNEES

Le formulaire de requêtage

PRDB est interfacé par le biais d'une page Web, l'utilisateur peut alors parcourir les TRs et formuler des requêtes incluant des filtres sur les critères décrits ci-dessus (figure 23). L'interface propose des fonctionnalités de filtrage sur tous les critères quantitatifs (le nombre de répétitions, la longueur de l'unité répétée, la longueur de la région TR, le niveau de perfection, le potentiel de formation de structure et la longueur du RCP) et les critères qualitatifs (la banque de données source, le règne et l'organisme, motif consensus, identifiant GI, les mots-clés, la localisation subcellulaire, les fonctions moléculaires et les domaines PFAM). Plusieurs génomes ont été exhaustivement séquencés, générant des données multiples pour une même espèce. Par conséquent, une option appelée «Strain redundancy filter» a été créé afin d'éviter la redondance desTRs.

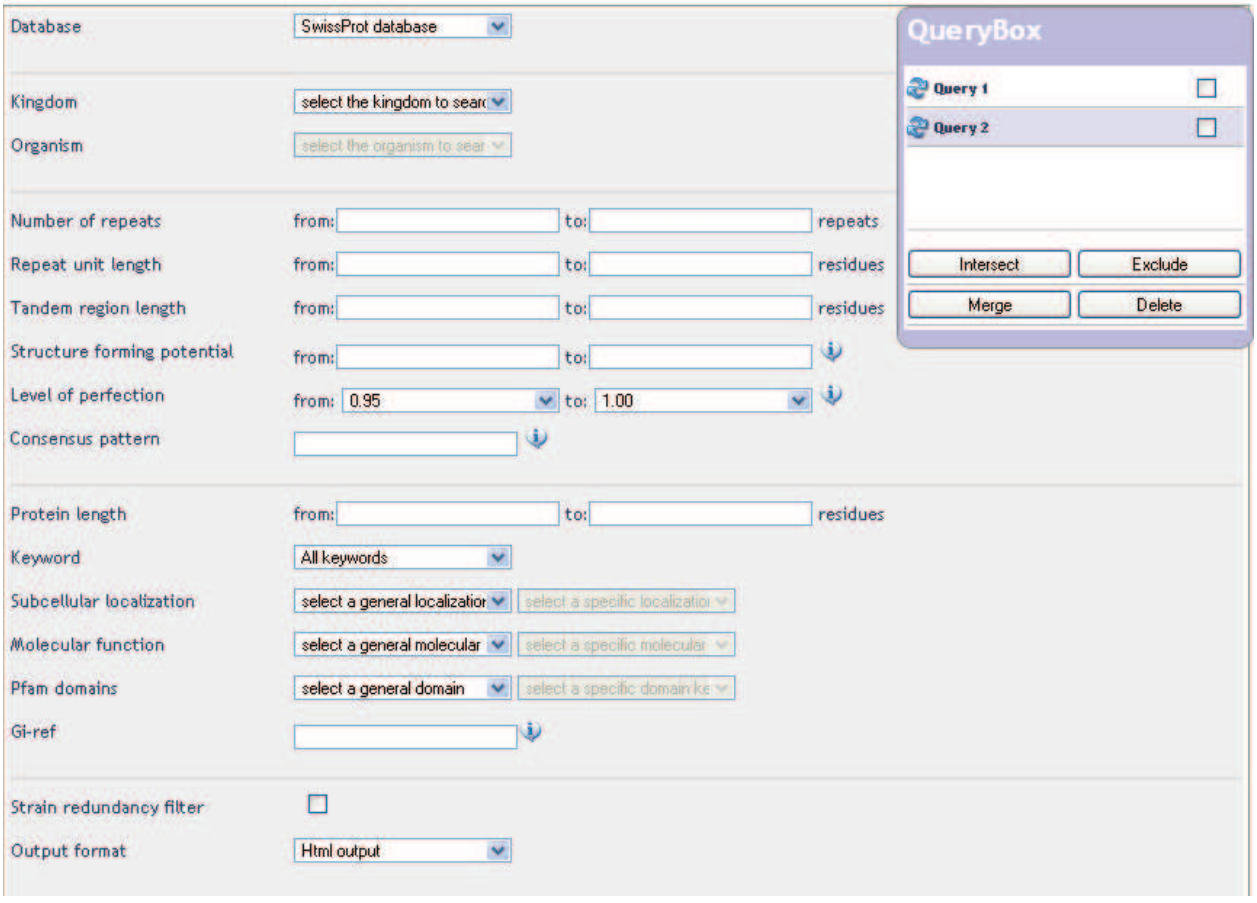


Figure 23. Vue de l'interface graphique de PRDB. Tous les champs du formulaire sont actifs exception faite des champs grisés, dépendant de leur champ « parent » (ici les champs « Organism », « Subcellular localization », « Molecular function » et «PFAM domains »). Le champ parent doit être défini préalablement pour autoriser la sélection des valeurs du champ fils.

Les résultats peuvent être récupérés sous trois formats, la valeur par défaut étant l'affichage des TRs dans une page web HTML, où les données sont visualisées dans une interface conviviale. D'autres formats ont été implémentés pour l'export des résultats dans des fichiers. PRDB propose le format FASTA, contenant les caractéristiques de base concernant la séquence du TR, et le format CSV dont les champs, définis par l'utilisateur, sont tabulés. La première ligne dans le fichier CSV est l'en-tête contenant les noms des champs tandis que les lignes suivantes sont les valeurs des TRs.

Page de résultats

Suite à la sélection du format HTML, les TRs résultants sont affichés dans un tableau récapitulant leurs principales caractéristiques (figure 24). En cliquant sur un champ de l'en-tête, le tableau est trié en fonction des critères sélectionnés. La localisation du TR dans la protéine est définie à partir des positions de son premier et dernier résidu, cette caractéristique a été simplifiée à l'affichage par trois caractères: 'N' pour N-terminal, «M» pour la région centrale et C pour C-terminal. Si le TR est situé dans deux ou toutes les régions, en conséquence la position sera une combinaison de 2 ou 3 caractères. Un système de pagination divise l'ensemble du jeu de données de TRs en pages, réduisant le nombre maximal de TR affichés dans le tableau à 400. Un lien sur les identifiants GI de chaque TR donne accès à la page Genpept de son RCP correspondant (voir la description de « Sequence Viewer »).

51101 tandem repeats

Page n°1 of 128 go!

Help page

Query 1 save

Get AA composition Statistics

select all

gi-ref	pattern	Psim	number	repeat length	position	prot. length	TOP-IDP	organism	kingdom	localization	alignment	details
41018019	P-LP	0.7	5	3	N	1406	-0.04	Yersinia pestis	Bacteria	unknown	See Alignment	
67473027	ELDAXXA	0.75	2	7	N	430	-0.00	Escherichia coli K-12	Bacteria	Cytoplasm (By similarity)	close Alignment	
ELDAAKA ELDALQA												
12644688	VLT-PX	0.722	3	5	N	286	0.050	Pseudomonas aeruginosa	Bacteria	Cell membrane	See Alignment	
66774161	IGLVX	0.733	3	5	N	451	0.188	Escherichia coli K-12	Bacteria	Cell inner membrane	See Alignment	
67468069	AEEQAKRE	0.75	7	8	N	890	-0.12	Escherichia coli K-12	Bacteria	Cytoplasm	See Alignment	
74834190	KSNQLXX	0.75	2	7	N	1738	-0.01	Dictyostelium discoideum	Eukaryota	Nucleus membrane	See Alignment	

Figure 24. La page de résultats au format HTML. Les caractéristiques sont affichées dans l'en-tête de chaque colonne, un clic sur une d'elles permet de trier les TRs en fonction.

L'analyse des caractéristiques des TRs résultant d'une requête peut être réalisée par les outils additionnels qui sont accessibles depuis la page de résultats.
Le premier outil donne la composition globale des acides aminés des TRs sous forme de graphique en «camembert», où les acides aminés sont classés et regroupés par couleurs en fonction de leurs propriétés physico-chimiques (Annexe I).

D'autres outils sont accessibles en cliquant sur le bouton "statistiques", incluant des graphes de distribution et les listes de critères principaux calculés à partir du jeu de données d'intérêt. Les critères quantitatifs (le nombre de répétitions, la longueur de l'unité répétée, la longueur de la région du TR, le niveau de perfection, le potentiel de formation de structure et la longueur du RCP) sont affichés dans des histogrammes tandis que les critères qualitatifs sont listés dans des tableaux et classés par leur fréquence, ou dans des graphes camembert.

La QueryBox

Lorsqu'une requête est soumise, elle peut être temporairement sauvegardée et être ainsi gérée a posteriori par la QueryBox. La QueryBox a été créée pour manipuler et combiner plusieurs jeux de données à travers trois opérateurs logiques :

Merge : renvoie tous les TRs des requêtes de sélectionnées dans une seule page de résultats.

Intersect : renvoie uniquement les TRs communs aux requêtes sélectionnées.

Exclude : affiche la différence symétrique des TRs issus des requêtes sélectionnées.

Sequence Viewer

La page de résultats ne recense qu'une partie des critères associés aux TRs, les informations relatives à leurs RCPs respectifs étant disponibles dans le module « Sequence Viewer ». Chaque TR dans la page de résultats propose un lien donnant accès au module. Sequence Viewer reprend l'ensemble des annotations et les informations supplémentaires du RCP en question, et génère une représentation graphique de la séquence avec l'ensemble de ses TRs identifiés. Les positions et les séquences des TRs sont visualisables sur une ligne modélisant la séquence du RCP (Annexe III). Pour un confort de visualisation, l'échelle de cette ligne peut être modifiée en temps réel. Les domaines PFAM identifiés dans le RCP par HMMPFAM sont disposés identiquement sous la séquence, disposant chacun d'un lien vers la description détaillée du domaine. D'autres zones cliquables existent, comme les régions répétées ou les régions liées à un domaine PFAM.

Par exemple, un clic sur un TR déclenche la génération d'un profil par la méthode d'utilisation de profils généralisés, cette méthode est mise en pratique par les programmes issus du package *pftools* (Bucher et al., 1996). Cette méthode crée un profil à partir de l'alignement multiple des copies du TR, ce même profil pourra ensuite être lancé contre les TRs de PRDB. En conséquence, tous les TRs détectés par le profil dont le score est supérieur à un seuil défini par l'utilisateur seront affichées dans Sequence viewer. D'une manière semblable, le clic sur une région avec un domaine PFAM permet à l'utilisateur d'afficher tous les RCPs présentant le même type de domaine, sous réserve que l'E-value associée dépasse un certain seuil. Ces deux outils sont utiles dans la comparaison inter-espèces des TRs dans les

RCPs, l'annotation de fonctions pour des protéines homologues non caractérisées, ou corréler des TRs spécifiques avec des domaines fonctionnels.

STATISTIQUES DES TRs

PRDB englobe environ 1 .25 millions de TRs, actuellement distribués parmi trois banques de données sources : autour de 1.2 millions pour NR (portés par 836670 RCPs), 52000 pour Swiss-Prot (33151 RCPs) et 100 pour PDB (45914 RCPs). La comparaison de la composition en acides aminés des TRs avec celle des séquences protéiques de Swiss-Prot (figure 25A) révèle une proportion accrue de résidus polaires (S, N, Q et E) et un appauvrissement en résidus apolaires (I, F, V et L) chez les TRs. Ces observations sont en accord avec les observations précédentes faites dans ce domaine (Simon and Hancock, 2009). De plus, les TRs présentent une abondance des résidus P, A et G. Ceci pourrait s'expliquer en partie par la présence d'homorepeats dans les données, car ils sont enrichis en ces résidus (Jorda and Kajava, 2010).

La figure 25B présente la fréquence pondérée des TRs et RCPs issus de Swiss-Prot, montrant que la prédisposition d'une séquence à contenir un TR augmente avec la longueur de la protéine. En effet, la fréquence des TRs portés par un RCP est relativement indépendante de la longueur de la protéine jusqu'à 500 résidus, tout comme la distribution de la fréquence des RCP dans les banques de données. Au delà de cette longueur, la probabilité de trouver un TR dans une protéine semble considérablement augmenter. Ces données peuvent être utiles à des applications futures, telles que des inférences statistiques permettant d'estimer la probabilité de la présence d'un TR dans une séquence.

Les diagrammes bâtons générés par PRDB dans la figure 26 montrent la distribution des TRs de Swiss-Prot en fonction de critères quantitatifs. Les Figures 26A et 26B exposent le fait qu'en général, la fréquence des TRs diminue quand la longueur ou le nombre de copies augmentent.

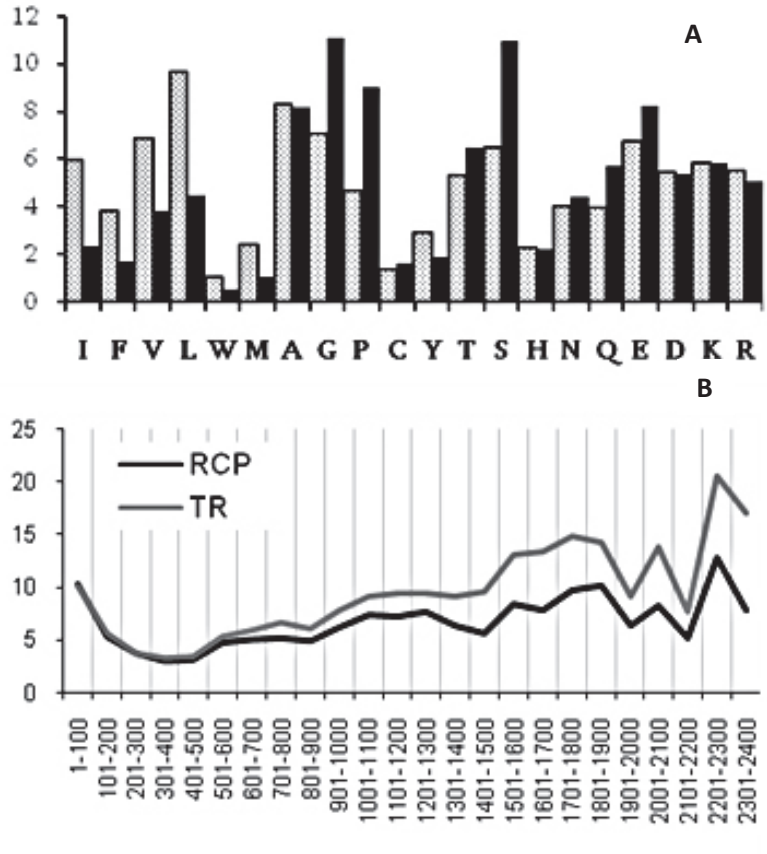


Figure 25. (A)Profil compositionnel des séquences protéiques (batons rayés) et des TRs (noir) issus de Swiss-Prot.(B) Distribution de la fréquence des TRs et des RCPs de Swiss-Prot (pondérés par le nombre de séquences et leurs longueurs) en fonction de la longueur de la protéine.

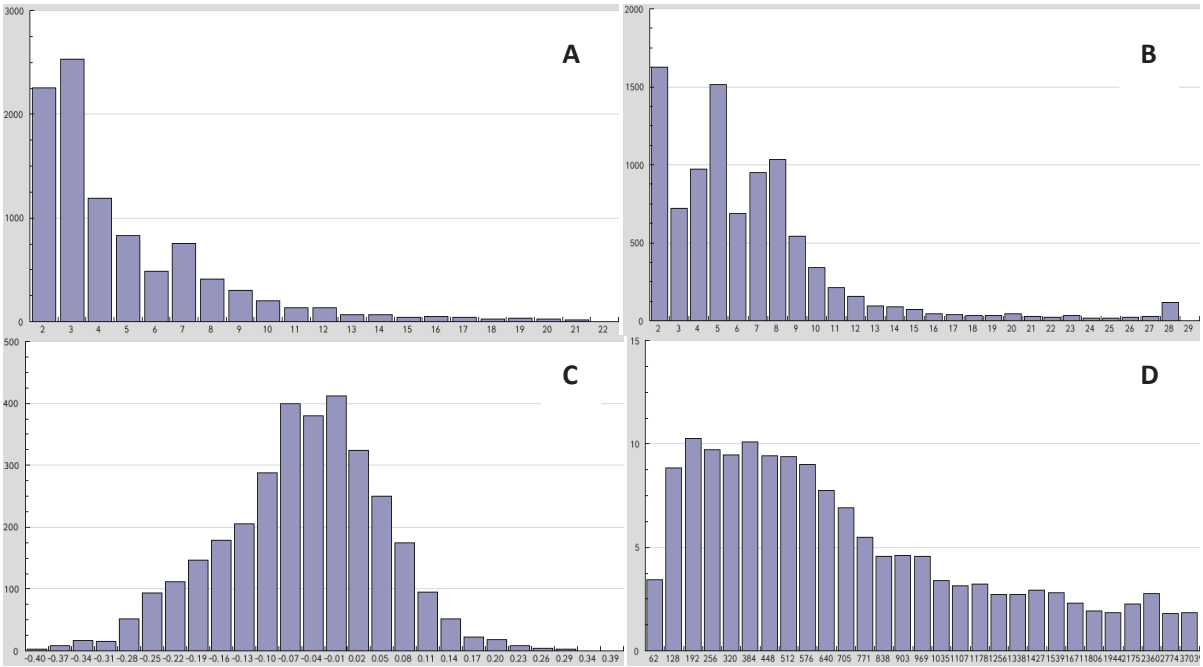


Figure 26. Distribution des TRs en fonction du nombre de copies(A), la longueur du motif unitaire (B), de l'échelle TOP-IDP(C) et la longueur de la séquence protéique(D). Le jeu de données utilisé correspond aux TRs de Swiss-Prot avec un Psim>=0.8 et exclue les homorepeats.

Cependant, dans la figure 26B on peut observer un biais pour les TRs dont la longueur est égale à 5 résidus. Ceci est du à la longueur minimale considérée lors de l'identification des TRs par T-REKS : En effet le seuil minimal étant de 14 résidus, le premier entier autorisé est donc 15, représenté soit par des homorepeats de cette même longueur (ceux-ci ont été exclus du jeu de données), soit par des tripeptides répétés 5 fois, soit par des TRs dont l'unité est longue de 5 résidus et répétée 3 fois, ce dernier exemple statistiquement plus fréquent menant au biais compositionnel décrit. De même, il est à noter que les TRs dont l'unité est longue de 28 résidus présentent une fréquence relativement élevée. Une analyse détaillée a pu révéler que la plupart étaient impliqués dans les domaines à doigts de zinc (Zinc-finger domain), où les 28 résidus correspondent à la longueur du motif structurel unitaire. Dans la figure 26C, la distribution des valeurs du TOP-IDP associées aux régions des TRs semble montrer que les TRs sont majoritairement désordonnées. Enfin, la distribution des TRs en fonction de la longueur de la protéine dans la figure 26D suit la même tendance que celle des séquences protéiques dans Swiss-Prot, signifiant que la distribution des TRs est indépendante de la longueur de la séquence protéique.

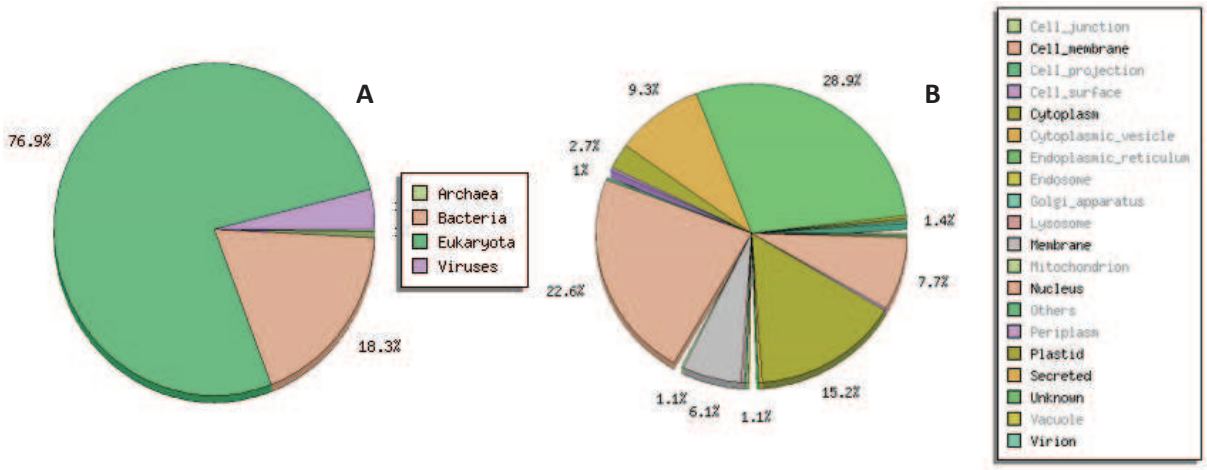


Figure 27. Répartition des TRs de Swiss-Prot en fonction des règnes et de leur localisation subcellulaire.

Les Figure 27A et 27B sont des représentations graphiques de la distribution de TRs respectivement en fonction des règnes et de leur localisation subcellulaire. La plupart des TR sont d'origine eucaryote dans PRDB, corroborant les observations antérieures (Marcotte et al., 1999) tandis que 29% des TRs n'ont aucune localisation associée dans la cellule. Les domaines PFAM listés dans la Table 5 indiquent que la majorité des TR sont associés à des protéines impliquées dans l'interaction avec les acides nucléiques, tels que les protéines

régulant l'expression des gènes, et les grands complexes impliqués dans la transcription ou la traduction. Ils abondent également dans les enzymes, en particulier chez les procaryotes qui représentent la moitié des domaines PFAM enzymatiques identifiés dans PRDB. La Table 6, qui comptabilise les fonctions moléculaires caractérisées dans les TRs de Swiss-Prot, confirment ces observations ainsi que celles précédemment proposées pour les homorepeats dans les protéines (Faux et al., 2007).

PFAM domains	Number of occurences
Enzymatic	2160
Repeat	1294
Unknown	1101
RNA related	540
Zinc finger	492
DNA related	412
Phosphate related	393
Transcription	328
ATP related	286
Virus	284
Ribosome related	281
Ribosome	281
Helix	247
Transport	200
Receptor	154
Membrane	153
Glycosyl related	100
Regulator	99
Inhibitor	99
Catalytic	98
Antigen	86
Activator	66
Synthesis related	65
Nucleus	61
Lipoprotein	47
Glycoprotein	44
Cytochrome	37
Coat	37
Toxin related	33
Tail	25
Tail related	25
Secretion	24
Non-structural	18
Mitochondrie	14
Replication	7

Table 5. Principaux domains PFAM associés aux RCPs et leurs nombres d'occurrences dans PRDB.

Molecular Functions	Number of occurences
Unknown	1658
protein binding	1205
ATP binding	964
zinc ion binding	909
DNA binding	836
RNA binding	554
transcription factor activity	345
calcium ion binding	284
structural constituent of ribosome	236
unfolded protein binding	233
nucleotide binding	232
structural molecule activity	219
GTP binding	218
sequence-specific DNA binding	211
translation initiation factor activity	208
GTPase activity	183

Table 6. Principales fonctions moléculaires associés aux RCPs et leurs nombres d'occurrences dans PRDB.

CHAPITRE 3

Caractérisation des propriétés physicochimiques, structurelles, évolutives et fonctionnelles des homorepeats

Résumé

Ce chapitre présente une synthèse sur les homorepeats qui s'appuie sur différents points : Tout d'abord leurs séquences, et plus particulièrement leurs biais compositionnels éventuels ainsi que leur distribution selon plusieurs critères (règnes, organismes, localisation subcellulaire). Une deuxième partie a consisté à décrire leurs propriétés structurelles, qu'ils soient synthétiques ou naturels. Une étude supplémentaire s'est axée sur leurs propriétés évolutives, tels que les mécanismes sous-jacents à leur génération ou encore les évènements mutationnels auxquels ils sont soumis. La dernière partie a eu pour objectif de regrouper les fonctions connues des protéines contenant les homorepeats, ainsi que de comprendre leur rôle dans la séquence même et leur influence sur la fonction de la protéine.

Ces travaux se sont basés principalement sur des articles scientifiques traitant des homorepeats, mais aussi sur nos propres observations et hypothèses émises sur les homorepeats de PRDB.

Cette étude est la première à tirer parti de la richesse des banques de données biologiques afin de caractériser de façon exhaustive les homorepeats dans les protéines. Elle a fait l'objet d'une revue :

REVIEW - Jorda J, Kajava AV.: **Protein homorepeats: sequences, structures, evolution, and functions.** Adv Protein Chem Struct Biol. 2010;79:59-88

PROTEIN HOMOREPEATS: SEQUENCES, STRUCTURES, EVOLUTION, AND FUNCTIONS

By JULIEN JORDA AND ANDREY V. KAJAVA

Centre de Recherches de Biochimie Macromoléculaire UMR 5237, CNRS, University of Montpellier 1 and 2, Montpellier, France

I. Introduction 60
II. Distribution of Homorepeats in Proteomes 62
A. What Is the Minimal Homorepeat Run That Should Be Taken into Consideration? 62
B. Homorepeats Are Predominantly Hydrophilic 63
C. Homorepeats Are Much More Prevalent in Eukaryotic Than in Viral or Prokaryotic Proteins 64
D. Frequencies and Types of Homorepeats Vary Greatly between Eukaryotes 66
E. Distribution and Localization of Homorepeats within Proteins 67
F. Protein Size and Tendency to Have Homorepeats 68
III. Structural Properties of Homorepeat Regions 70
A. Synthetic Homopeptides Form Crystallites 70
B. Living Organisms Avoid Structure-Forming Homorepeats 74
IV. Evolutionary Dynamics of Homorepeats 75
V. Functional Associations of Homorepeat Regions 76
A. Functions of Proteins Containing Homorepeat Regions 76
B. Do Homorepeats Represent Junk Protein Regions or Have Essential Functional Roles? 77
C. Preferences of Homorepeat Types for Certain Localizations and Functions 78
D. Functions Specifically Related to Homorepeat Regions 80
E. Function of a Homorepeat Region Strongly Depends on the Adjacent Domains and Subcellular Localization of the Protein 83
F. Homopeptides as Building Blocks of New Materials and Supramolecular Nanostructures 83
References 84

Abstract

The vast majority of protein sequences are aperiodic; they do not have any strong bias in the amino acid composition, and they use a subtle mixture of all or most of the 20 amino acid residues to code a great

Copyright 2010, Elsevier Inc.
All rights reserved.

number of various structures and functions. In this context, homorepeats, runs of a single amino acid residue, represent unusual, eye-catching motifs in proteins. Despite the sequence simplicity and relatively small size, the homorepeat runs have a strong potential for molecular interactions due to the excessively high local concentration of a certain physico-chemical property. Appearance of such runs within proteins may give them new structural and functional features. An increasing number of studies demonstrate the abundance of these motifs in proteins, their important roles in biological processes, and their link to a number of hereditary and age-related diseases. In this chapter, we summarize data on the distribution of homorepeats in proteomes and on their structural properties, evolution, and functions.

I. Introduction

Amino acid homorepeats, also called homopeptide repeats, single amino acid repeats, or amino acid runs, represent protein regions with reiterations of a single residue. The homorepeats look unusually simple compared to typical protein sequences which are characterized by high complexity and an absence of strong bias in the amino acid composition.

A series of studies suggests that these regions represent the protein equivalent of junk DNA sequences that mostly serve as linkers between functionally important domains (Huntley and Golding, 2002; Newfeld et al., 1994; Gatti et al., 1994; Sumiyama et al., 1996). However, it is becoming apparent that homorepeats play crucial roles in key biological processes. They tend to belong to specific functional classes of proteins such as transcription factors, protein kinases, and proteins of ion storage and exchange (Karlin and Burge, 1996). Variation of homorepeat length which modulate transcriptional activation can be the major source of phenotypic diversity (Gerber et al., 1994; Fondon and Garner, 2004). The runs of poly-A and poly-G (here and in the rest of the chapter the one-letter amino acid code is used) are indispensable to provide necessary mechanical properties to a number of extracellular structural proteins (Fraser and MacRae, 1973; Altman et al., 2003)

Homorepeat regions are also attracting considerable attention because of their links to a number of hereditary and age-related diseases. In particular, several poly-Q-containing proteins are associated

with inherited neurodegenerative disorders, including Huntington's disease, spinobulbar muscular atrophy, dentatorubral-pallidoluysian atrophy, and different forms of spinocerebellar ataxia (Orr and Zoghbi, 2007). Expansion of poly-A in polyadenine-binding protein 2 is associated with oculopharyngeal muscular dystrophy (Brais et al., 1998). Several other genes have been identified in which poly-A expansions may cause human diseases, including the Artistaless-related homeobox protein in X-linked mental retardation and epilepsy (Stromme et al., 2002) and the SRY-box 3 in X-linked mental retardation (Laumonnier et al., 2002). A growing number of studies suggest that homorepeats may have a broader role in human diseases than was previously recognized (Siwach and Ganesh, 2008). In accordance with the Online Mendelian Inheritance in Man (OMIM) annotation (Hamosh et al., 2002), more than 40% of proteins containing multiple runs of homorepeats are associated with different human diseases (Karlin et al., 2002).

Another unusual property of the homorepeat regions is their exceptionally high length polymorphism. This polymorphism is observed within populations of the same organisms, between different cell types within an organism, and even depends on the age of an individual (Ellegren, 2000; Brouwer et al., 2009). The growing attention to homorepeats reflects both the discovery of homorepeat instability and a high capacity for cytotoxicity of these regions. It is still unknown why potentially harmful and genetically unstable homorepeats have been maintained throughout evolution.

In living organisms, homopeptides can also be of nonribosomal origin. Over the last few years, a large family of enzymes have been characterized whose members catalyze attachment of multiple amino acids to a γ-carboxyl group of a glutamate residue of target proteins (Edde et al., 1990; Redeker et al., 1994; Janke et al., 2005).

Unusual structural properties of homorepeats, including their ability to form crystalline structures, and to impart high concentrations of the same physico-chemical properties (hydrophobicity, charge, rigidity, etc.) within a given region, create an opportunity for successful application of homorepeat-containing molecules in biotechnology and materials science. Known examples include the poly-H tag, which is widely used to purify recombinant proteins, poly-K coating of culture dishes, which provides the properties required for cell adhesion and

proliferation (Ai et al., 2003), and poly-G molecules, which can self-assemble into different supramolecular structures (Kajava, 1999; Bovin et al., 2008).

The present survey was inspired by the fact that genome-sequencing projects are revealing a large number of protein sequences, and this information base has made it possible to conduct a detailed bioinformatics analysis of the homorepeat runs (Singh et al., 2004; Faux et al., 2005). In this chapter we use a protein repeat database (PRDB) that is accessible at http://bioinfo.montp.cnrs.fr/?r=repeatDB (Jorda and Kajava, 2009) to confirm or supplement the previous observations and to generate new conclusions. This bioinformatics tool can be used to analyze amino acid compositions, length distributions, cellular localization, functions, and other properties of the homorepeats.

II. Distribution of Homorepeats in Proteomes

A. What Is the Minimal Homorepeat Run That Should Be Taken into Consideration?

The exponential growth of genomic data has led to increasing efforts to analyze the distribution of homorepeats in coding regions of the genomes (Marcotte et al., 1999; Katti et al., 2001; Karlin et al., 2002; Faux et al., 2005). One of the first questions is "What is the minimum length of the homorepeat runs that should be taken into consideration?" The frequency of repetitive regions in proteins increases exponentially with the decrease of their length (Marcotte et al., 1999); therefore, the choice of the minimum length of the run can affect significantly the size and content of the analyzed set of homorepeats. The shorter the run, the less impact it may have on the structure and function of the protein. Very short runs may occur simply by chance, and not because of their functionality. Depending on the statistical models and assumptions, the minimum length used in the genome analyses is 5–7 residues (Karlin, 1995; Katti et al., 2001; Karlin et al., 2002; Faux et al., 2005). These values, based on standard random models of sequence evolution, may not be the most accurate in view of recent data that the frequency distribution of genomic motifs may arise from duplicative evolutionary processes, and not necessarily from selection based on point mutations and deletions

(Csuros et al., 2007). Nevertheless, the calculated minimum length of 5–7 residues is close to the limit at which some of the homorepeats start to affect structure and function. For example, proteins containing alanine runs of more than 10 residues demonstrate an enhanced tendency to aggregate (Fan et al., 2001). Furthermore, fusion of six histidines with a protein is sufficient for efficient purification of this recombinant protein via binding of the histidine tag to nickel-containing columns (Hengen, 1995).

B. Homorepeats Are Predominantly Hydrophilic

Several general conclusions became apparent across all the data. For example, homorepeats from all kingdoms are predominantly hydrophilic or composed of small residues such as Gly, Pro, and Ala, while large aliphatic, aromatic residues and Cys are rare (Green and Wang, 1994; Katti et al., 2001; Karlin et al., 2002; Faux et al., 2005) (Fig. 1). Furthermore, if runs of a hydrophobic residue do occur, their length tends to be less than 15 residues, while some hydrophilic runs can be longer than 50 residues (Karlin et al., 2002; Faux et al., 2005).

The most frequent runs made of large hydrophobic residues contain leucine (Fig. 1). A majority of poly-L runs (e.g., 90% in human proteins) occurs within the first 40 amino acid residues of the amino terminus identified as a part of signal-peptide sequences (Karlin et al., 2002). Other homorepeats of large hydrophobic residues do not have preferred locations within protein sequences. The great bulk of these hydrophobic runs represent transmembrane regions of cell membrane proteins.

The large number of hydrophilic runs can be explained by their lower propensity to aggregate in comparison to the hydrophobic stretches. Interestingly, runs of small nonpolar alanine are very frequent, despite their hydrophobic properties, the ability to form transmembrane regions, and their link to cytotoxicity (Fan et al., 2001). The abundance of poly-A may be associated with the small size of the side chain that attenuates the aggregation potential of the homorepeat thereby allowing a wider range of functions, including membrane anchoring, binding to proteins and DNA, and contributing to the rigidity of extracellular tissues such as elastins.

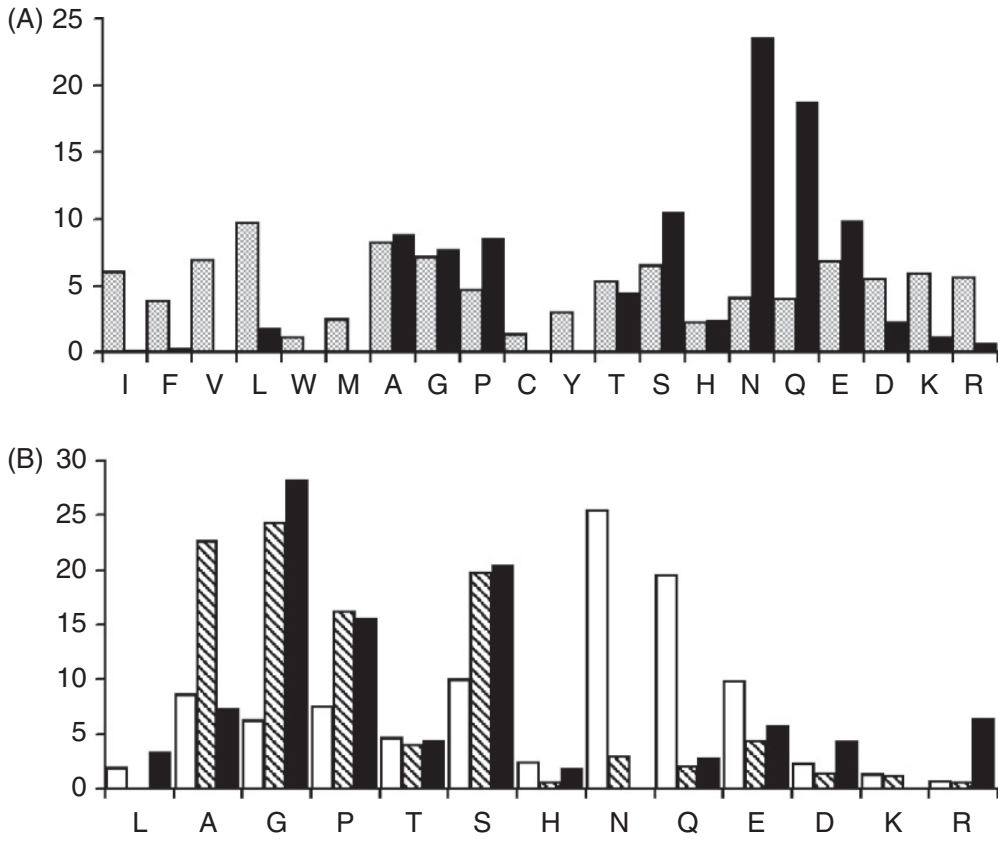


FIG. 1. Amino acid frequencies (%) ranged in order of decreasing hydrophobicity (Eisenberg et al., 1984). (A) Black and gray bars represent homorepeats ($\geq$7 residues) and all protein sequences, correspondingly, from Swiss-Prot (release July, 2009). The homorepeat data used in this and the other figures were taken from the PRDB. (B) Open bars represent homorepeats from eukaryotic proteins, the hatched bars show homorepeats from bacterial proteins, and the black bars correspond to viral proteins of Swiss-Prot. Poly-I, poly-F, poly-V, poly-W, poly-M, and poly-C homorepeats are omitted due to their low occurrences in all tested organisms. The remaining amino acid residues are arranged in accordance with the hydrophobic scale.

C. Homorepeats Are Much More Prevalent in Eukaryotic Than in Viral or Prokaryotic Proteins

Another conclusion revealed from numerous studies is that eukaryotes show a much higher frequency of homorepeats compared to prokaryotes (Karlin and Burge, 1996; Marcotte et al., 1999; Huntley and Golding, 2000; Faux et al., 2005). The analysis of Swiss-Prot homorepeats from PRDB shows that on an average 4.5% of the eukaryotic proteins have at least one homorepeat run (Fig. 2). Among the noneukaryotic organisms,

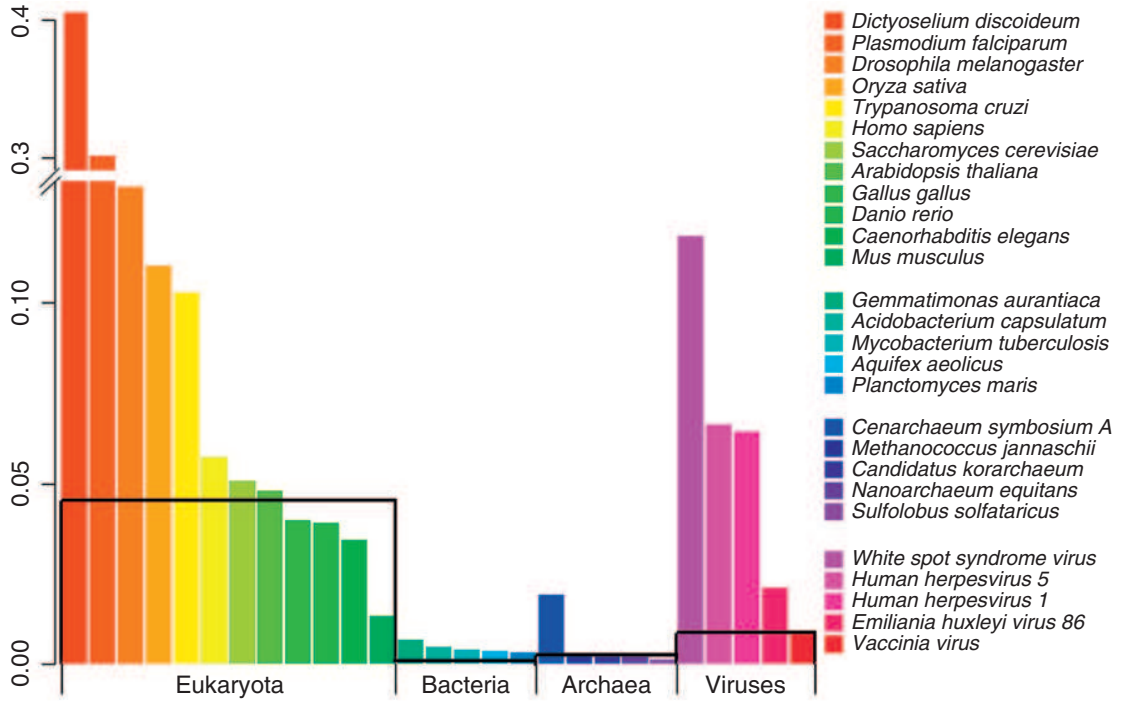


FIG. 2. Frequencies of homorepeat-containing proteins (HCP) in different proteomes. A representative set of 27 proteomes was clustered according to taxonomic domains (eukaryota, bacteria, archaea, and viruses). They have been gathered from the RefSeq database of NCBI. Within each taxonomic group, the proteomes are arranged in order of decreasing HCP frequencies. Open bars show the HCP frequencies in Swiss-Prot databank grouped by their taxonomic domains. Within the taxonomic groups, many HCP frequencies of individual proteomes are higher than the average Swiss-Prot frequency. This can be explained by the fact that our selection of the proteomes was biased toward organisms with a high number of homorepeats.

viruses have the highest frequency of homorepeat-containing proteins (about 1%) followed by archaea (0.2%) and bacteria (0.06%).

In eukaryotes, the most common homorepeats (in order of increasing frequency) are poly-Q, poly-N, poly-A, poly-S, poly-G, and poly-P. In prokaryotes, they are poly-S, poly-G, poly-A, and poly-P, and in viruses, they are poly-G, poly-S, poly-P, and poly-A (Fig. 1). A noticeable disparity is the rare occurrence of poly-Q and poly-N in prokaryotes and viruses that are the most common homorepeats in eukaryotes.

The acidic runs (poly-E and poly-D) exceed the basic runs (poly-K and poly-R) by a factor of 2 to 4 in eukaryotes while in bacterial genomes the proportions of the charged runs are almost equal (Karlin et al., 2002).

According to PRDB data, the proportion of basic runs is similar in eukaryotes and prokaryotes (about 6–7%) (Fig. 1). This indicates that the disproportion of the charged runs in eukaryotes is due to an increase of acidic runs. The acidic runs are not only more numerous in eukaryotes, but also far longer than the basic runs. The longest acidic runs are about 40–70 residues in contrast to the longest basic runs of 20–25 residues. Our analysis shows that a set of viral homorepeats from PRDB displays a prevalence of negatively charged runs compared to positively charged runs in a way similar to eukaryotes (Fig. 1). Surprisingly, in viruses, positively charged homorepeats are represented only by poly-R. The reason for the absence of poly-K is unclear.

D. Frequencies and Types of Homorepeats Vary Greatly between Eukaryotes

The higher frequency of homorepeats in eukaryotic proteins compared to prokaryotic ones is in an agreement with the previous conclusion regarding an increase of the homorepeat frequency with the genome size (Hancock and Simon, 2005). However, subsequent analysis of several complete genomes revealed that this correlation is not observed in either prokaryotic or eukaryotic proteomes (Huntley and Golding, 2000; Karlin et al., 2002; Singh et al., 2004; Faux et al., 2005). Table I shows how the percentages of proteins with at least one run of homorepeats in eukaryotes vary in order of the increase of the genome size. This comparison indicates a striking difference between the organisms. *Plasmodium falciparum* and *Dictyostelium discoideum* have

TABLE I
Genome size and percentage of homorepeat-containing proteins in different eukaryotes

Organism	Homorepeats (%)[a]	Size of genome (Mb)	Number of proteins
Saccharomyces cerevisiae	5	13	5,884
Plasmodium falciparum	30	23	5,271
Plasmodium yoelii	9.4	23	7,353
Caenorhabditis elegans	3.4	100	23,906
Arabidopsis thaliana	4.8	140	33,402
Drosophila melanogaster	13.2	165	21,099
Homo sapiens	5.7	3,300	37,392
Dictyostelium discoideum	40.4	34	13,408

[a] Percentage of homorepeats from PRDB.

relatively small genomes in the listed organisms but encode the highest numbers of proteins with homorepeats (36% and 40.4% of proteins, correspondingly). When grouping organisms differently, for example, considering only vertebrates, the variability is also high. The homorepeats are 10 times more frequent in mammals than in amphibians, birds, or fishes. Surprisingly, a noticeable difference is also observed within the same family of malaria parasites. The rodent parasite *Plasmodium yoelii* contains almost three times less proteins with homorepeats than human parasite *P. falciparum* (Singh et al., 2004). Currently there is no plausible explanation for this variation. Comparison of homorepeats within either bacterial, archaeal, or viral genomes also shows strong variations of the repeat frequencies (Fig. 2), and this variation does not correlate with the size of the genomes.

The other remarkable difference observed in homorepeats of eukaryotes is organism-specific diversity in the proportions of the types of the amino acid runs (Karlin et al., 2002; Faux et al., 2005). Although poly-Q, poly-N, poly-A, poly-S, poly-G, and poly-P runs account for a significant proportion of the runs in almost all eukaryotes, human proteins are most abundant (in increasing order) in poly-Q, poly-P, poly-L, and poly-A runs. Fly proteins are most abundant in poly-Q, poly-A, poly-S, and poly-G, *Caenorhabditis elegans* proteins in poly-S, poly-T, poly-P, and poly-Q, proteins from *Saccharomyces cerevisiae* in poly-S, poly-Q, poly-E, poly-D, and poly-N, *Arabidopsis thaliana* proteins in poly-S, poly-E, poly-G, and poly-P, and *P. falciparum* proteins in poly-S, poly-T, and poly-N. *Drosophila melanogaster* is enriched in poly-Q repeats (3.5 times more than *Homo sapiens*) but poly-Q runs are absent in *P. falciparum* (Faux et al., 2005). The most striking observation is that, for unknown reasons, poly-N runs are extremely rare in vertebrates while that of the chemically and structurally similar poly-Q are very frequent. In contrast, in nonvertebrates, poly-N represents an important fraction of homorepeats but not poly-Q (Kreil and Kreil, 2000; Karlin et al., 2002; Faux et al., 2005).

E. Distribution and Localization of Homorepeats within Proteins

The occurrence of more than one run in a given protein (called multiple homorepeat runs) is quite frequent, especially in eukaryotes. For example, about 23% of all homorepeat-containing proteins have more than one run including 24% of these types of proteins in eukaryotes and only 9% for prokaryotes (Faux et al., 2005). In most cases the multiple

runs in a given protein consist of the same amino acid residue. Among the pairs of different types of homorepeats in one protein, the most frequent pairs include either poly-N or poly-Q runs, which also individually represent the most frequent homorepeats. Strong association is observed for the pairs poly-A and poly-Q, poly-H and poly-Q, poly-D and poly-N, poly-K and poly-N, poly-F and poly-N, and poly-T and poly-N (Faux et al., 2005).

Homorepeats are preferably located at the N- (43%) and/or C-terminal (30%) flanking regions of well-defined structural domains and they are seldom found within (16%) or between (10%) of the domains (Faux et al., 2005). The flanking regions tend to be disordered (Alba et al., 2007). Poly-L runs mainly occur within the first 40 residues corresponding to the signal-peptide sequence (Karlin et al., 2002). The other homorepeats do not have such a strong preference for specific regions within proteins.

F. Protein Size and Tendency to Have Homorepeats

It was shown that the percentage of proteins containing at least one homorepeat rises with an increase in protein size (Singh et al., 2004). It may appear obvious that it is more probable to find a homorepeat as the length of a protein increases. However, this tendency is also observed when the frequency of homorepeats per unit of length is plotted, but it is less pronounced (Fig. 3). Remarkably, at the beginning of the plot, the smallest proteins (<100 residues) have a high propensity for homorepeats, which drastically decreases for proteins with sizes of about 200 residues. As the protein size increases, the homorepeat propensity gradually rises and reaches its maximum at around 900 residues followed by another maximum at 1,700 residues. Interestingly, according to this plot, proteins of sizes less than 100 residues can be distinguished as a special group by their strong tendency to have homorepeats. Amino acid compositions of these homorepeats support the conclusion that they have distinctive characteristics (Fig. 3). They have un unusually high proportion of positively charged runs (20.8% vs. on the average 1.8% for all proteins) and have almost no negatively charged runs (3.2% vs. 12.1%). The proportion of the hydrophobic runs is unusually high (8.7% vs. 2%). The majority of the proteins containing positively charged runs bind to DNA in the nucleus, while ones with hydrophobic runs are located in the

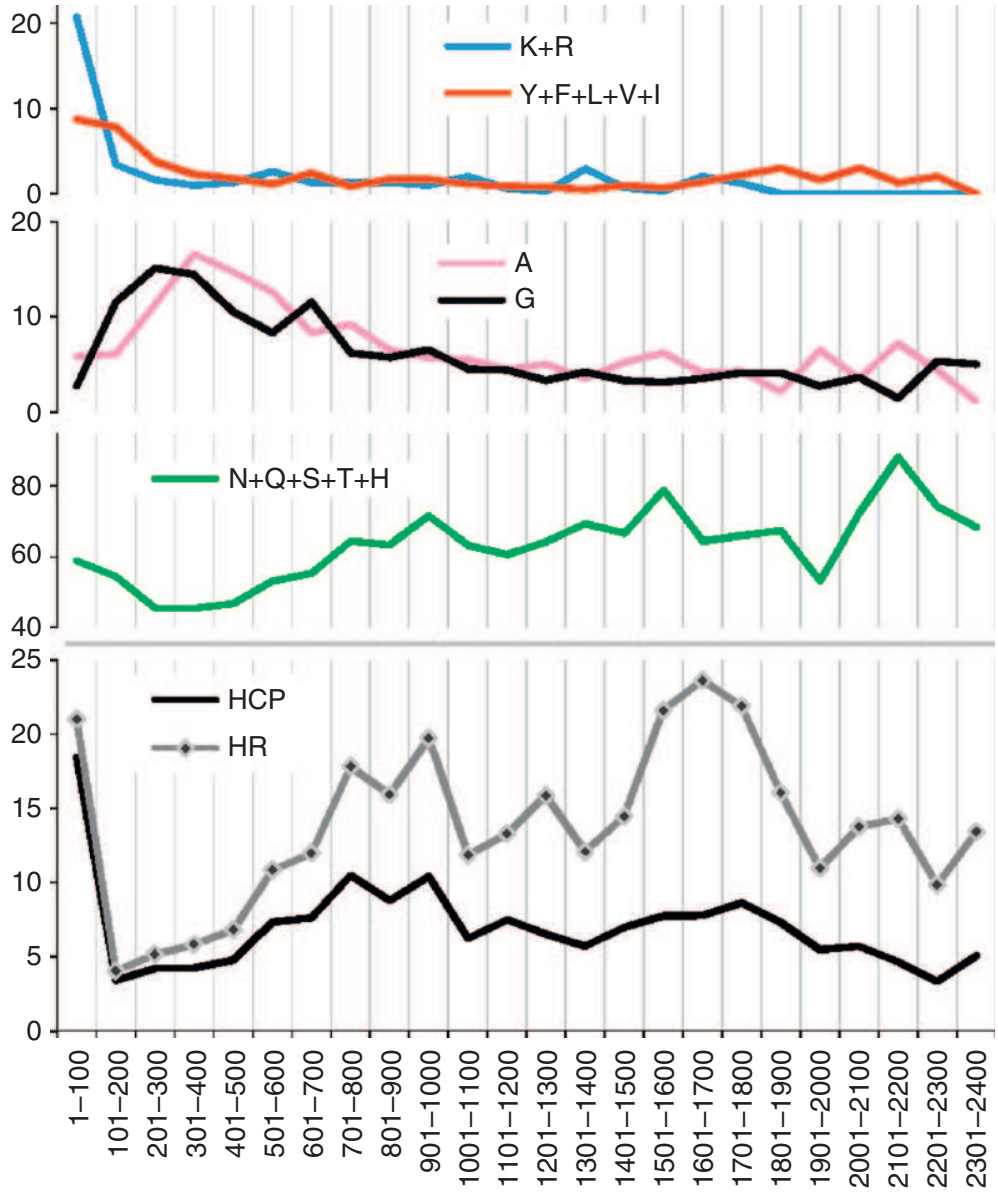


FIG. 3. Frequencies per unit length (in arbitrary units) of Swiss-Prot homorepeat-containing proteins (HCP) and homorepeats (HR) as a function of protein size (at the bottom). Above, corresponding amino acid compositions of HRs are shown.

membranes. Among these small proteins with homorepeats, cytoplasmic proteins are practically absent. Proteins of size between 100 and 600 residues also have biased compositions of homorepeats with increased proportions of poly-A and poly-G and with decreased proportions of polar homorepeats (Fig. 3).

III. Structural Properties of Homorepeat Regions

A. Synthetic Homopeptides Form Crystallites

Synthetic homopeptides were among the first objects of structural studies of proteins due to the availability of such specimens and the relative simplicity of data interpretation. It was shown that under physiological conditions, homopeptides containing charged or polar amino acid residues tend to be in a random coil conformation while many nonpolar homopeptides form aggregates (Bloom et al., 1962; Schulz and Schirmer, 1979). Some correlations were found between the types of amino acid residues and the preference of the homopeptides to be in α-helix or β-structure (Bloom et al., 1962; Auer and Doty, 1966). For example, it was noticed that C_β-branched amino acid residues prefer the β-structure to α-helical conformation. Nonpolar solutions that model the environment within biological membranes favor the α-helical conformation for most homopeptides. Studies of homopeptides also revealed their ability to form well-ordered crystallites suitable for X-ray fiber diffraction studies (Fraser and MacRae, 1973). This aptitude of the homopeptides was used to determine the geometry, basic conformation, and interchain packing of the polypeptide chain (Fraser and MacRae, 1973). An important outcome of these studies was a description of the principal elements of the secondary structure found in proteins: the α-helix, the parallel and antiparallel pleated β-sheet structures and the polyproline II helices (Pauling and Corey, 1951; Fraser and MacRae, 1973; Fraser et al., 1973; Eisenberg, 2003) (Fig. 4A and 4B).

Depending on the amino acid type, peptide length, and conditions of the experiment (procedure of crystallization, nature of solution, peptide concentration, ionic strength, pH, temperature), homopeptides can form different types of crystallites. For example, poly-A and poly-K peptides can self-assemble into α-helical or β-structure crystallites (Blout, 1962; Auer and Doty, 1966). Poly-V and poly-S exclusively form the β-structured assemblies while poly-Y peptides favor α-helical structures (Schulz and Schirmer, 1979).

In α-helices, the peptide groups of the backbone are involved in intrahelical H-bonds. The side chains are on the surface and play a critical role in the intermolecular interactions (Fig. 4). In the crystalline form, the α-helices are tightly packed in the hexagonal arrays interacting via the side chains. This favors a platelet-like morphology with hexagonal habit

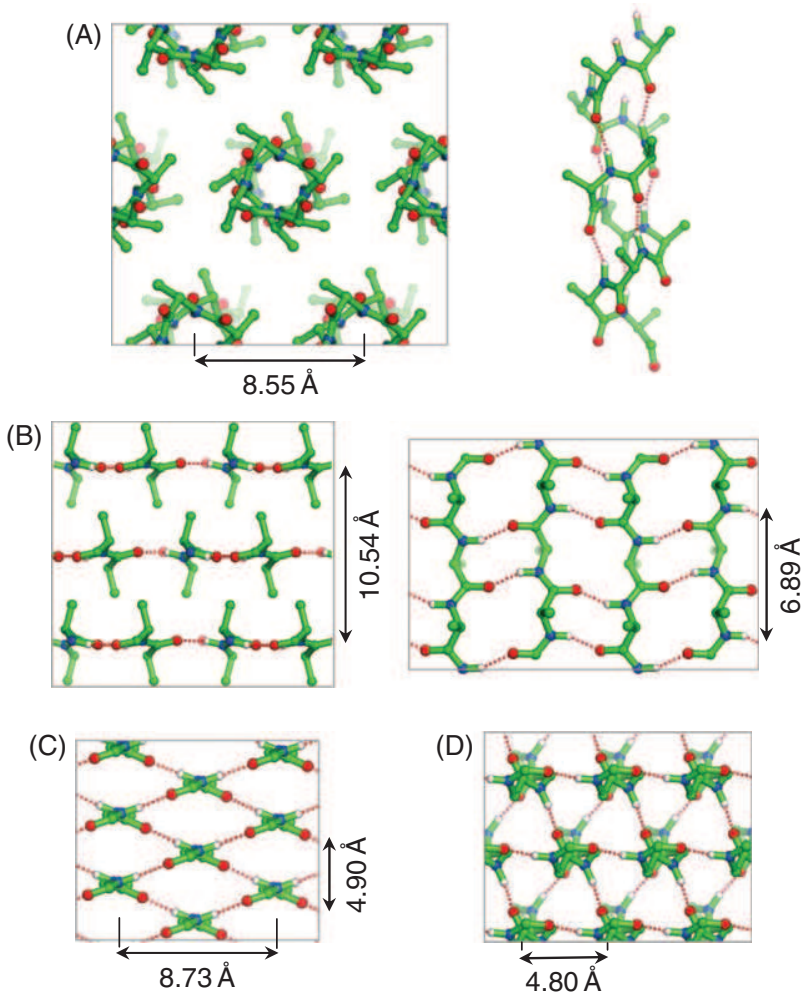


FIG. 4. Principal types of crystal structures formed by homopeptides. The structures are in ball-and-stick representations with carbon atoms depicted in green, nitrogens in blue, and oxygens in red. Only peptide group hydrogens are shown. H-bonds are shown as red dotted lines. (See color plate 4) (A) On the left, axial projection of the crystal structure formed by poly-A α-helices. The hexagonal unit cell has the parameters: $a = 8.55$ Å and $c = 70.3$ Å (Arnott and Dover, 1967). On the right, a lateral view of poly-A α-helix is shown. (B) Axial projection of the antiparallel β-structure of poly-A with an orthogonal unit cell $a = 9.46$ Å, $b = 10.54$ Å, and $c = 6.89$ Å (chain axis) (Arnott et al., 1967). On the right, a lateral view on one antiparallel β-sheet of this crystal structure. (C and D) Axial projections of crystal structures proposed for poly-G form I (Kajava, 1999) and form II (Crick and Rich, 1955). The structure of form I shown in this figure is formed by low molecular weight poly-G molecules. It has a monoclinic unit cell with $a = 4.9$ Å, $b = 8.73$ Å, $c = 7.18$ Å (chain axis), and $\beta = 130.1°$. Form I for high molecular weight molecules is formed by stacking of rippled layers that are similar to β-structural layers. Form II has a hexagonal unit cell ($a = 4.8$ Å, $c = 9.3$, chain axis). The molecular structures were generated with Pymol (De Lano, 2002) (See color plate 5).

(Padden and Keith, 1965). In the α-helical crystallites, all alanines cannot have equivalent environments due to the 3.6 residue per turn helical symmetry. In contrast, within the β-structure crystallites, each residue of a homopeptide has an equivalent environment (Fig. 4). β-Structures are formed by polypeptide chains in extended conformations. The chains interact with each other via peptide group H-bonds and form a flat β-sheet. Within the β-sheets the chains can be arranged in either a parallel or an antiparallel manner. The side chains project above and below the plane of the sheet. The β-sheets are stacked and tightly packed to form stable crystallites. In general, the high stability of these structures is provided by close packing of nonpolar side chains. However, it was also shown that polar poly-Q homopeptides can self-assemble into very stable and high-density β-crystallites (Sikorski and Atkins, 2005) due to the ability of these side chains to form networks of H-bonds between β-sheets (Fig. 5). Given the fact that the expansion of poly-Q runs causes a

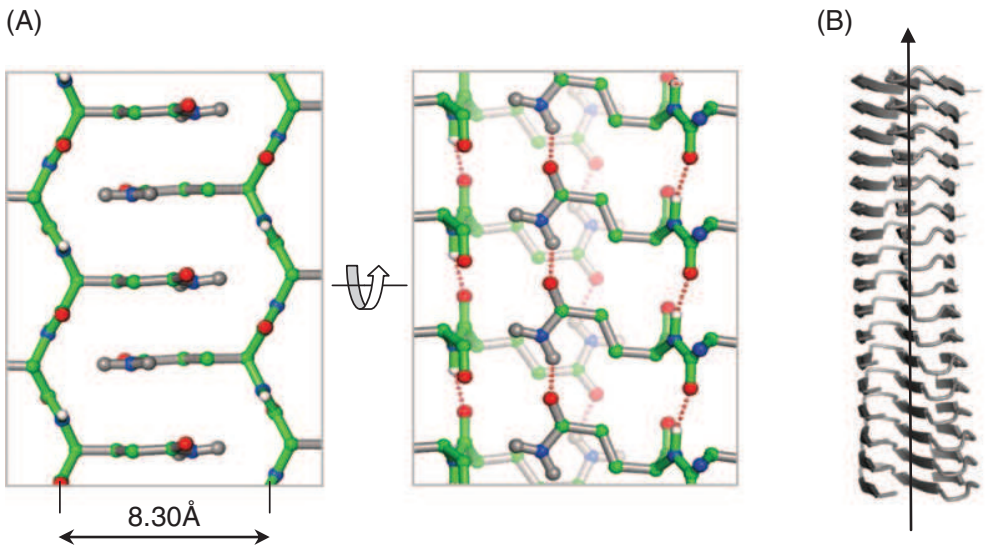


FIG. 5. (A) A model of β-crystallite of poly-Q with an orthogonal unit cell with $a = 9.5$ Å, $b = 16.60$ Å, and $c = 6.95$ Å (chain axis) (Sikorski and Atkins, 2005). On the left, a view along the H-bond directions shows interdigitation of Q side chains. On the right, a view along the β-strands illustrates how the H-bonding of Q side chains creates ladders inside the β-structure. Sticks of the main chains are in green and of the side chains are in gray. (B) A schematic diagram of a cross-β amyloid fibril formed by twisted β-sheets. Beta-strands are shown as arrows. The twist can delimit lateral dimensions and favor axial growth of the homopeptide aggregates. When considered locally, side chain packings (e.g., the H-bonded Q-ladders) within the crystallite (A) and fibril-like assemblies (B) can be similar. (See color plate 5).

number of neurodegenerative disorders, including Huntington's disease (Orr and Zoghbi, 2007), this crystallite structure is important in understanding the role of poly-Q segments in neurodegenerative disorders.

It is worth mentioning that, in known protein structures, β-sheets are rarely flat. Rather, they exhibit a slight twist due to the chirality of their amino acids (Chothia, 1973). This twist delimits the length of the β-strands and the number of β-sheets in the stacks, and by doing so it can favor formation of long fibrous cross-β structures (Nelson and Eisenberg, 2006) (Fig. 5) instead of crystallites.

The crystallites formed by poly-P and poly-G homopeptides differ from both the α-helical and the β-structure crystals. The characteristic properties of poly-P are derived from the fact that Pro residue has a cyclic side chain and its ϕ dihedral angle is fixed at approximately –75° providing exceptional rigidity. On the other hand, it lacks a hydrogen on the amide group and, therefore, cannot form interpeptide H-bonds between poly-P chains. These constraints make poly-P incompatible with the α-helical and β-structure conformations. Two forms of poly-P have been found (Harrington and Sela, 1958): form I consists of regular poly-P helices with the *cis*-configuration of the peptide bonds and form II is a regular helix with the *trans*-peptide groups.

Poly-G also stands apart from other homopeptides. Glycine has a hydrogen atom as its side chain and does not contain an asymmetric C_α-atom. Therefore, the mirror-symmetrical aggregates of poly-G can be formed with equal probability. Furthermore, the small size of glycine allows these polypeptides to approach each other so closely in the crystals that their peptide groups form a spatial network of H-bonds. Poly-G can form at least three crystalline forms. The crystal modification "polyglycine I" has two different forms depending on the molecular weight. In the low molecular weight (~10–15 residues) form I, the molecules have an unusual extended conformation generated by the alternation of two mirror-symmetrical residual conformations along the chain (Kajava, 1999). The molecules are parallel with respect to the direction from N- to C-terminus and each chain forms interpeptide H-bonds with four adjacent chains (Fig. 4). The high molecular weight (~90 residues) form I crystal represents a composition of twinning crystallites. The crystallites themselves consist of antiparallel enantiomorphous chains united by H-bonds to form rippled sheets (Pauling and Corey, 1953; Kajava, 1999). In the high molecular weight form I the orientation of the chains is antiparallel, but it is parallel in the low

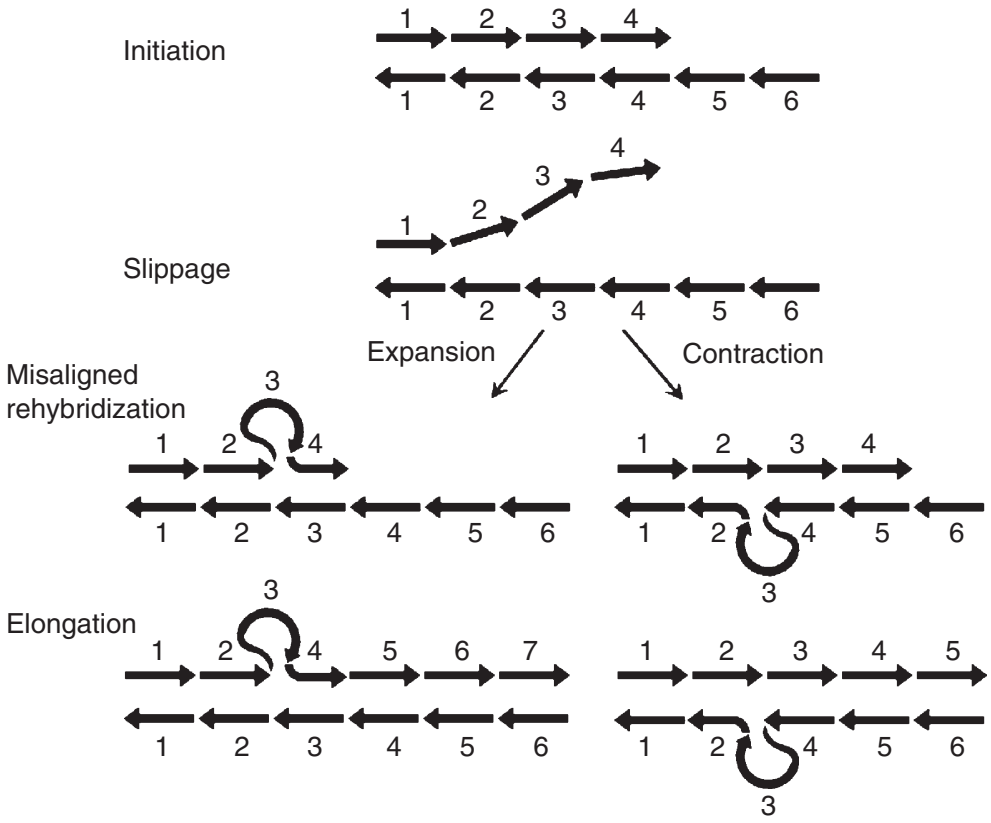


FIG. 6. Model of molecular mechanism of microsatellite mutation based on replication slippage. DNA repeat units are shown by arrows and numbered within each strand.

molecular weight structure. Therefore, one crystal modification cannot be transformed into the other by mechanical or other means. In the crystal form "polyglycine II," each molecule has a threefold screw axis and conformation similar to the poly-P II helix (Crick and Rich, 1955) (Fig. 4). The chains are parallel with respect to the N- to C-terminus direction, packed in a hexagonal array and each is H-bonded to its six nearest neighbors.

B. Living Organisms Avoid Structure-Forming Homorepeats

Self-assembly itself is not toxic for the cell and is widely used by natural proteins. It is not the assembly itself but the inherent property of homorepeat structures to form stable aggregates of unlimited size that may be a basis of homorepeat cytotoxicity. Both properties (unlimited size and high stability) of the homopeptide structures do not correspond to requirements of biological molecules: a) to be able to control their

dimensions and b) when it is necessary, to be readily degraded (Kajava, 2001). Thus, the homopeptide structures are predisposed to form deposits that are harmful to living organisms (Perutz, 1999).

When homorepeats with the potential to self-assemble occur in natural proteins, they are usually flanked by large aperiodic sequences typical of globular structures (Kreil and Kreil, 2000; Kajava, 2001). The flanking globular structures may prevent the formation of homorepeat associates by steric impedance or interactions with the homorepeat region. As a rule, those rare proteins that contain homorepeat runs of bulky nonpolar residues target membranes after their biosynthesis and are not released to the aqueous solution. Poly-L, the most frequent among them, is preferentially located at the N-terminal leader peptide region, which is proteolytically digested after translocation of its protein across the membrane (Karlin et al., 2002). Most of the rare proteins that are composed of multiple stretches of structure-forming homorepeats are extracellular. For example, production of silk fibers is designed so that poly-A-containing silk fibroin forms its final structure only when outside the organism (Hinman et al., 2000).

IV. Evolutionary Dynamics of Homorepeats

Homorepeat regions of proteins are coded by tandemly repeated stretches of DNA known as microsatellites (Charlesworth et al., 1994). The microsatellites differ from most other types of DNA sequences in their unusually high polymorphism (Goldstein and Schlötterer, 1999). Depending on the organism, the mutation rate of microsatellites is several orders of magnitude higher (10^3–10^6 per locus per generation) than the rate of the unique DNA sequence (Weber and Wong, 1993; Ellegren, 2000). The rate also depends on the total length of the individual microsatellite—long loci mutate more often than short ones (Wierdl et al., 1997; Brinkmann et al., 1998). A series of studies suggests that the homorepeat instability can be both intergenerational and somatic (Brouwer et al., 2009).

The mutation mechanisms contributing to the length polymorphism at microsatellite loci are still a matter of debate. However, an increasing amount of evidence suggests a model based on replication slippage (Levinson and Gutman, 1987; Ellegren, 2000). According to this model, when slipping-strand mispairing occurs within a microsatellite region during DNA synthesis, it can result in the gain or loss of one or more

repeat units depending on whether the newly synthesized DNA chain or the template chain respectively loops out (Fig. 6). Some data suggest that the loops of highly expandable repeats can form hairpin structures with both Watson–Crick and mismatched base pairs (Mirkin, 2006).

The homorepeat regions of proteins consist predominantly of homocodons (the same type of codons) and they do not have an apparent G+C or A+T nucleotide base bias relative to the organism's transcriptome (Faux et al., 2007). In general, transition substitutions within the codons are more common than transversions (Collins and Jukes, 1994; Morton, 1995); however, most of the homorepeats comprise a homocodon interrupted by a single base transversion (Faux et al., 2007). Disruptions of homocodons are likely to reduce the rate of slippage mutation as shown in *S. cerevisiae* (Rolfsmeier and Lahue, 2000), probably by reduction of DNA loop formation during tandem repeat replication.

Thus, on one hand, the homorepeat regions evolve more rapidly than the rest of the gene but, on the other hand (see Section III.C), their expansion may be toxic to the cells. This suggests that the homorepeat expansion and contraction is under a strong control by a number of factors including DNA replication, repair, and modification machineries (Karlin et al., 2002; Cleary and Pearson, 2003; Feschenko et al., 2003). Given the fact that the homorepeat regions are harmful to the organisms, it is unclear as to why the instability of homorepeat regions be maintained throughout evolution? Recently, it was shown that homorepeat instability can have positive phenotypic effects. For example, homorepeat expansion and contraction within *cis*-regulatory elements of developmental genes is linked to morphological changes in dog breeds (Fondon and Garner, 2004). This observation offers the hypothesis that the repeat length instability is a major source of phenotypic variation in evolution (Fondon and Garner, 2004).

V. Functional Associations of Homorepeat Regions

A. *Functions of Proteins Containing Homorepeat Regions*

Several studies have indicated an association between homorepeat-containing proteins and their functional properties (Singh et al., 2004). These proteins are involved in a large variety of functions including

catalytic activity, enzyme regulation, transport, channeling, adhesion, signaling, translation, and structural functions. A survey of proteins with homorepeat runs indicates that despite a diversity of functions, they tend to belong to specific functional classes of intracellular proteins, particularly transcription factors and protein kinases (Karlin and Burge, 1996; Alba and Guigo, 2004). This is especially true for eukaryotes. Prokaryotes, on the other hand, do not have such overrepresentation of homorepeat-containing proteins of transcription/translation and signaling. They account for 24% of enzymes and 12% of adhesion molecules among the proteins considered (Faux et al., 2005). In eukaryotes, the majority of the homorepeat-containing proteins are associated with large multiprotein and/or nucleic acid complexes such as ribosomes and their factors, complexes involved in transcriptions and chromatin condensation and signaling complexes (e.g., MAP kinase) (Faux et al., 2005). According to the PRDB, in viruses, the greatest number of homorepeats is found among structural proteins of capsids and envelopes followed by transcriptional factors.

B. Do Homorepeats Represent Junk Protein Regions or Have Essential Functional Roles?

Attempts to find a specific functional role for protein homorepeats have produced contradictory conclusions. Some authors suggest that the homorepeats serve primarily as spacers between functionally important domains (Huntley and Golding, 2000). This point of view is based on the facts that a) the majority of the homorepeat runs are predicted to be intrinsically unfolded (or unstructured) (Huntley and Golding, 2002), b) the rate of evolution is higher within than outside of the homorepeat, c) the repeat region length as well as type of amino acid residue can vary between species (Newfeld et al., 1994) and, d) homorepeats of some proteins may be deleted without affecting the function (Gatti et al., 1994; Sumiyama et al., 1996). At the same time, other studies clearly demonstrate that homorepeats themselves are critical to the function of some repeat-containing proteins. For example, stretches of poly-Q or poly-P can activate transcription when fused to the DNA-binding domain of GAL4 factor (Gerber et al., 1994). The transcriptional activation can be modulated by variation of the length and amino acid type of the homorepeats. Comparative genomics and morphological studies suggested that length variation in homorepeats can be a major source of

morphological variations in dog breeds (Fondon and Garner, 2004). The role of neither poly-A nor poly-G runs that are found in extracellular proteins such as silk fibroin and elastin can be reduced to spacers between the domains (Altman et al., 2003).

The most abundant homorepeats of poly-Q and poly-N are prone to cytotoxic aggregation. However, they were not replaced during evolution by less harmful hydrophilic homorepeats. This suggests that poly-Q or poly-N may be involved in specific binding with other molecules rather than serving as a neutral cross-link between functional domains.

These and other examples clearly demonstrate the diversity of homorepeat functional roles rather than the existence of one function. The frequent absence of a stable folded structure of homorepeat regions (Huntley and Golding, 2002) cannot be used as a proof in favor of the insignificance of the homorepeat functions. The intrinsically unfolded regions have a number of important functions (Radivojac et al., 2007). At the same time, often, expansion or contraction of the homorepeat regions provide subtle regulation and modulation of the basic function of their proteins (Gerber et al., 1994; Fondon and Garner, 2004). Therefore, the rapid evolution of these regions (Newfeld et al., 1994) should not endanger the population but rather provide means of attaining the necessary diversity in the face of evolutionary pressure.

C. Preferences of Homorepeat Types for Certain Localizations and Functions

Homorepeats of some amino acid residues occur more frequently than others depending on the function or subcellular localization of their proteins. According to the PRDB, almost all proteins containing poly-R are located in the nucleus. This relationship also holds, but is less pronounced, in proteins containing poly-K. This observation is supported by experiments with recombinant proteins fused with poly-R or poly-K tracts that are found in the nucleus (Nam et al., 2001; Oma et al., 2004). Poly-K runs often occur in membrane proteins. In contrast, the negatively charged poly-E and poly-D are abundant among cytoplasmic proteins (Fig. 7). The cytoplasmic proteins are also enriched in poly-P and poly-G. Nuclear proteins have a high proportion of poly-A and a low proportion of polar homorepeats. As expected, membrane proteins are overrepresented by nonpolar runs. Nonpolar runs are also frequent in secreted proteins.

Clustering of homorepeat-containing proteins by their functions also reveals preferences for certain types of amino acid residues (Fig. 7). For example, the occurrence of poly-A runs is greater-than-average in transcription factors while negatively and (more surprisingly) positively charged residues are almost absent among these proteins. Structural proteins have many poly-G and poly-A, but very few polar homorepeat runs.

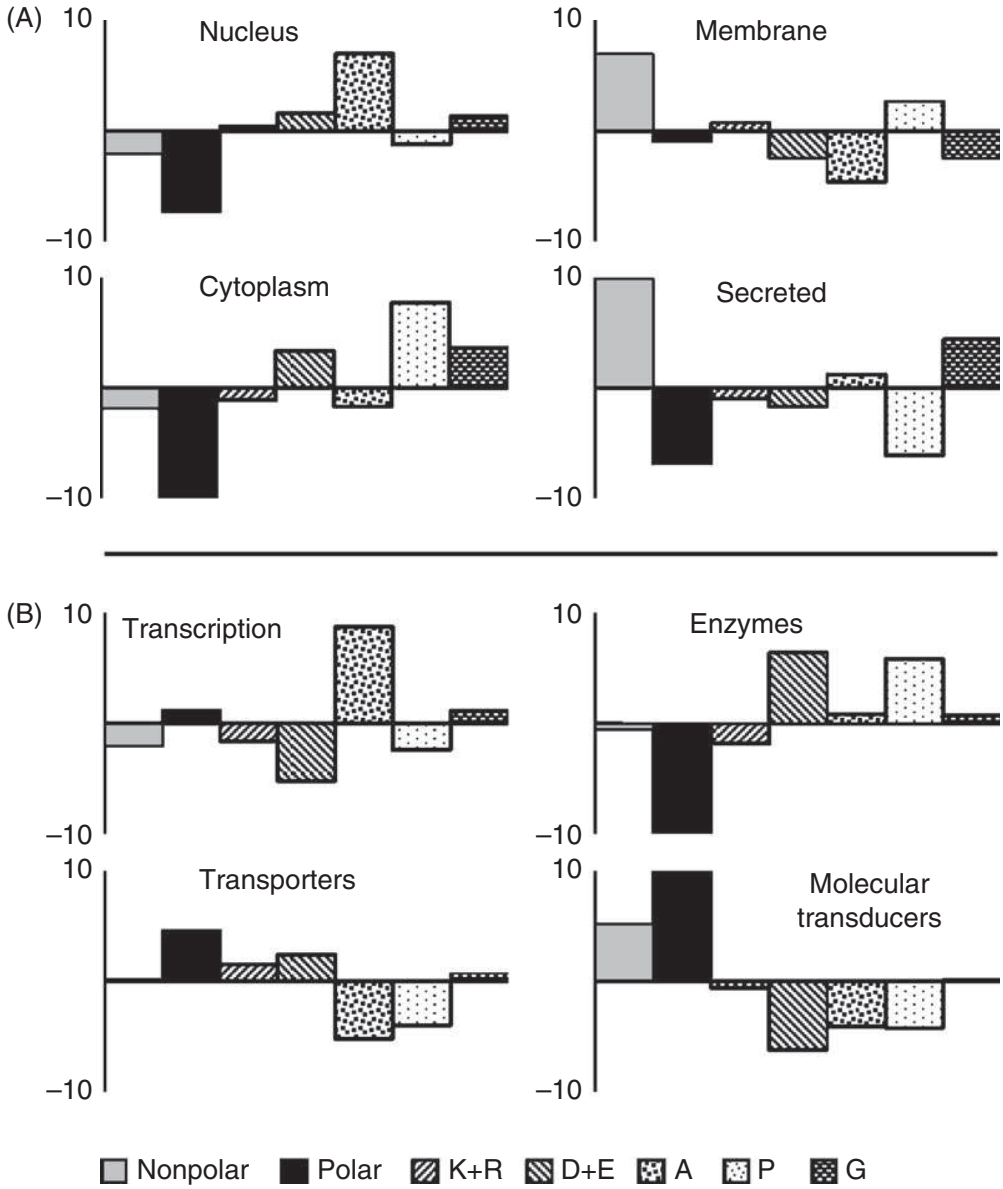


FIG. 7. Histograms presenting differences between frequencies of types of homorepeats assigned to particular localizations (A) and functions (B) and the average frequency of all homorepeats from Swiss-Prot.

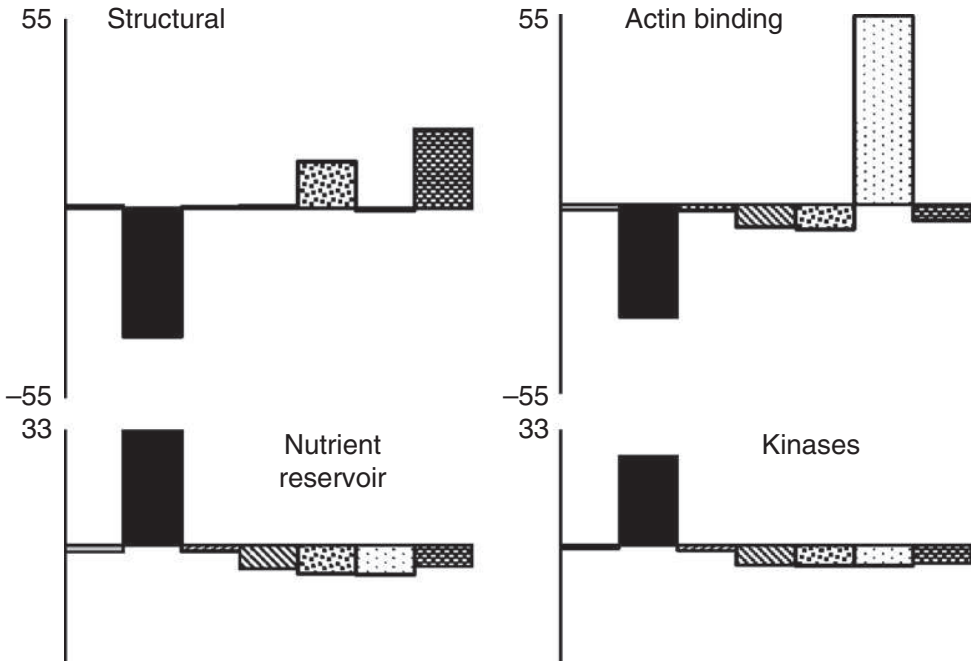


FIG. 7. (Continued)

In contrast, kinases and proteins involved in nutrient reservoir activity have an extremely high proportion of polar homorepeats. Actin-binding proteins have an unusually high content of poly-P runs and a low proportion of polar homorepeats. This observation correlates with the ability of several actin-binding proteins (profilin, SH3, WW-domains, and others) to bind proline-rich regions. Enzymes have many negatively charged runs and poly-P, but a "less-than-average" proportion of polar runs.

D. Functions Specifically Related to Homorepeat Regions

Usually, homorepeats occur in large proteins containing several functional domains. Therefore, it is difficult to separate the homorepeat function from the remaining part of the protein. Knowledge of the unique properties of homorepeat sequences, structures, and evolution (see previous sections) in combination with the functional data and mutational analysis of the repeat-containing proteins allows us to speculate about functions that are specifically linked to homorepeats.

The exact repetitiveness and homogeneity of the sequences suggest that the homorepeats can interact well in a multivalent manner with

partners having regular structures, such as DNA, RNA, carbohydrates, repetitive protein structures, or nonorganic crystals (Fig. 8). In several cases when a homorepeat-containing protein interacts with DNA, the variation of the homorepeat length provides subtle modulation of the basic protein function (Gerber et al., 1994; Fondon and Garner, 2004). This suggests that the interactions of the homorepeat with DNA are nonspecific and are characterized by weak energy contributions per amino acid residue. DNA grooves may provide such regular binding surfaces for the homorepeat regions. Interactions are likely between the positively charged homorepeats and the negatively charged backbone of the nucleic acid. It can also be mediated by H-bonding between the DNA/RNA grooves and poly-Q, poly-N, poly-A, poly-P, or poly-G in extended conformations. The fact that different types of homorepeats differ in their optimal length for transcriptional activation (Gerber et al., 1994) can be explained by the amino acid residue-specific differences in the strengths of interaction with DNA. The mode of a linear, nonspecific interaction can tolerate a certain level of homorepeat length variation

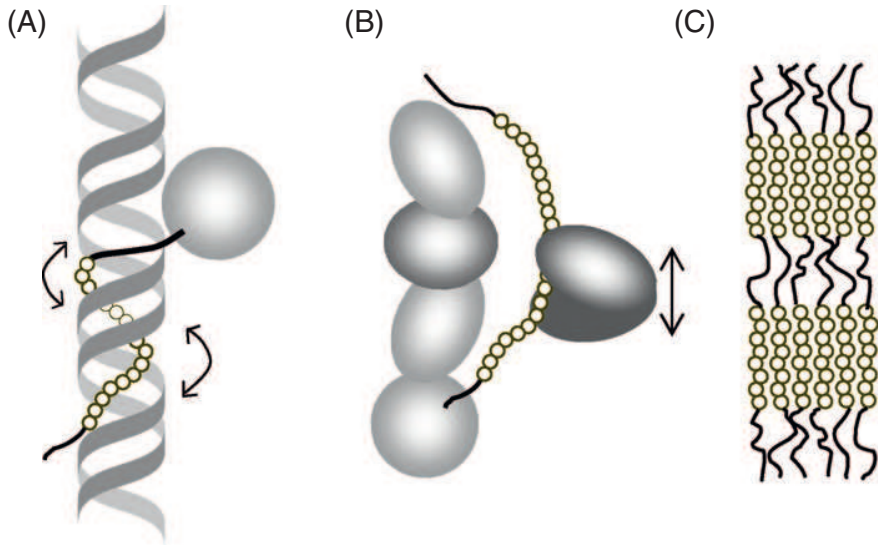


FIG. 8. Examples of homorepeat-specific functions. Globular protein structures are shown as spheres and homorepeat regions as "beads on string." (A) DNA–homopeptide interaction is shown as an example of multivalent binding to repetitive structures. (B) Sliding of molecules along homorepeat regions. This movement can provide necessary mobility to subunits of a supramolecular machine. (C) Formation of homorepeat crystalline structures.

quite well. Therefore, it can also efficiently attenuate the evolutionary instability of these regions.

The exact repetitiveness of the homorepeats also provides a basis for efficient interaction with nonorganic crystals or catalysis of their crystallization. For example, poly-E runs of bone sialoprotein are involved in the hydroxyapatite-nucleating activity (Tye et al., 2003).

The homogeneity of the homorepeats can also provide another set of unique functional properties. It is known that homorepeat-containing proteins participate in large suparmolecular machines and mediate and regulate protein–protein interaction networks (Remacha et al., 1995; Korschen et al., 1999; Mayer, 2001; Poetsch et al., 2001; Hancock and Simon, 2005). When the size of a homorepeat stretch is larger than its binding site to a protein partner, the bound protein can slide along the homorepeat region while maintaining contact. *Sliding* along the homorepeat region may impart dynamic properties to the supramolecular machine, which can be critical for its functioning (Fig. 8). Such homorepeats may tolerate the extension–contraction by a few residues that may occur during evolution. Complexes between homorepeats and nucleic acid molecules can also be very dynamic due to the sliding mechanism. The sliding might be critical for proper functioning of large protein–nucleic acid complexes involved in transcription, translation, and recombination processes.

The homogeneity of homorepeats provides a predisposition to interact not only with other periodic molecules (e.g. DNA, RNA) but also with themselves forming crystalline structures or amyloid fibrils (Fig. 8). Frequently, this aggregation is associated with cytotoxicity. However, nature also uses this property for beneficial biological functions. For example, the homorepeat crystallites are required to stiffen silk fibers or elastin tissue (Hinman et al., 2000; Altman et al., 2003). Furthermore, recent studies have identified amyloid fibers in bacteria, fungi, insects, invertebrates, and humans that are necessary for the proper functioning (Fowler et al., 2007; Maji et al., 2009).

Previously it was proposed that homorepeats, due to their inherent predispositions to rapid evolution, may be used to increase the overall size of proteins (Lavorgna et al., 2001). When such a possibility exists, we believe that it represents a rather minor, collateral role for homorepeats. It does not require specific properties of the homorepeats and can be fulfilled even more efficiently by multiplication of longer (more than

1 residue) repetitive units. The principal role of the homorepeats as a flexible neutral linker between folded domains (Karlin and Burge, 1996; Faux et al., 2005) is also unlikely. This function does not require perfect repetitiveness of the link and can be provided by other less regular and intrinsically unfolded regions. The fact that only a minor portion (about 10%) of homorepeat regions occurs between folded functional domains (Faux et al., 2005) supports this conclusion.

E. Function of a Homorepeat Region Strongly Depends on the Adjacent Domains and Subcellular Localization of the Protein

In contrast to the globular domains, homorepeats of the same type and length can have very different functions depending on the adjacent domains and subcellular localization of the protein. For example, runs of poly-A or poly-G fused with DNA-binding domains of transcriptional factors can modulate activity of these proteins (Gerber et al., 1994). On the other hand, the same poly-A or poly-G as a part of silk fibroin or elastin tend to self-assemble into the crystallites that are necessary to support the mechanical properties of the proteins (Fraser and MacRae, 1973; Altman et al., 2003). Depending on the protein context, poly-P homorepeat can serve as cross-links, modulate transcriptional activity (Gerber et al., 1994), or provide a binding site for proteins (e.g., SH3 domain) interactions (Mayer, 2001). Thus, susceptibility to flanking domains and cellular localization are important factors that should be taken into consideration in predicting homorepeat functions.

F. Homopeptides as Building Blocks of New Materials and Supramolecular Nanostructures

Some extracellular proteins such as silk fibroin, elastin, plant cell wall proteins, and mollusc shell framework proteins contain runs of structure-forming homorepeats such as poly-A, poly-G, or poly-P. Structural studies also revealed that two crystal modifications of polyglycine molecules are relevant to artificial nylon materials (Tormo et al., 1992; Kajava, 1999). Synthetic polyglycine-containing molecules can self-assemble in an aqueous medium into supramolecular structures of different shapes and properties ranging from plane layers to vesicles (Bovin et al., 2008). This opens new avenues in the development of technology designed to

coat surfaces with biocompatible materials, and the creation of supramolecular structures as antiviral agents or drug-delivery vectors. Thus, because of their impressive self-assembly properties and mechanical characteristics, homorepeat runs can provide an important set of material options in the fields of controlled release, biomaterials, and scaffolds for tissue engineering (Altman et al., 2003).

Acknowledgments

The authors thank A. Ahmed, A. Nesterova, and T. Vasselon for critical reading of the manuscript and suggestions. This work was supported by a doctoral fellowship to J.J. from Ministère de l'Education Nationale, de la Recherche et de la Technologie.

References

Ai, H., Lvov, Y. M., Mills, D. K., Jennings, M., Alexander, J. S., Jones, S. A., (2003). Coating and selective deposition of nanofilm on silicone rubber for cell adhesion and growth. *Cell Biochem. Biophys.* **38**, 103–114.

Alba, M. M., Guigo, R., (2004). Comparative analysis of amino acid repeats in rodents and humans. *Genome Res.* **14**, 549–554.

Alba, M. M., Tompa, P., Veitia, R. A., (2007). Amino acid repeats and the structure and evolution of proteins. *Genome Dyn.* **3**, 119–130.

Altman, G. H., Diaz, F., Jakuba, C., Calabro, T., Horan, R. L., Chen, J., et al., (2003). Silk-based biomaterials. *Biomaterials* **24**, 401–416.

Arnott, S., Dover, S.D., Elliott, A., (1967). Structure of beta-poly-L-alanine: refined atomic co-ordinates for an anti-parallel beta-pleated sheet. *J. Mol. Biol* **30**, 201–208.

Arnott, S., Dover, S.D., (1967). Refinement of bond angles of an alpha-helix. *J. Mol. Biol* **30**, 209–212.

Auer, H. E., Doty, P., (1966). The conformational stability of alpha-helical nonpolar polypeptides in solution. *Biochemistry* **5**, 1716–1725.

Bloom, S. M., Fasman, G. D., de Lozé, C., Blout, E. R., (1962). The effect of amino acid composition on the conformations of synthetic polypeptides, polymers and copolymers of L-methionine S-methyl-L-cysteine and L-valine. *J. Am. Chem. Soc.* **84**, 458–463.

Blout, E. R., (1962). Polyamino acids, polypeptides and proteins. In: Stahmann, M. A. (Eds.), University of Wisconsin Press, Madison, p. 275.

Bovin, N. V., Tuzikov, A. B., Chinarev, A. A., (2008). Oligoglycines: materials with unlimited potential for nanotechnologies. *Nanotechnol. Russ.* **3**, 291–302.

Brais, B., Bouchard, J. P., Xie, Y. G., Rochefort, D. L., Chretien, N., Tome, F. M., et al., (1998). Short GCG expansions in the PABP2 gene cause oculopharyngeal muscular dystrophy. *Nat. Genet.* **18**, 164–167.

Brinkmann, B., Klintschar, M., Neuhuber, F., Huhne, J., Rolf, B., (1998). Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. *Am. J. Hum. Genet.* **62**, 1408–1415.

Brouwer, J. R., Willemsen, R., Oostra, B. A., (2009). Microsatellite repeat instability and neurological disease. *Bioessays* **31**, 71–83.

Charlesworth, B., Sniegowski, P., Stephan, W., (1994). The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature* **371**, 215–220.

Chothia, C., (1973). Conformation of twisted beta-pleated sheets in proteins. *J. Mol. Biol.* **75**, 295–302.

Cleary, J. D., Pearson, C. E., (2003). The contribution of cis-elements to disease-associated repeat instability: clinical and experimental evidence. *Cytogenet. Genome Res.* **100**, 25–55.

Collins, D. W., Jukes, T. H., (1994). Rates of transition and transversion in coding sequences since the human-rodent divergence. *Genomics* **20**, 386–396.

Crick, F. H., Rich, A., (1955). Structure of polyglycine II. *Nature* **176**, 780–781.

Csuros, M., Noe, L., Kucherov, G., (2007). Reconsidering the significance of genomic word frequencies. *Trends Genet.* **23**, 543–546.

DeLano, W.,L., (2002). The PyMOL Molecular Graphics System, DeLano Scientific, San Carlos, CA.

Edde, B., Rossier, J., Le Caer, J. P., Desbruyeres, E., Gros, F., Denoulet, P., (1990). Posttranslational glutamylation of alpha-tubulin. *Science* **247**, 83–85.

Eisenberg, D., Schwarz, E., Komaromy, M., Wall, R., (1984). Analysis of membrane and surface protein sequences with hydrophobic moment plot. *J. Mol. Biol.* **179**, 125–142.

Eisenberg, D., (2003). The discovery of the alpha-helix and beta-sheet, the principal structural features of proteins. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 11207–11210.

Ellegren, H., (2000). Microsatellite mutations in the germline: implications for evolutionary inference. *Trends Genet.* **16**, 551–558.

Fan, X., Dion, P., Laganiere, J., Brais, B., Rouleau, G. A., (2001). Oligomerization of polyalanine expanded PABPN1 facilitates nuclear protein aggregation that is associated with cell death. *Hum. Mol. Genet.* **10**, 2341–2351.

Faux, N. G., Bottomley, S. P., Lesk, A. M., Irving, J. A., Morrison, J. R., de la Banda, M. G., et al., (2005). Functional insights from the distribution and role of homopeptide repeat-containing proteins. *Genome Res.* **15**, 537–551.

Faux, N. G., Huttley, G. A., Mahmood, K., Webb, G. I., de la Banda, M. G., Whisstock, J. C., (2007). RCPdb: an evolutionary classification and codon usage database for repeat-containing proteins. *Genome Res.* **17**, 1118–1127.

Feschenko, V. V., Rajman, L. A., Lovett, S. T., (2003). Stabilization of perfect and imperfect tandem repeats by single-strand DNA exonucleases. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 1134–1139.

Fondon, J. W. 3rd, Garner, H. R., (2004). Molecular origins of rapid and continuous morphological evolution. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 18058–18063.

Fowler, D. M., Koulov, A. V., Balch, W. E., Kelly, J. W., (2007). Functional amyloid–from bacteria to humans. *Trends Biochem. Sci.* **32**, 217–224.

Fraser, R. D., Gillespie, J. M., Macrae, T. P., (1973). Tyrosine-rich proteins in keratins. *Comp. Biochem. Physiol. B* **44**, 943–947.

Fraser, R. D.B., MacRae, T. P., (1973). Conformation in Fibrous Proteins and Related Synthetic Polypeptides. Academic Press, London and New York.

Gatti, E., Popolo, L., Vai, M., Rota, N., Alberghina, L., (1994). O-linked oligosaccharides in yeast glycosyl phosphatidylinositol-anchored protein gp115 are clustered in a serine-rich region not essential for its function. *J. Biol. Chem.* **269**, 19695–19700.

Gerber, H. P., Seipel, K., Georgiev, O., Hofferer, M., Hug, M., Rusconi, S., et al., (1994). Transcriptional activation modulated by homopolymeric glutamine and proline stretches. *Science* **263**, 808–811.

Goldstein, D. B., Schlötterer, C., (1999). Microsatellites Evolution and Applications. Oxford University Press, New York.

Green, H., Wang, N., (1994). Codon reiteration and the evolution of proteins. *Proc. Natl. Acad. Sci. U.S.A.* **91**, 4298–4302.

Hamosh, A., Scott, A. F., Amberger, J., Bocchini, C., Valle, D., McKusick, V. A., (2002). Online Mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **30**, 52–55.

Hancock, J. M., Simon, M., (2005). Simple sequence repeats in proteins and their significance for network evolution. *Gene* **345**, 113–118.

Harrington, W. F., Sela, M., (1958). Studies on the structure of poly-L-proline in solution. *Biochim. Biophys. Acta* **27**, 24–41.

Hengen, P., (1995). Purification of his-tag fusion proteins from *Escherichia coli*. *Trends Biochem. Sci.* **20**, 285–286.

Hinman, M. B., Jones, J. A., Lewis, R. V., (2000). Synthetic spider silk: a modular fiber. *Trends Biotechnol.* **18**, 374–379.

Huntley, M., Golding, G. B., (2000). Evolution of simple sequence in proteins. *J. Mol. Evol.* **51**, 131–140.

Huntley, M. A., Golding, G. B., (2002). Simple sequences are rare in the protein data bank. *Proteins* **48**, 134–140.

Janke, C., Rogowski, K., Wloga, D., Regnard, C., Kajava, A. V., Strub, J. M., et al., (2005). Tubulin polyglutamylase enzymes are members of the TTL domain protein family. *Science* **308**, 1758–1762.

Jorda, J., Kajava, A. V., (2009). T-REKS: identification of tandem REpeats in sequences with a K-meanS based algorithm. *Bioinformatics* **25**, 2632–2638.

Kajava, A. V., (1999). Dimorphism of polyglycine I: structural models for crystal modifications. *Acta Crystallogr. D Biol. Crystallogr.* **55**, 436–442.

Kajava, A. V., (2001). Review: proteins with repeated sequence–structural prediction and modeling. *J. Struct. Biol.* **134**, 132–144.

Karlin, S., (1995). Statistical significance of sequence patterns in proteins. *Curr. Opin. Struct. Biol.* **5**, 360–371.

Karlin, S., Brocchieri, L., Bergman, A., Mrazek, J., Gentles, A. J., (2002). Amino acid runs in eukaryotic proteomes and disease associations. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 333–338.

Karlin, S., Burge, C., (1996). Trinucleotide repeats and long homopeptides in genes and proteins associated with nervous system disease and development. *Proc. Natl. Acad. Sci. U.S.A.* **93**, 1560–1565.

Karlin, S., Chen, C., Gentles, A. J., Cleary, M., (2002). Associations between human disease genes and overlapping gene groups and multiple amino acid runs. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 17008–17013.

Katti, M. V., Ranjekar, P. K., Gupta, V. S., (2001). Differential distribution of simple sequence repeats in eukaryotic genome sequences. *Mol. Biol. Evol.* **18**, 1161–1167.

Korschen, H. G., Beyermann, M., Muller, F., Heck, M., Vantler, M., Koch, K. W., et al., (1999). Interaction of glutamic-acid-rich proteins with the cGMP signalling pathway in rod photoreceptors. *Nature* **400**, 761–766.

Kreil, D. P., Kreil, G., (2000). Asparagine repeats are rare in mammalian proteins. *Trends Biochem. Sci.* **25**, 270–271.

Laumonnier, F., Ronce, N., Hamel, B. C., Thomas, P., Lespinasse, J., Raynaud, M., et al., (2002). Transcription factor SOX3 is involved in X-linked mental retardation with growth hormone deficiency. *Am. J. Hum. Genet.* **71**, 1450–1455.

Lavorgna, G., Patthy, L., Boncinelli, E., (2001). Were protein internal repeats formed by "bricolage"? *Trends Genet.* **17**, 120–123.

Levinson, G., Gutman, G. A., (1987). Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol. Biol. Evol.* **4**, 203–221.

Maji, S. K., Perrin, M. H., Sawaya, M. R., Jessberger, S., Vadodaria, K., Rissman, R. A., et al., (2009). Functional amyloids as natural storage of peptide hormones in pituitary secretory granules. *Science* **325**, 328–332.

Marcotte, E. M., Pellegrini, M., Yeates, T. O., Eisenberg, D., (1999). A census of protein repeats. *J. Mol. Biol.* **293**, 151–160.

Mayer, B. J., (2001). SH3 domains: complexity in moderation. *J. Cell. Sci.* **114**, 1253–1263.

Mirkin, S. M., (2006). DNA structures, repeat expansions and human hereditary disorders. *Curr. Opin. Struct. Biol.* **16**, 351–358.

Morton, B. R., (1995). Neighboring base composition and transversion/transition bias in a comparison of rice and maize chloroplast noncoding regions. *Proc. Natl. Acad. Sci. U.S.A.* **92**, 9717–9721.

Nam, Y. S., Petrovic, A., Jeong, K. S., Venkatesan, S., (2001). Exchange of the basic domain of human immunodeficiency virus type 1 rev for a polyarginine stretch expands the RNA binding specificity, and a minimal arginine cluster is required for optimal RRE RNA binding affinity, nuclear accumulation, and trans-activation. *J. Virol.* **75**, 2957–2971.

Nelson, R., Eisenberg, D., (2006). Structural models of amyloid-like fibrils. *Adv. Protein Chem.* **73**, 235–282.

Newfeld, S. J., Tachida, H., Yedvobnick, B., (1994). Drive-selection equilibrium: homopolymer evolution in the Drosophila gene mastermind. *J. Mol. Evol.* **38**, 637–641.

Oma, Y., Kino, Y., Sasagawa, N., Ishiura, S., (2004). Intracellular localization of homopolymeric amino acid-containing proteins expressed in mammalian cells. *J. Biol. Chem.* **279**, 21217–21222.

Orr, H. T., Zoghbi, H. Y., (2007). Trinucleotide repeat disorders. *Annu. Rev. Neurosci.* **30**, 575–621.

Padden, F. J., Keith, H. D., (1965). *J. Appl. Phys.* **36**, 2987–2995.

Pauling, L., Corey, R. B., (1951). The structure of fibrous proteins of the collagen-gelatin group. *Proc. Natl. Acad. Sci. U.S.A.* **37**, 272–281.

Pauling, L., Corey, R. B., (1953). Two rippled-sheet configurations of polypeptide chains, and a note about the pleated sheets. *Proc. Natl. Acad. Sci. U.S.A.* **39**, 253–256.

Perutz, M. F., (1999). Glutamine repeats and neurodegenerative diseases: molecular aspects. *Trends Biochem. Sci.* **24**, 58–63.

Poetsch, A., Molday, L. L., Molday, R. S., (2001). The cGMP-gated channel and related glutamic acid-rich proteins interact with peripherin-2 at the rim region of rod photoreceptor disc membranes. *J. Biol. Chem.* **276**, 48009–48016.

Radivojac, P., Iakoucheva, L. M., Oldfield, C. J., Obradovic, Z., Uversky, V. N., Dunker, A. K., (2007). Intrinsic disorder and functional proteomics. *Biophys. J.* **92**, 1439–1456.

Redeker, V., Levilliers, N., Schmitter, J. M., Le Caer, J. P., Rossier, J., Adoutte, A., et al., (1994). Polyglycylation of tubulin: a posttranslational modification in axonemal microtubules. *Science* **266**, 1688–1691.

Remacha, M., Jimenez-Diaz, A., Bermejo, B., Rodriguez-Gabriel, M. A., Guarinos, E., Ballesta, J. P., (1995). Ribosomal acidic phosphoproteins P1 and P2 are not required for cell viability but regulate the pattern of protein expression in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **15**, 4754–4762.

Rolfsmeier, M. L., Lahue, R. S., (2000). Stabilizing effects of interruptions on trinucleotide repeat expansions in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **20**, 173–180.

Schulz, G. E., Schirmer, R. H., (1979). Principles of Protein Structure. Springer-Verlag, New York.

Sikorski, P., Atkins, E., (2005). New model for crystalline polyglutamine assemblies and their connection with amyloid fibrils. *Biomacromolecules* **6**, 425–432.

Singh, G. P., Chandra, B. R., Bhattacharya, A., Akhouri, R. R., Singh, S. K., Sharma, A., (2004). Hyper-expansion of asparagines correlates with an abundance of proteins with prion-like domains in *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **137**, 307–319.

Siwach, P., Ganesh, S., (2008). Tandem repeats in human disorders: mechanisms and evolution. *Front. Biosci.* **13**, 4467–4484.

Stromme, P., Mangelsdorf, M. E., Shaw, M. A., Lower, K. M., Lewis, S. M., Bruyere, H., et al., (2002). Mutations in the human ortholog of aristaless cause X-linked mental retardation and epilepsy. *Nat. Genet.* **30**, 441–445.

Sumiyama, K., Washio-Watanabe, K., Saitou, N., Hayakawa, T., Ueda, S., (1996). Class III POU genes: generation of homopolymeric amino acid repeats under GC pressure in mammals. *J. Mol. Evol.* **43**, 170–178.

Tormo, J., Puiggali, J., Vives, J., Fita, I., Lloveras, J., Bella, J., et al., (1992). Crystal structure of a helical oligopeptide model of polyglycine II and of other polyamides: acetyl-(glycyl-beta-alanyl)2-NH propyl. *Biopolymers* **32**, 643–648.

Tye, C. E., Rattray, K. R., Warner, K. J., Gordon, J. A., Sodek, J., Hunter, G. K., et al., (2003). Delineation of the hydroxyapatite-nucleating domains of bone sialoprotein. *J. Biol. Chem.* **278**, 7949–7955.

Weber, J. L., Wong, C., (1993). Mutation of human short tandem repeats. *Hum. Mol. Genet.* **2**, 1123–1128.

Wierdl, M., Dominska, M., Petes, T. D., (1997). Microsatellite instability in yeast: dependence on the length of the microsatellite. *Genetics* **146**, 769–779.

CHAPITRE 4

La corrélation entre la perfection des répétitions et leur désordre intrinsèque

Résumé

Les régions parfaitement répétées forment une catégorie à part parmi les répétitions en tandem protéiques. En effet, rares sont celles qui ne subissent pas de mutations susceptibles de dégénérer la séquence d'origine au cours de leur évolution. Cette perfection peut représenter une importance structurelle ou fonctionnelle hautement conservée par une pression sélective, ou être assimilée à des événements évolutifs récents, comme par exemple une réponse rapide à des changements environnementaux. Lors de leur analyse, il a pu être observé que la plupart des répétitions identifiées dans PDB ne présentaient pas de structures associées. Ce chapitre a permis d'étayer ces observations grâce à une approche bioinformatique consistant en la prédiction du désordre intrinsèque dans les répétitions en tandem stockées dans PRDB, et à l'étude des variations du niveau de perfection des répétitions en fonction du degré de désordre.

the FEBS Journal

Protein tandem repeats – the more perfect, the less structured

Julien Jorda[1], Bin Xue[2,3], Vladimir N. Uversky[2,3,4,5] and Andrey V. Kajava[1]

1 Centre de Recherches de Biochimie Macromoléculaire, CNRS UMR-5237, University of Montpellier 1 and 2, France
2 Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN, USA
3 Institute for Intrinsically Disordered Protein Research, Indiana University School of Medicine, Indianapolis, IN, USA
4 Institute for Biological Instrumentation, Russian Academy of Sciences, Pushchino, Moscow Region, Russia
5 Department of Biochemistry and Molecular Biology, Indiana University School of Medicine, Indianapolis, IN, USA

Keywords
bioinformatics; disordered conformation; evolution; protein structure; sequence analysis

Correspondence
A. V. Kajava, Centre de Recherches de Biochimie Macromoléculaire, CNRS, 1919 Route de Mende, 34293 Montpellier, Cedex 5, France
Fax: +33 4 67 521559
Tel: +33 4 67 61 3364
E-mail: andrey.kajava@crbm.cnrs.fr

(Received 23 February 2010, revised 7 April 2010, accepted 12 April 2010)

doi:10.1111/j.1742-4658.2010.07684.x

We analysed the structural properties of protein regions containing arrays of perfect and nearly perfect tandem repeats. Naturally occurring proteins with perfect repeats are practically absent among the proteins with known 3D structures. The great majority of such regions in the Protein Data Bank are found in the proteins designed *de novo*. The abundance of natural structured proteins with tandem repeats is inversely correlated with the repeat perfection: the chance of finding natural structured proteins in the Protein Data Bank increases with a decrease in the level of repeat perfection. Prediction of intrinsic disorder within the tandem repeats in the Swiss-Prot proteins supports the conclusion that the level of repeat perfection correlates with their tendency to be unstructured. This correlation is valid across the various species and subcellular localizations, although the level of disordered tandem repeats varies significantly between these datasets. On average, in prokaryotes, tandem repeats of cytoplasmic proteins were predicted to be the most structured, whereas in eukaryotes, the most structured portion of the repeats was found in the membrane proteins. Our study supports the hypothesis that, in general, the repeat perfection is a sign of recent evolutionary events rather than of exceptional structural and (or) functional importance of the repeat residues.

Introduction

Genome sequencing projects are producing knowledge about a large number of protein sequences. Understanding the biological role of many of these proteins requires information about their 3D structure as well as their evolutionary and functional relationships. At least 14% of all proteins and more than one-third of human proteins carrying out fundamental functions contain arrays of tandem repeats (TRs) [1]. The 3D structures of many of these proteins have already been determined by X-ray crystallography and NMR methods. Fibrous proteins with repeats of two to seven residues (collagen, silk fibroin, keratin, and tropomyosin) were the first objects studied by structural biology methods [2]. Proteins with repeat lengths from 5 to 50 residues gained special interest in the 1990s, when several unusual structural folds, including β-helices [3], β-rolls [4], the horseshoe-shaped structure of leucine-rich-repeat proteins [5], β-propellers [6], and α-helical solenoids [7], were resolved by X-ray crystallography. Many proteins with repeats longer than 30 residues have a 'beads-on-a-string' organization, with each repeat being folded into a globular domain, e.g. zinc

Abbreviations
IDP, intrinsically disordered protein; IDR, intrinsically disordered region; PDB, Protein Data Bank; SCA, spinocerebellar ataxia; TR, tandem repeat.

finger domains [8], immunoglobulin domains [9], and human matrix metalloproteinase [10]. It was noticed that, frequently, proteins with repeats do not have unique, stable 3D structures [11]. Rough estimates propose that half of the regions with TRs may be naturally unfolded [12,13]. Low-complexity regions of eukaryotic proteins that are enriched in repetitive motifs are rare among the known 3D structures from the Protein Data Bank (PDB) [14]. The common structural features, functions and evolution of proteins with TRs have been summarized in several reviews [7,11,15–18].

Perfect TRs occupy a special place among protein repeats, which are usually imperfect because of mutations (substitutions, insertions, and deletions) that have accumulated during evolution. The high level of perfection of repeats can indicate substantial structural and functional importance for each residue in the repeat, as was observed in collagen molecules and some β-roll structures [2,19]. It can also indicate recent evolutionary events that, for example, in pathogens can allow a rapid response to environmental changes and can thus lead to emerging infection threats, and in higher organisms can lead to rapid morphological effects [20].

Perfect and nearly perfect repeats occur in a significant portion of proteins. Recently, by using a newly developed algorithm for *ab initio* identification of TRs, we detected this type of repeat in 9% of proteins in the SwissProt database [21]. To estimate the level of perfection of the TRs, we used a parameter called P_{sim}, which is based on the calculation of Hamming distances between the consensus sequence and aligned repeats of the TR (see Experimental procedures). In this work, we analysed perfect and nearly perfect TRs with $P_{sim} \geq 0.7$.

Specific structural and evolutionary properties of the perfect repeats pose challenges for the annotation of genomic data. First, unlike with the aperiodic globular proteins, prediction of structure–function relationships by sequence similarity cannot be directly applied to the perfect or nearly perfect repeats, owing to their different evolutionary mechanisms. Second, although *ab initio* structural prediction for proteins with TRs generally yields reliable results [11], the very high fidelity of sequence periodicity decreases the accuracy and reliability of the information obtained from the sequence alignment of the repeats. Each position of the perfect repeats is conserved, and this makes it difficult to distinguish between residues that form the interior of the structure and those that face the solvent.

TRs are often found in proteins associated with various human diseases. For example, expansion of homorepeats is the molecular cause of at least 18 human neurological diseases, including myotonic dystrophy 1, Huntington's disease, Kennedy disease (also known as spinal and bulbar muscular atrophy), dentatorubral–pallidoluysian atrophy, and a number of spinocerebellar ataxias (SCAs), such as SCA1, SCA2, Machado–Joseph disease (SCA3), SCA6, SCA7, and SCA17 [22,23]. A number of clinical disorders, including prostate cancer, benign prostatic hyperplasia, male infertility, and rheumatoid arthritis, are associated with polymorphisms in the length of the polyglutamine and polyglycine repeats of the androgen receptor [24].

Thus, proteins with perfect or nearly perfect TRs play important functional roles, are abundant in genomes, are related to major health threats, and, at the same time, represent a challenge for *in silico* identification of their structures and functions. The objective of this work was a systematic bioinformatics analysis of arrays of perfect or nearly perfect TRs to obtain a global view of their structural properties.

Results and Discussion

The 3D structures of naturally occurring proteins with perfect repeats are practically absent in the PDB

Our analysis shows that, among 20 800 sequences of the nonredundant PDB (95% identity), only nine naturally occurring proteins (0.04%) have perfect TRs with $P_{sim} = 1$ (Table 1). Furthermore, these arrays of TRs are short (less than 19 residues), and they are missing from the determined structures representing regions with blurred electron density. A common reason for missing electron density is that the unobserved atom, side chain, residue or region fails to scatter X-rays coherently, because of variation in position from one protein to the next; for example, the unobserved atoms can be flexible or disordered. Two proteins are exceptions to this: (a) an antibody molecule in which the

Table 1. Number of structured and unstructured regions found for each range of P_{sim} values in the PDB TR dataset. The following tags were assigned to each analysed region with TRs: Sn and Sd, fragments containing secondary structures from natural and designed proteins, respectively; Ln and Ld, fragments connecting secondary structures from natural and designed proteins, respectively; Un and Ud, fragments whose structure was not determined from natural and designed proteins, respectively.

P_{sim} ranges	Sn	Ln	Un	Sd	Ld	Ud
$P_{sim} = 1.0$	0	2	7	16	4	14
$0.9 \leq P_{sim} < 1.0$	1	2	8	20	2	5
$0.8 \leq P_{sim} < 0.9$	17	8	31	24	1	12

Gly-rich TR represents a crosslink between two domains (PDB code: 1F3R) [25]; and (b) a substrate with an $(Arg\text{-}Ser)_8$ tract that was cocrystallized with protein kinase (PDB code: 3BEG) [26]. This Arg-rich peptide, being alone in solution, will most probably be unstructured, owing to the absence of nonpolar residues and the presence of eight Arg residues carrying a charge of the same sign. Thus, this analysis suggested that regions of natural proteins with perfect repeats have a tendency to be unstructured.

To investigate this tendency, we analysed further the regions with less perfect TRs. The TRs with $0.9 \leq P_{sim} < 1.0$ are also rare among natural proteins of the PDB. Furthermore, the conformations of almost all of them have not been resolved by X-ray crystallography, because they are located in regions with missing electron density. Only one of them, human CD3-e/d dimer (PDB: 1XIW) [27], has a short region of two nine-residue repeats corresponding to a loop followed by β-strand. We also analysed TRs with $0.8 \leq P_{sim} < 0.9$, and found 17 TRs of natural proteins with the 3D structures (Table 1). In addition to relatively short regions of fewer than 20 residues, corresponding to the α-helical elements, we also found longer regions that form an immunoglobulin-like structure (1D2P) [28], a β-roll (1GO7) [29], an α-solenoid (2AJA) [30], and an unusual long β-hairpin (1JHN) [31] (Fig. 1). Three of these four structures are formed by bacterial proteins.

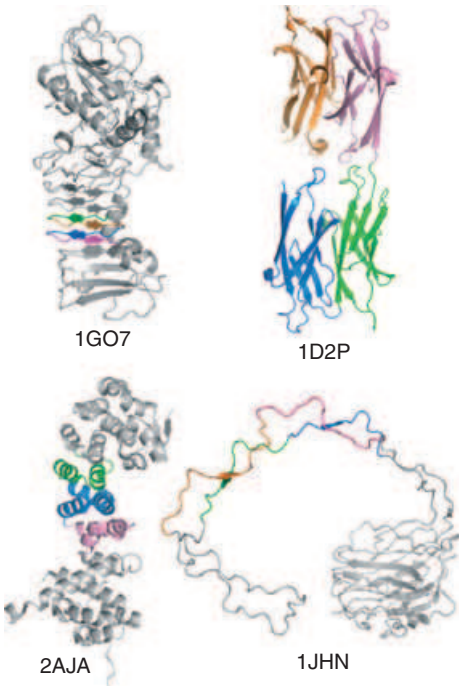


Fig. 1. The 3D structures of proteins with almost perfect TRs. Repeat regions are shown in colour.

De novo designed proteins with perfect repeats fold into stable 3D structures

In the PDB, majority (80%) of the proteins with perfect TRs are proteins designed *de novo* (Table 1). The TR of a large proportion of these proteins fold into the well-defined repetitive 3D structures such as collagen triple helices, α-helical coiled coils, and α-helical solenoids [2,17]. The fact that the designed perfect TRs can form the stable 3D structures indicates that the absence of such structures in natural proteins results from evolution and not from problems with their folding propensities *per se*.

Prediction of intrinsically disordered regions in SwissProt supports the tendency of TRs to be unfolded

The ability of TRs to be structured or disordered was further tested by using a larger dataset extracted from SwissProt. The analysed dataset of TRs from the Protein Repeat DataBase (http://bioinfo.montp.cnrs.fr/?r=repeatDB) was filled in by the T-REKS program [21]. The TRs with P_{sim} values ranging from 0.7 to 1 consist of 51 685 repeats found in 33 151 proteins, which represent 9.1% of all proteins in the SwissProt release of January 2009 (364 403 sequences). The level of intrinsic disorder in these repeats and repeat-containing proteins was evaluated by using several computational tools.

Compositional profiling

Intrinsically disordered proteins (IDPs) and intrinsically disordered regions (IDRs) are known to be different from structured globular proteins and domains with regard to many attributes, including amino acid composition, sequence complexity, hydrophobicity, charge, flexibility, and type and rate of amino acid substitutions over evolutionary time. For example, IDPs/IDRs are significantly depleted in a number of so-called order-promoting residues, including bulky hydrophobic (Ile, Leu, and Val) and aromatic (Trp, Tyr, and Phe) residues, which would normally form the hydrophobic core of a folded globular protein, and also possess low contents of Cys and Asn residues. On the other hand, IDPs/IDRs were shown to be substantially enriched in so-called disorder-promoting residues: Ala, Arg, Gly, Gln, Ser, Pro, Glu, and Lys [32–36]. These biases in the amino acid composition of

IDPs and IDRs can be visualized using a normalization procedure known as compositional profiling [32,33,37]. In brief, compositional profiling is based on the evaluation of the $(C_{s1} - C_{s2})/C_{s2}$ values, where C_{s1} is the content of a given residue in a set of interest (regions and proteins with TRs), and C_{s2} is the corresponding value for the reference dataset (set of ordered proteins or set of well-characterized IDPs). Negative values of the profiling correspond to residues that are depleted in a given dataset in comparison with a reference dataset, and the positive values correspond to residues that are overrepresented in the set of interest.

Figure 2 compares the amino acid compositions of (a) all TRs analysed in this study, (b) proteins containing these TRs and (c) a dataset of IDPs with the compositions of ordered proteins. The datasets of IDPs and fully structured proteins were taken from our previous analysis [38,39]. This shows that the compositions of proteins containing TRs and of TRs themselves are different from the compositions of ordered proteins. They follow the trend for IDPs, being generally depleted in major order-promoting residues. This tendency for disorder is stronger for the TRs, indicating that they contribute to this trend. At the same time, the amino acid compositions of the TRs have a bias when compared with the compositions of 'typical' disordered proteins (Fig. 2). TRs have an especially low occurrence of order-promoting Met and the disorder-promoting charged residues Asp, Glu, and Lys. On the other hand, TRs are highly enriched in Cys and the disorder-promoting Pro, Gly, Ser, and His.

To test the tendency of TRs to be disordered as a function of their level of perfection, the TRs were subdivided into four subsets according to their P_{sim} values [$0.7 < P_{sim} \leq 0.8$ (32691 TRs), $0.8 < P_{sim} \leq 0.9$ (8322 TRs), $0.9 < P_{sim} \leq 1.0$ (1471 TRs), and homorepeats with $P_{sim} = 1.0$ (5259 TRs). Homorepeats were analysed separately from the other TRs, because they significantly outnumber the other types of repeats, and having them in the same group would obscure the effect related to the other repeats. The amino acid compositions of these subsets were compared with the compositions of fully structured proteins. Figure 3 represents the results of compositional profiling for TRs with different level of perfection. Both homorepeats and the other TRs show the same trend. With the increase in the perfection of the repeated segment, the amount of order-promoting residues is gradually reduced, whereas the relative contents of disorder-promoting polar residues are gradually increased.

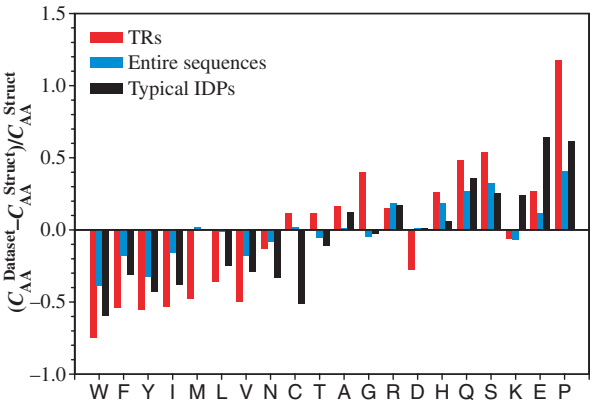


Fig. 2. Compositional profiling of TRs, entire sequences of proteins containing these TRs, and a set of fully disordered proteins from DisProt in comparison with the composition of fully structured proteins from the PDB. C_{AA}^{Struct} is the content of a given amino acid in the set of structured proteins; $C_{AA}^{Dataset}$ is the content of this amino acid in the dataset of interest. Amino acids are arranged in order of decreasing structure-promoting ability as suggested by the TopIDP scale [37].

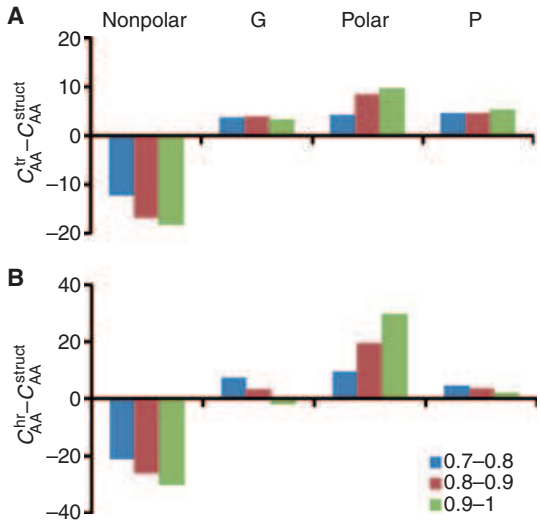


Fig. 3. (A) Differences in amino acid compositions between TRs, subdivided into groups with different levels of repeat perfection and fully structured proteins. The homorepeats are analysed separately (B), owing to their unusually high occurrence in comparison to the other TRs. For this purpose, a dataset of perfect and cryptic homorepeats was created and subdivided into three groups depending on the P_{sim} values. C_{AA}^{tr} and C_{AA}^{hr} are the contents of a given amino acid in the set of TRs (excluding homorepeats) and only homorepeats, respectively. Amino acids are arranged in four sets: order-promoting aromatic and aliphatic amino acids (Trp, Phe, Tyr, Ile, Met, Leu, Val, and Ala) which are denoted as nonpolar; order-neutral Gly, disorder-promoting polar residues (Asn, Cys, Thr, Gln, Ser, Arg, Asp, His, Glu, and Lys) and disorder-promoting nonpolar Pro.

The contents of Gly and Pro residues do not change significantly.

Prediction of intrinsic disorder

As the compositional profiling showed that TRs and repeat-containing proteins have a noticeable increase in the number of disorder-promoting residues, we further analysed the abundance of predicted intrinsic disorder in these sequences with several computational tools, including the PONDR® VLXT [34,40] and VSL2 [41,42] algorithms, as well as predictors such as IUPRED [43,44], FOLDINDEX [45], and TOPIDP [37]. The results of this analysis are summarized in Table 2, which clearly shows that both TRs and repeat-containing proteins are highly disordered. Furthermore, TRs have higher percentage of disordered residues than the entire TR-containing sequences. Prediction of intrinsic disorder also confirmed an observation that the amounts of disorder in both datasets increase with increases in the repeat perfection (Table 2).

This observation is further illustrated by the distributions of values representing the number of predicted disorder residues divided by the number of residues in the considered region (Fig. 4). These distributions are generated for TR regions of different levels of perfection (Fig. 4A) and for the corresponding repeat-containing proteins (Fig. 4B). Figure 4A shows that all analysed TRs are highly disordered, irrespective of the level of their perfection. At the same time, as the perfection of TRs increases, the relative content of disorder also increases. For example, at least 70% of TRs with $0.7 < P_{sim} \leq 0.8$ are predicted to have disorder ratios of more than 0.95. For TRs with $0.8 < P_{sim} \leq 0.9$, this percentage increases to 85%, for those with $0.9 < P_{sim} \leq 1.0$ it is 86%, and for perfect homorepeats it reaches 97% (Fig. 4A). Figure 4B shows that only 6% of the whole sequences of proteins containing perfect repeats are well structured (disorder ratio less than 0.2). The rest of these sequences have widespread disorder ratios, ranging from 0.25 to 1. Proteins containing the least perfect repeats ($0.7 < P_{sim} \leq 0.8$), about 5%, are almost evenly distributed among the various disorder ratios. Thus, perfect repeats preferentially occur in proteins that have disorder ratios of more than 0.2 and are poorly represented in more structured proteins, whereas less perfect repeats are equally probable in sequences with different disorder ratios.

Intrinsic disorder of tandem repeats across species and subcellular localizations

The PONDR® VLXT predictor and TopIDP index were used to establish variation of the disorder level among TRs of viral, eukaryotic and prokaryotic proteins. The tested dataset included TRs with $P_{sim} \geq 0.9$ identified in SwissProt. The homorepeats were excluded and analysed separately from the other TRs, because their predominant occurrence in eukaryotic proteins would obscure the results. Prior to the analysis, the redundancy of the dataset related to the existence of protein sequences from different strains of the same species (especially for bacteria and viruses) had been filtered out by using the species name, consensus motif, and number and location of repeats. As a result, the dataset contained 245 repeats from prokaryotic proteins, 1059 repeats from eukaryotic proteins, and 70 repeats

Table 2. Analysis of intrinsic disorder distribution in TRs and TR-containing proteins.

	P_{sim} = 0.7–0.8	P_{sim} = 0.8–0.9	P_{sim} = 0.9–1	Homorepeats
TRs				
Total no.	34 286	5519	1382	5259
Average length	25.5	41.0	59.1	13.8
Intrinsic disorder ratio (%): VSL2	80.4	88.6	88.9	98.4
Intrinsic disorder ratio (%): IUPRED	56.0	62.7	67.2	86.5
Intrinsic disorder ratio (%): FOLDINDEX	62.4	68.6	70.3	79.9
Intrinsic disorder ratio (%): TOPIDP	85.6	88.8	91.1	74.4
Sequences[a]				
Total no.	25 649	4915	1295	3663
Average length	643.4	752.0	840.2	790.4
Intrinsic disorder ratio (%): VSL2	49.3	58.6	57.0	61.6
Intrinsic disorder ratio (%): IUPRED	32.1	41.7	41.7	45.4
Intrinsic disorder ratio (%): FOLDINDEX	46.6	52.3	52.3	52.7
Intrinsic disorder ratio (%): TOPIDP	71.2	75.3	72.2	74.9

[a] Whole proteins containing these TRs.

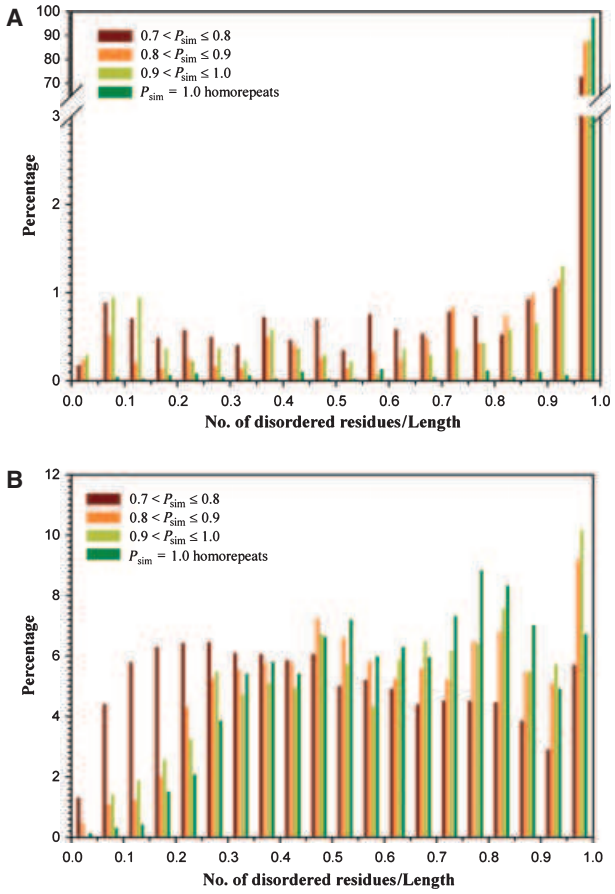


Fig. 4. Length distribution of predicted disordered segments. (A) Length distribution of predicted disorder for four groups of TRs. (B) Length distribution of predicted disorder for whole protein sequences containing the TRs in four groups.

from viral proteins. Our analysis shows that TRs from all species have a tendency to be unstructured (Table 3). At the same time, TRs from eukaryotic proteins have ratios of disordered proteins that are slightly higher than those of TRs from viral or prokaryotic proteins.

Table 3. Variation in the disorder level among TRs of viral, eukaryotic and prokaryotic proteins.

	Prokaryotes (%)	Viruses (%)	Eukaryotes (%)
PONDR® VLXT[a]	84	85.0	88.4
TopIDP[b]	71.4	72.4	77.8

[a] Protein regions with VLXT cumulative distribution function distances of less than 0 are identified as disordered. The P_{sim} range for this dataset is 0.9–1. Disorder level is estimated as percentage of residues predicted to be disordered. [b] Protein regions with TopIDP values of less than 0 are identified as disordered. The P_{sim} range for this dataset is 0.9–1. The disorder level is estimated as the percentage of TRs with negative TopIDP values.

The ratio of disordered repeats was also investigated as a function of the subcellular localization of corresponding repeat-containing proteins. We performed this analysis separately for homorepeats and the other TRs of SwissProt with $P_{sim} \geq 0.8$. The obtained distributions among cellular compartments were similar in these two datasets; therefore, Table 4 represents the combined results for both types of repeat. The lowest proportion of disordered repeats (54.3%) was found in the cytoplasmic proteins of prokaryotes (Table 4). The ratio increases from the cytoplasm to the cellular exterior, being equal to 72.3% and 83.6% in membrane and secreted proteins, respectively. A survey of amino acid sequences of the bacterial cytoplasmic repeats that were predicted to be structured revealed a large number (90 TRs) of $(GGM)_n$ repeats. These repeats are located at the C-terminal extremity of the GroEL chaperone and play important roles in the refolding of proteins [46]. In the crystal structure of the GroEL complex, these C-terminal tails have blurred electron density inside the complex chamber. This suggests that, inside the GroEL complex, they are disordered. Such repeats are also found in mitochondria of eukaryotes in HSP60, a eukaryotic homolog of GroEL. The cytoplasmic TRs of prokaryotes with excluded GGM repeats still have the highest percentage of predicted structured regions among the cellular compartments.

In eukaryotes, the ratio of disorder varies with cellular localization. The lowest level of TR disorder is found in membrane proteins, followed by secreted and nuclear proteins. The cytoplasmic TRs are the most disordered in eukaryotes (82%). The high percentage of ordered TRs in membrane proteins suggests that they may form part of transmembrane regions. However, our analysis revealed that only 12% of them were predicted to be within the transmembrane regions.

Conclusions

TRs of proteins with known 3D structures are generally imperfect. They have consensus sequences with both conserved and variable residues. Analysis of these 3D structures reveals that each sequence repeat corresponds to a repetitive structural unit and that their tandem arrangement yields elongated regular structures [11]. The conserved residues of repeats are frequently located inside the structure, because they are important for its stability, whereas variable residues are exposed on the protein surface. This might lead one to expect that all residues of highly perfect TRs would be conserved, because of their important structural roles. However, our present study shows that this rule does

Table 4. Abundance of disordered repeats as a function of the subcellular localization of corresponding repeat-containing proteins. Membrane localization for eukaryotes combines 'membrane' and 'cell membrane' terms from SwissProt.

	Prokaryotes			Eukaryotes			
	Cytoplasm	Membrane	Secreted	Nucleus	Cytoplasm	Membrane	Secreted
Ratio of TopIDP (%)	54.3	72.3	83.6	74	81.2	60.2	72.7
Number of TRs	459	264	140	3650 (among them 1898 homorepeats)	1181 (476 homorepeats)	1436 (637 homorepeats)	782 (178 homorepeats)

not apply for perfect or almost perfect repeats. We have shown that increasing repeat perfection correlates with a stronger tendency to be unstructured. This result is in agreement with the previous conclusion about a strong association between homorepeats and unstructured regions [13]. Coding for protein disorder is more permissive, and does not require exact sequence motifs, in contrast to the coding for the 3D structures. It allows higher variability in amino acid sequences. Therefore, TR perfection cannot be explained by the need to encode disordered conformations. The other reason for high conservation of residues may be their functional importance, such as the involvement of all or almost all residues of the repeat in interactions with the other molecule. This scenario is also unlikely, because only some residues of the repeat motif can be in contact with the other molecule and will therefore be conserved owing to the specific functional interactions. Thus, the structural role and functional interactions of TRs, even when they are considered together, cannot explain repeat perfection. This consideration favours explanations based on evolutionary reasons. For example, the perfection of TRs may reflect their recent appearance during evolution. It is known that the repetitive regions, such as microsatellites, evolve more rapidly (mutational rate is 10^6-fold higher) than the unique parts of genes [47,48]. This generic instability of TRs, together with the structurally permissive nature of their disordered state, may increase the probability of newly emerged repeats being fixed during evolution, and allow a rapid response to environmental changes [12,49,50]. The evolutionary explanation for repeat perfection is in line with the previously suggested hypothesis that intrinsically disordered proteins may evolve by repeat extension [12]. Functional constraints, such as the ability of TRs to bind to the repetitive surfaces of other molecules or to provide a spacer that can vary in length in rapid response to environmental threats, may play a role in their selection during evolution.

Our results suggest that, up to a certain level of repeat perfection, there are structural reasons for conservation of residues and that these types of residue may stabilize the unique 3D structure. However, when a certain threshold of the conserved residues in the repeat is exceeded, the repetitive regions of proteins are predominantly disordered, and the main reason for residue conservation in TRs may change from a structural to an evolutionary one. This hypothesis can be tested by further evolutionary analysis. The results of our analysis also lead to a practical recommendation for prediction of the structures and functions of proteins. If one sees a perfect TR in a protein of interest, this region is most probably unstructured by itself but still may adopt 3D structures upon binding to the other molecular partners.

Methods

Detection of protein tandem repeats

The program T-REKS was used for *ab initio* identification of the TRs in protein sequences (http://bioinfo.montp.cnrs.fr/?r=t-reks) [21]. This method is based on clustering of lengths between identical short strings by use of a *K*-means algorithm. Benchmarks on several sequence datasets showed that T-REKS detects the TRs in protein sequences better than the other tested software. Several parameters of the program can be defined by users. Among them are the allowed percentage of length variability, Δl (the default value of Δl used in this analysis is equal to 20% of the repeat length). It was chosen on the basis of analysis of known repeats of biological importance. The program also evaluates the level of sequence similarity between the identified repeats of each run by using the following approach. On the basis of multiple sequence alignment of the repeats constituting a given tandem array, T-REKS deduces a consensus sequence and uses it as a reference for similarity calculation. In this alignment, an indel is considered as an additional 21st type of residue. We calculate a Hamming distance, D_i [51], between the consensus sequence and a repeat, R_i, with $1 \leq i \leq m$, where m is the number of repeats in one run. Then, we define a similarity coefficient for the whole alignment as $P_{\mathrm{sim}} = (N - \sum_{i=1}^{m} D_i)/N$, with $N = ml$ (l is the repeat length). The P_{sim} value can be used to estimate the level of perfection of the TR. The maximal value, $P_{\mathrm{sim}} = 1$, corresponds to the run of the perfect repeats. In

this work, we analysed TRs with $P_{sim} \geq 0.70$. The minimal length of TR regions was determined by estimation of the expected number of perfect TRs found by chance in a random sequence dataset (of the SwissProt size), which follows a binomial distribution approximated by a Poisson distribution [21]. The lengths for which the expected number of perfect TRs is equal or close to zero correspond, respectively, to nine residues for homorepeat regions and 14 residues for the other repeats.

Two databases were analysed: (a) a nonredundant databank of sequences (with less than 95% identity) from the July 2008 release of the PDB [52]; and (b) SwissProt, release of January 2009 [53]. During analysis of the PDB, artificial His-tags attached to proteins were not taken into consideration. Short peptides of fewer than 20 residues that represent ligands bound to proteins were also not taken into consideration. Several errors in PDB sequence annotations were found and excluded from the analysis. The 3D structures of the remaining 164 repeats, divided into three groups by the level of perfection ($P_{sim} = 1$, $1 > P_{sim} \geq 0.9$, and $0.9 > P_{sim} \geq 0.8$), were analysed manually (Table 1). The identified TRs were stored in the Protein Repeat DataBase (http://bioinfo.montp.cnrs.fr/?r = repeatDB).

Compositional profiling

Biases in the amino acid compositions of IDPs and IDRs can be visualized by using a normalization procedure known as compositional profiling [32,33,37]. Compositional profiling is based on the evaluation of the $(C_{s1} - C_{s2})/C_{s2}$ values, where C_{s1} is the content of a given residue in a set of interest (regions and proteins with TRs), and C_{s2} is the corresponding value for the reference dataset (set of ordered proteins or set of well-characterized IDPs). Datasets of fully disordered and structured proteins were taken from the DisProt and PDB databases [38,39].

Prediction of disordered regions

Two disorder predictors from the PONDR® family, VLXT [34,40] and VLS2 [41,42], as well as a set of orthogonal predictors such as IUPRED [43,44], FOLDINDEX [45], and TopIDP [37], were used to analyse the differences between the above-described datasets. PONDR® VLXT is an integration of three artificial neural networks that were designed for each of the termini and the internal part of the sequences, respectively. Each individual predictor was trained in a dataset containing only the corresponding part of sequences. The inputs of the neural networks were amino acid composition, hydropathy, net charge, flexibility, and coordination number. The final prediction result was an average over the overlapping regions of three independent predictors [34,40].

PONDR® VSL2 utilized support vector machines to train on long sequences with length ≥ 30 and on short sequences of length ≤ 30, separately. The inputs included hydropathy, net charge, flexibility, coordination number, the position-specific score matrix from PSI-BLAST [54], and predicted secondary structures from PHDSEC [55] and PSIPRED [56]. The final output was a weighted average with the weights determined by a metapredictor [41,42]. VSL2 is accurate in detecting both short and long disordered sequences.

IUPRED assumes that globular proteins have larger interresidue interactions than disordered proteins [43,44]. Hence, it is possible to derive a sequence-based pairwise interaction matrix from globular proteins of known structures. The averaged energy based on this pairwise interaction matrix for globular proteins should be different from that of disordered proteins.

FOLDINDEX was developed from the charge–hydrophobicity plot [35] by adding the technique of sliding windows [45]. The charge–hydrophobicity plot was designed to determine whether a protein is disordered or not as a whole [35]. By application of a sliding window of 21 amino acids centred at a specific residue, the position of this segment on the charge–hydrophobicity plot can be calculated, and the distance of this position from the boundary line is taken as an indication of whether the central residue is disordered or not [45].

The TopIDP index is an amino acid scale that discriminates between order and disorder [37]. It is based on a set of general intrinsic properties of amino acids that are responsible for the absence of ordered structure in IDPs. The corresponding TopIDP score for each amino acid along the sequence is an average over a sliding window of 21 residues. It reflects the conditional possibility of disordered status for the central amino acid in the sliding window [37].

All of these predictors calculate a prediction score for each residue in the sequence. When the threshold value of the prediction score was set up, all of the residues whose prediction scores were higher than the threshold value were assigned as disordered, and the lower-score residues were assigned as structured.

Acknowledgements

This work was supported in part by grants R01 LM007688-01A1 (to V. N. Uversky) and GM071-714-01A2 (to V. N. Uversky) from the National Institute of Health, grant EF 0849803 (to V. N. Uversky) from the National Science Foundation and the Program of the Russian Academy of Sciences for 'Molecular and Cellular Biology' (to V. N. Uversky). We gratefully acknowledge the support of the IUPUI Signature Centres Initiative. This work was also supported by Ministère de l'Education Nationale, de la Recherche et de la Technologie (MENRT) grant to

J. Jorda. We thank A. Ahmed for critical reading of the manuscript and suggestions.

References

1 Pellegrini M, Marcotte EM & Yeates TO (1999) A fast algorithm for genome-wide analysis of proteins with repeated sequences. *Proteins* **35**, 440–446.
2 Fraser RDB & MacRae TP (1973) *Conformation in Fibrous Proteins and Related Synthetic Polypeptides*. Academic Press, London.
3 Yoder MD, Lietzke SE & Jurnak F (1993) Unusual structural features in the parallel beta-helix in pectate lyases. *Structure* **1**, 241–251.
4 Baumann U, Wu S, Flaherty KM & McKay DB (1993) Three-dimensional structure of the alkaline protease of *Pseudomonas aeruginosa*: a two-domain protein with a calcium binding parallel beta roll motif. *EMBO J* **12**, 3357–3364.
5 Kobe B & Kajava AV (2001) The leucine-rich repeat as a protein recognition motif. *Curr Opin Struct Biol* **11**, 725–732.
6 Fulop V & Jones DT (1999) Beta propellers: structural rigidity and functional diversity. *Curr Opin Struct Biol* **9**, 715–721.
7 Groves MR & Barford D (1999) Topological characteristics of helical repeat proteins. *Curr Opin Struct Biol* **9**, 383–389.
8 Lee MS, Gippert GP, Soman KV, Case DA & Wright PE (1989) Three-dimensional solution structure of a single zinc finger DNA-binding domain. *Science* **245**, 635–637.
9 Sawaya MR, Wojtowicz WM, Andre I, Qian B, Wu W, Baker D, Eisenberg D & Zipursky SL (2008) A double S shape provides the structural basis for the extraordinary binding specificity of Dscam isoforms. *Cell* **134**, 1007–1018.
10 Elkins PA, Ho YS, Smith WW, Janson CA, D'Alessio KJ, McQueney MS, Cummings MD & Romanic AM (2002) Structure of the C-terminally truncated human ProMMP9, a gelatin-binding matrix metalloproteinase. *Acta Crystallogr D Biol Crystallogr* **58**, 1182–1192.
11 Kajava AV (2001) Review: proteins with repeated sequence – structural prediction and modeling. *J Struct Biol* **134**, 132–144.
12 Tompa P (2003) Intrinsically unstructured proteins evolve by repeat expansion. *Bioessays* **25**, 847–855.
13 Simon M & Hancock JM (2009) Tandem and cryptic amino acid repeats accumulate in disordered regions of proteins. *Genome Biol* **10**, R59.1–R59.16.
14 Huntley MA & Golding GB (2002) Simple sequences are rare in the Protein Data Bank. *Proteins* **48**, 134–140.
15 Andrade MA & Bork P (1995) HEAT repeats in the Huntington's disease protein. *Nat Genet* **11**, 115–116.
16 Heringa J (1998) Detection of internal repeats: how common are they? *Curr Opin Struct Biol* **8**, 338–345.
17 Kobe B & Kajava AV (2000) When protein folding is simplified to protein coiling: the continuum of solenoid protein structures. *Trends Biochem Sci* **25**, 509–515.
18 Matsushima N, Yoshida H, Kumaki Y, Kamiya M, Tanaka T, Izumi Y & Kretsinger RH (2008) Flexible structures and ligand interactions of tandem repeats consisting of proline, glycine, asparagine, serine, and/or threonine rich oligopeptides in proteins. *Curr Protein Pept Sci* **9**, 591–610.
19 Aachmann FL, Svanem BI, Guntert P, Petersen SB, Valla S & Wimmer R (2006) NMR structure of the R-module: a parallel beta-roll subunit from an *Azotobacter vinelandii* mannuronan C-5 epimerase. *J Biol Chem* **281**, 7350–7356.
20 Fondon JW III & Garner HR (2004) Molecular origins of rapid and continuous morphological evolution. *Proc Natl Acad Sci USA* **101**, 18058–18063.
21 Jorda J & Kajava AV (2009) T-REKS: identification of Tandem REpeats in sequences with a K-meanS based algorithm. *Bioinformatics* **25**, 2632–2638.
22 Cummings CJ & Zoghbi HY (2000) Trinucleotide repeats: mechanisms and pathophysiology. *Annu Rev Genomics Hum Genet* **1**, 281–328.
23 Cummings CJ & Zoghbi HY (2000) Fourteen and counting: unraveling trinucleotide repeat diseases. *Hum Mol Genet* **9**, 909–916.
24 McEwan IJ (2001) Structural and functional alterations in the androgen receptor in spinal bulbar muscular atrophy. *Biochem Soc Trans* **29**, 222–227.
25 Kleinjung J, Petit MC, Orlewski P, Mamalaki A, Tzartos SJ, Tsikaris V, Sakarellos-Daitsiotis M, Sakarellos C, Marraud M & Cung MT (2000) The third-dimensional structure of the complex between an Fv antibody fragment and an analogue of the main immunogenic region of the acetylcholine receptor: a combined two-dimensional NMR, homology, and molecular modeling approach. *Biopolymers* **53**, 113–128.
26 Ngo JC, Giang K, Chakrabarti S, Ma CT, Huynh N, Hagopian JC, Dorrestein PC, Fu XD, Adams JA & Ghosh G (2008) A sliding docking interaction is essential for sequential and processive phosphorylation of an SR protein by SRPK1. *Mol Cell* **29**, 563–576.
27 Arnett KL, Harrison SC & Wiley DC (2004) Crystal structure of a human CD3-epsilon/delta dimer in complex with a UCHT1 single-chain antibody fragment. *Proc Natl Acad Sci USA* **101**, 16268–16273.
28 Deivanayagam CC, Rich RL, Carson M, Owens RT, Danthuluri S, Bice T, Hook M & Narayana SV (2000) Novel fold and assembly of the repetitive B region of the *Staphylococcus aureus* collagen-binding surface protein. *Structure* **8**, 67–78.
29 Hege T, Feltzer RE, Gray RD & Baumann U (2001) Crystal structure of a complex between *Pseudomonas*

aeruginosa alkaline protease and its cognate inhibitor: inhibition by a zinc-NH2 coordinative bond. *J Biol Chem* **276**, 35087–35092.
30 Kuzin AP, Chen Y, Acton T, Xiao R, Conover KMC, Kellie R, Montelione GT, Tong L & Hunt JF (2010) X-Ray structure of an ankyrin repeat family protein Q5ZSV0 from *Legionella pneumophila*., doi:10.2210/pdb2aja/pdb.
31 Schrag JD, Bergeron JJ, Li Y, Borisova S, Hahn M, Thomas DY & Cygler M (2001) The structure of calnexin, an ER chaperone involved in quality control of protein folding. *Mol Cell* **8**, 633–644.
32 Vacic V, Uversky VN, Dunker AK & Lonardi S (2007) Composition Profiler: a tool for discovery and visualization of amino acid composition differences. *BMC Bioinformatics* **8**, 211.1–211.7.
33 Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, Oldfield CJ, Campen AM, Ratliff CM, Hipps KW *et al*. (2001) Intrinsically disordered protein. *J Mol Graph Model* **19**, 26–59.
34 Romero P, Obradovic Z, Li X, Garner EC, Brown CJ & Dunker AK (2001) Sequence complexity of disordered protein. *Proteins* **42**, 38–48.
35 Uversky VN, Gillespie JR & Fink AL (2000) Why are 'natively unfolded' proteins unstructured under physiologic conditions? *Proteins* **41**, 415–427.
36 Radivojac P, Iakoucheva LM, Oldfield CJ, Obradovic Z, Uversky VN & Dunker AK (2007) Intrinsic disorder and functional proteomics. *Biophys J* **92**, 1439–1456.
37 Campen A, Williams RM, Brown CJ, Meng J, Uversky VN & Dunker AK (2008) TOP-IDP-scale: a new amino acid scale measuring propensity for intrinsic disorder. *Protein Pept Lett* **15**, 956–963.
38 Xue B, Li L, Meroueh SO, Uversky VN & Dunker AK (2009) Analysis of structured and intrinsically disordered regions of transmembrane proteins. *Mol Biosyst* **5**, 1688–1702.
39 Xue B, Oldfield CJ, Dunker AK & Uversky VN (2009) CDF it all: consensus prediction of intrinsically disordered proteins based on various cumulative distribution functions. *FEBS Lett* **583**, 1469–1474.
40 Romero P, Obradovic Z, Kissinger C, Villafranca J & Dunker A (1997) Identifying disordered regions in proteins from amino acid sequence. *Proc IEEE Int Conf Neural Networks* **1**, 90–95.
41 Peng K, Radivojac P, Vucetic S, Dunker AK & Obradovic Z (2006) Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics* **7**, 208.1–208.17.
42 Obradovic Z, Peng K, Vucetic S, Radivojac P & Dunker AK (2005) Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Proteins* **61**(Suppl 7), 176–182.
43 Dosztanyi Z, Csizmok V, Tompa P & Simon I (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* **21**, 3433–3434.
44 Dosztanyi Z, Csizmok V, Tompa P & Simon I (2005) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J Mol Biol* **347**, 827–839.
45 Prilusky J, Felder CE, Zeev-Ben-Mordehai T, Rydberg EH, Man O, Beckmann JS, Silman I & Sussman JL (2005) FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics* **21**, 3435–3438.
46 Tang YC, Chang HC, Roeben A, Wischnewski D, Wischnewski N, Kerner MJ, Hartl FU & Hayer-Hartl M (2006) Structural features of the GroEL–GroES nano-cage required for rapid folding of encapsulated protein. *Cell* **125**, 903–914.
47 Buard J & Vergnaud G (1994) Complex recombination events at the hypermutable minisatellite CEB1 (D2S90). *EMBO J* **13**, 3203–3210.
48 Weber JL & Wong C (1993) Mutation of human short tandem repeats. *Hum Mol Genet* **2**, 1123–1128.
49 Ellegren H (2000) Microsatellite mutations in the germline: implications for evolutionary inference. *Trends Genet* **16**, 551–558.
50 Williamson MP (1994) The structure and function of proline-rich regions in proteins. *Biochem J* **297** (Pt 2), 249–260.
51 Hamming R (1950) Error detecting and error correcting codes. *AT&T Tech J* **29**, 147–160.
52 Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN & Bourne PE (2000) The Protein Data Bank. *Nucleic Acids Res* **28**, 235–242.
53 Bairoch A & Apweiler R (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res* **28**, 45–48.
54 Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W & Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389–3402.
55 Rost B, Sander C & Schneider R (1994) PHD – an automatic mail server for protein secondary structure prediction. *Comput Appl Biosci* **10**, 53–60.
56 McGuffin LJ, Bryson K & Jones DT (2000) The PSIPRED protein structure prediction server. *Bioinformatics* **16**, 404–405.

4- DISCUSSION et PERSPECTIVES

IV-1 - Récapitulatif

Cette thèse s'inscrit principalement dans le domaine de l'analyse de séquences, et requiert l'utilisation d'outils et de méthodes issues de la Bioinformatique. En particulier, elle consiste en l'analyse exhaustive de répétitions en tandem dans les séquences protéiques, par (i)le développement d'un outil adapté à l'exploration de grandes banques de données protéiques, et le stockage des résultats issus de son utilisation dans une base de données, (ii) l'utilisation de cette base afin d'analyser les répétitions d'un seul résidu appelées homorepeats, et enfin, (iii) la mise en évidence de la corrélation entre la perfection des répétitions et leur désordre intrinsèque.

La première étape de la thèse nous place dans une phase de collecte des données visant à regrouper de manière globale les répétitions en tandem existants dans les séquences protéiques. Elle s'est concrétisée par la création de deux outils : T-REKS, pour l'exploration des séquences et PRDB, pour le stockage et l'analyse. T-REKS a fait l'objet d'une publication scientifique (Jorda and Kajava, 2009), tandis que le manuscrit décrivant PRDB est actuellement en cours de soumission. Il a pu être montré que T-REKS détectait plus efficacement les répétitions en tandem que d'autres programmes existants. De même, PRDB représente une source de données majeure quant à son exhaustivité et l'utilisation d'outils d'analyse implémentés sur le serveur. Dans le chapitre 3, une étude a été menée sur les homorepeats, répétitions dont les propriétés physico-chimiques en font une classe marginale, afin de les caractériser de façon succincte à partir des publications précédentes et d'observations provenant de PRDB(Jorda and Kajava, 2010).Enfin, une seconde classe assez particulière regroupe les répétitions parfaites dont l'analyse a été l'objectif du travail présenté dans le chapitre 4(Jorda et al., 2010). En résumé, les travaux de cette thèse apportent une vision globale des répétitions en tandem, et les caractérisent afin de mieux comprendre leur rôle dans les séquences protéiques. Ces travaux ne constituant qu'une partie des possibilités d'analyses offertes par PRDB, les perspectives seront discutées dans cette partie.

IV-2 - T-REKS et PRDB : Deux nouveaux outils pour l'analyse exhaustive des répétitions en tandem dans les séquences protéiques

De nombreux programmes ont été développés en vue de faire face à la problématique qu'est l'identification des répétitions en tandem dans les séquences. Néanmoins, jusqu'à récemment, peu d'entre eux étaient axés sur l'exploration des séquences protéiques, et n'étaient pas adaptés à la recherche à grande échelle du fait de leur complexité. De cette nécessité est né T-REKS, programme d'identification qui s'appuie sur une méthode originale comparée à celles habituellement utilisées dans le domaine. En effet, les propriétés des séquences protéiques comme la courte longueur des séquences (350 résidus en moyenne), ou encore leur composition (20 monomères dans les séquences protéiques au lieu de 4 dans l'ADN) font que des heuristiques différentes de celles appliquées aux séquences nucléotidiques peuvent être envisagées. Le cœur de TREKS s'appuie sur le partitionnement (« clustering ») des distances calculées entre deux occurrences d'un même motif appelé « SS » (par défaut long de deux résidus) issu de la séquence. La phase de clustering, réalisée par l'algorithme des K-Means (MacQueen, 1967), a une complexité en $O((n-1).K)$ où n est le nombre d'occurrences pour un même SS dans une séquence de longueur l et K le nombre de clusters (défini à 10 dans T-REKS). Elle est appliquée sur chacun des SS, amenant la complexité de l'algorithme de clustering pour la séquence entière à $O(l.K)$. L'extraction des séquences hypothétiquement répétés en tandem se fait par un glissement de fenêtre sur la séquence des distances d'un même cluster, réordonnées en fonction des positions de leurs occurrences respectives. Des étapes successives de filtrage basées sur des estimations statistiques affinent les résultats, tout en gardant une complexité compatible avec des explorations à grande échelle dans les séquences protéiques.

Cette approche permet ainsi une identification *ab initio* des répétitions, condition pré-requise pour une exploration exhaustive des banques de données.

D'autres programmes dont l'article ne fait pas mention ont été publiés et n'ont pu être confrontés à des tests comparatifs de performance sur des séquences protéiques. C'est le cas de SWELFE (Abraham et al., 2008) et de TRStalker (Pellegrini et al., 2010), tous deux ayant été publiés après la création de T-REKS. Le premier s'appuie sur le même précepte que T-REKS : Si à partir d'une séquence donnée on calcule les distances entre des sous-séquences

identiques, il y a de fortes probabilités que les distances issues d'une répétition en tandem soient les plus fréquentes. Pellegrini *et al* propose une approche similaire de calcul de distances de sous séquences, et recourt à la méthode des «gapped q-grams » (Burkhardt and Karkkainen, 2002) afin de déterminer la longueur de la répétition. Toutefois, TRStalker a été développé spécifiquement pour l'exploration des séquences nucléotidiques, les divers essais d'exploration de séquences protéiques réalisés sur son interface web n'ont abouti à aucun résultat. SWELFE quant à lui est capable d'explorer les répétitions sur trois niveaux : les séquences nucléotidiques, protéiques et la structure tridimensionnelle. Il intègre un algorithme de programmation dynamique pour l'analyse de séquence, et analyse la similarité entre les différents angles diédraux formés par quatre Cα consécutifs. On peut supposer que l'utilisation de la programmation dynamique, détectant aspécifiquement les répétitions en tandem comme les répétitions dispersées, couplée à l'exploration des structures 3D dans SWELFE risque de poser des problèmes de performance lors de l'exploration à grande échelle des séquences. Un test préliminaire réalisé sur la séquence de l'involucrine humaine a pu confirmer cette hypothèse (Table 7).

	SWELFE	T-REKS
Meilleur match	MSQQHTLPVTLSPALSQELLKTVPPPVNTHQEQMKQPTPLPPPCQKVPVELPVEVPSKQE EKHMTAVKGLPEQECEQQQKEPQEQELQQQHWEQHEEYQKAENPEQQLKQEKTQRDQQL N KQLEEEKKLLDQQLDQELVKRDEQLGMKKEQLLELPEQQEGHLKHLEQQEGQLKHPEQ QE GQLELPEQQEGQLELPEQQEGQLELPEQQEGQLELPEQQEGQLELPEQQEGQLELPQ QQE GQLELSEQQEGQLELSEQQEGQLKHLEHQEGQLEVPEEQMGQLKYLEQQEGQLKHL DQQE KQPELPEQQMGQLKHLEQQEGQPKHLEQQEGQLEQLEEQEGQLKHLEQQEGQLEH LEHQE GQLGLPEQQVLQLKQLEKQQGQPKHLEEEEGQLKHLVQQEGQLKHLVQQEGQLE QQERQV EHLEQQVGQLKHLEEQEGQLKHLEQQQGQLEVPEQQVGQPKNLEQEEKQLELP EQQEGQV KHLEKQEAQLELPEQQVGQPKHLEQQEKHLEHPEQQDGQLKHLEQQEGQLKD LEQQKGQL EQPVFAPAPGQVQDIQPALPTKGEVLLPVEHQQQKQEVQWPPKHK	Length: 10 res - nb: 21 from 251 to 460 - Psim:0.732 region Length:210 GQLELSEQQE--- GQLKHLEHQE--- GQLEVPEEQM--- GQLKYLEQQE--- GQLKHLDQQE--- KQPELPEQQM--- GQLKHLEQQE--- GQPKHLEQQE--- GQLEQLEEQE--- GQLKHLEQQE--- GQLEHLEHQE--- GQLGLPEQQV--- LQLKQLEKQQ--- GQPKHLEEEE--- GQLKHLVQQE--- GQLKHLVQQE--- GQLEQQERQV--- ---EHLEQQVGQL ---KHLEEQEGQL ---KHLEQQQGQL ---EVPEQQVGQP
Temps d'exécution	3 min	2 secondes

Table 7. Comparatif de deux résultats issus de l'exploration par SWELFE et T-REKS de la séquence de l'involucrine humaine (Swiss-Prot Accessing number : P07476).

On préfèrera l'utilisation de SWELFE lors d'analyses ponctuelles d'une séquence donnée, car il offre une vue générale des répétitions sur trois niveaux, ce qu'actuellement la plupart des programmes, T-REKS inclus, ne sont en mesure de proposer.

T-REKS n'en reste pas moins perfectible et plusieurs améliorations sont d'ores et déjà envisageables. La plupart des paramètres numériques de T-REKS ont été fixés sur des valeurs définies via des tests statistiques. Or, il n'est pas concevable de déterminer optimalement tous les types de répétitions dans tous les types de séquences si ces paramètres ne prennent pas en compte les variabilités survenant dans les répétitions.

Nous avons pu montrer dans l'article du chapitre 1 que la probabilité de trouver une répétition parfaite par chance dans une séquence aléatoire suivait une loi binomiale (approximée par une loi de Poisson). Afin d'éviter les faux positifs, un seuil de longueur de répétition minimale a été fixé à 14 résidus, valeur pour laquelle le nombre attendu de répétitions dans une banque telle que Swiss-Prot est proche de zéro. On peut voir sur la figure 3 de l'article que ce même nombre varie selon la longueur du motif unitaire. Il est à noter qu'une exception a pu être faite pour les homorepeats, pour lesquels le seuil est de 9 résidus. Il est ainsi possible d'étendre cette modulation de seuil plutôt que de la fixer à une valeur, elle aura pour conséquence une détection précise des répétitions comprises entre 9 et 14 résidus, tout en minimisant les faux positifs.

Dans une perspective plus poussée, le modèle statistique énoncé plus haut nécessiterait d'être complété afin d'être plus proche de la réalité. En effet, celui-ci ayant été élaboré pour les régions parfaitement répétées, le calcul du seuil minimum ne prend pas en compte la variabilité due aux événements mutationnels. De même, nous considérons dans ce modèle que tous les acides aminés sont équiprobables (1 chance sur 20 qu'un acide aminé soit présent à une position dans la séquence), ce qui est loin d'être le cas lorsque l'on regarde la composition des séquences protéiques dans Swiss-Prot (Annexe II).

Enfin, dans l'éventualité où T-REKS pourrait un jour être appliqué aux séquences nucléotidiques (cf les tests préliminaires dans l'article), le modèle devra être adapté afin de faire face aux difficultés survenant avec ce type de séquence. La taille de l'alphabet, la longueur des séquences peuvent être des facteurs limitants à l'application de l'algorithme des K-Means et la définition du seuil minimum de longueur.

Grâce à T-REKS, nous avons pu construire une des bases des données les plus complètes en termes de répétitions en tandem dans les protéines. Les outils additionnels et la variété des données proposées à l'utilisateur ont fait évoluer PRDB d'une simple base de données vers une plate-forme centrale pour l'analyse des répétitions. L'interface de PRDB met à disposition un système de formulation de requêtes ainsi qu'un formatage des données adéquat pour des analyses ultérieures dans d'autres logiciels. Les outils statistiques, les générateurs de graphes, et les capacités de combinaison de requêtes de la QueryBox ont été abondamment utilisés lors des études dirigées dans les chapitres 3 et 4.

Une des applications de PRDB a consisté en la recherche des corrélations entre les différentes variables caractérisant les répétitions en tandem. Par exemple est-ce que la longueur d'une répétition varie en fonction de la longueur de la séquence protéique et est elle dépendante de sa localisation subcellulaire? Pour répondre à ce genre de question nous avons réalisé des travaux préliminaires basés sur la construction d'une matrice de corrélation entre les différentes variables. En premier lieu, elle a nécessité la récupération de tous les critères associés aux répétitions en tandem de Swiss-Prot, qu'ils soient quantitatifs ou qualitatifs. L'export des données au format CSV a permis de les réutiliser dans le programme de référence dédié à l'analyse statistique du nom de R (R Development Core Team, 2010). Il est à noter

qu'une étape intermédiaire visant à transformer les données qualitatives en des données quantitatives (cette méthode est appelée le codage disjonctif complet) a été effectuée grâce à PRDB. Grâce aux fonctions implémentées dans R, il a été possible de calculer la matrice des corrélations issue de toutes les variables, et de les visualiser sous forme d'un plot (figure 28).

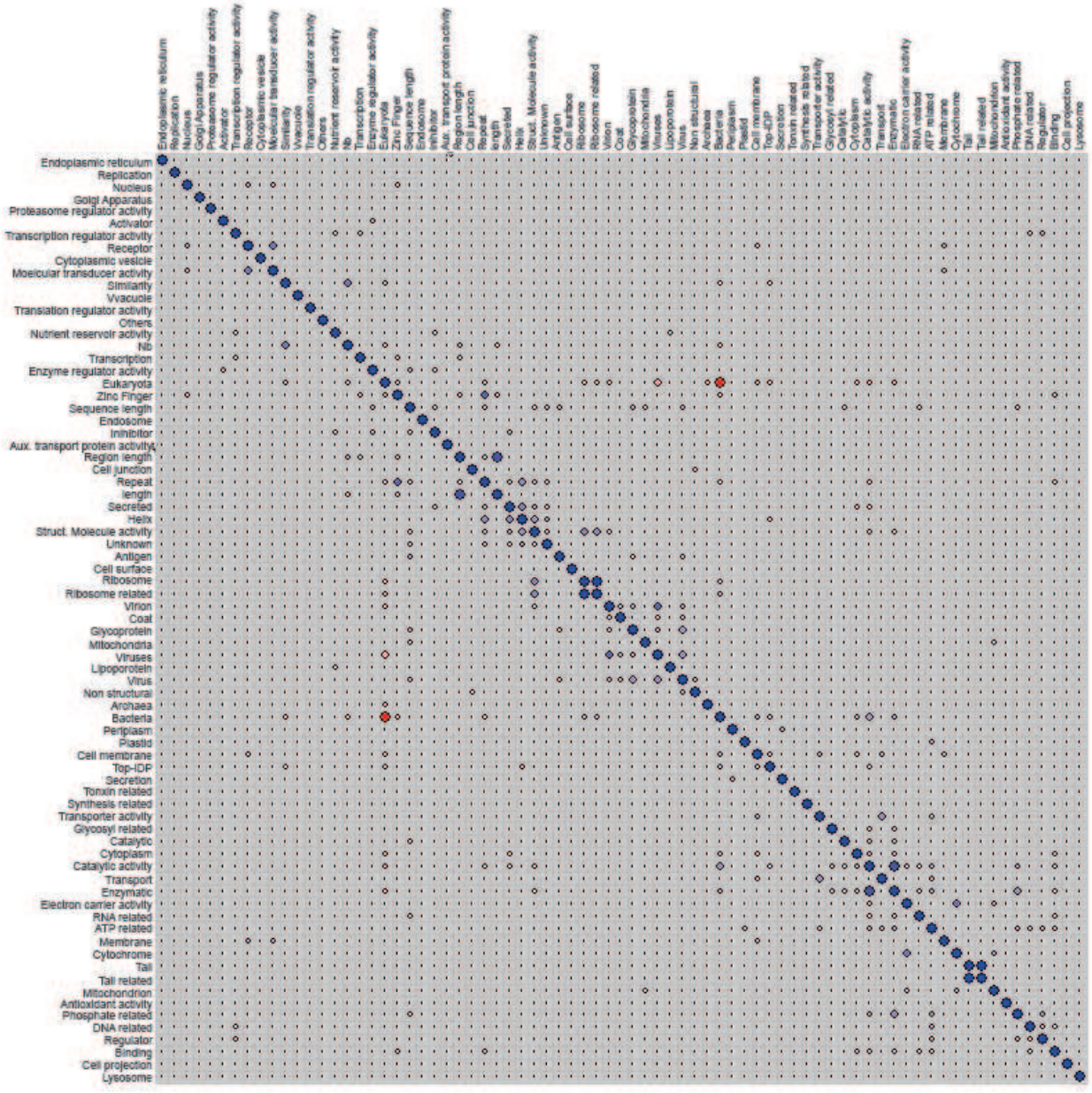


Figure 28. Interprétation graphique de la matrice de corrélation des variables issues de PRDB. Un cercle correspond à la corrélation entre deux variables, leur diamètre est proportionnel à l'intensité de la corrélation. Les corrélations positives sont représentées en bleu, les corrélations négatives (ou anti corrélations) en rouge.

Suite à une première observation, on peut noter que les corrélations les plus fortes sont assez évidentes : A l'exception des cercles sur la diagonale qui correspondent à la comparaison d'un critère avec lui-même, on retrouve une forte corrélation négative entre les bactéries et les eucaryotes (lorsqu'une répétition en tandem n'est pas d'origine bactérienne elle est très

souvent eucaryote ou inversement), ainsi qu'une forte corrélation positive de la longueur du motif avec la longueur totale de la région répétée. On retrouve également dans une moindre mesure la corrélation négative entre le coefficient *TOP-IDP* et le *Psim*, démontrée dans le chapitre 4. D'autres corrélations sont aussi présentes et peuvent faire l'objet de futures études (telles que la corrélation « bactérie - nombre de répétitions » ou encore « sécrété – hélice »). De plus, il est envisageable d'utiliser d'autres méthodes de mesure de corrélation entre deux variables: la matrice présentée étant basée sur le calcul du coefficient de Bravais-Pearson (Fisher, 1914), elle ne peut être employée que pour des données quantitatives. Néanmoins, il existe des tests statistiques capable de comparer des variables qualitatives comportant plusieurs modalités avec des variables quantitatives, tels que le test de Kruskal-Wallis (Kruskal and Wallis, 1952), qu'il serait intéressant d'appliquer à notre jeu de données.

Une autre application envisageable serait d'analyser les régions flanquantes des régions répétées dans les séquences protéiques, les données étant récupérables sur PRDB. La recherche des motifs consensus dans ces régions, ou de biais compositionnels permettrait de comprendre si certaines parties sont plus favorables que d'autres à l'accueil des répétitions en leur sein.

Dans le prolongement des perspectives de T-REKS, PRDB pourrait à terme incorporer les résultats issus d'exploration de séquences nucléotidiques, et mettre en juxtaposition les répétitions protéiques qui leur sont associées. Dans cette optique, il serait pertinent de confronter ces résultats avec les critères propres aux séquences génomiques : par exemple, évaluer s'il y a une corrélation entre les SNPs et les répétitions en tandem, tout deux étant des points chauds mutationnels.

IV-3 - Caractérisation des homorepeats

Les homorepeats ont été choisis comme sujet d'analyse du fait de leurs propriétés exceptionnelles conférées par la répétition d'un seul et même résidu. L'objectif de ces travaux a été de réaliser un état de l'art dans le domaine des homorepeats, en regroupant les informations issues de publications précédentes, mais aussi de le compléter avec de nouvelles données issues de notre analyse des homorepeats retrouvés dans PRDB. Dans une première phase, nous avons tenté de corroborer les hypothèses et observations précédentes (Huntley and Golding, 2000; Karlin et al., 2002; Faux et al., 2005) et de les organiser dans une revue cohérente et détaillée sur plusieurs niveaux : composition et distribution, structure, évolution et fonction. Ainsi, nous avons pu discuter de la longueur minimale des homorepeats (le plu souvent 5 à 7 résidus, inférieur à l'estimation de notre modèle statistique), de leur composition (ils sont majoritairement constitués de résidus polaires), de leur prédominance chez les eucaryotes et de la variabilité de leur fréquence entre les espèces dans un même règne. Il a pu être également mis en évidence que la composition en acides aminés variait selon les espèces (par exemple on retrouve les poly-Q majoritairement chez les vertébrés tandis que les poly-N, pourtant proches chimiquement, sont quasi absents). Aussi frappantes que soient certaines de ces préférences compositionnelles, nous n'avons pu expliquer leur existence.

Structurellement parlant, la revue souligne la capacité des homorepeats à constituer *in vitro* des cristallites α-hélicaux ou à structures β selon la nature de leurs acides aminés et des conditions expérimentales. Les homorepeats poly-P peuvent former une troisième famille de structures dites hélices polyproline (de type I et II), de même que les poly-G ont la capacité de s'organiser en trois types de cristallites (2 formes de type I et une forme de type II*). In vivo*, la prépondérance des résidus hydrophiles dans les homorepeats incite à dire qu'ils sont souvent désordonnés, confirmant les observations de Alba *et al* sur les homorepeats localisés dans les régions flanquantes de domaines structuraux définis (Mularoni et al., 2006). Néanmoins, les homorepeats peuvent s'autoagréger et former des structures fibrillaires stables, leur expansion pouvant mener à une cytotoxicité et être responsable de pathologies. Par ailleurs, certains homorepeats de résidus apolaires peuvent acquérir une structure qui n'aboutira pas à des agrégats, c'est le cas des répétitions poly-L localisées le plus souvent dans les protéines membranaires ou extracellulaires.

La corrélation entre les préférences compositionnelles, la localisation subcellulaire des protéines et leurs fonctions a été étudiée grâce à PRDB. Il en découle certains biais qui ont pu être mis à jour, comme l'absence de résidus polaires dans les protéines structurelles, l'abondance de répétitions poly-A dans les facteurs de transcription, le ratio élevé des résidus apolaires dans les répétitions de protéines membranaires ou extracellulaires, ou encore la grande proportion de prolines dans les homorepeats associés à des fonctions de liaison à l'actine.

Dans une troisième partie, la dynamique évolutive des homorepeats à l'origine de leur polymorphisme a été mise en exergue. Des études ont rapporté que les « gaps » (souvent dus aux phénomènes de contraction/expansion) dans les régions répétées étaient présents à un ratio 5 fois plus important que dans les régions adjacentes. Le taux élevé de ces variations et le fait que les répétitions sont souvent originaires du même codon (on parle d'homogénéité de codon ; par exemple les répétitions poly-Q sont principalement codées par le triplet CAG), suggèrent que le mécanisme d'expansion des homorepeats est principalement basé sur le glissement de l'ADN polymérase lors de la réplication. Parallèlement, il a pu être mis en évidence que les homorepeats encodés par des répétitions pures d'un même codon sont moins conservés entre les espèces que les répétitions dont les codons sont hétérogènes (Alba et al., 1999). Il a été souligné à plusieurs reprises qu'une expansion anormale pouvait être source d'agrégations dangereuses pour la cellule. Il est envisageable que des mécanismes régulateurs permettent de limiter ces effets et de contrôler la balance expansion/contraction (Karlin et al., 2002; Cleary and Pearson, 2003; Feschenko et al., 2003). L'étude des codons à la périphérie des répétitions, à la différence des codons les plus internes qui sont purs, a révélé qu'ils étaient très souvent soumis à des événements de transversion, limitant ainsi les bégaiements de la polymérase (Faux et al., 2007). Cependant, ceci a été spécifiquement mis en évidence pour les répétitions poly-Q, et mérite une étude approfondie des autres types de répétitions dont l'expansion est variable. Par exemple, les répétitions poly-P sont une des plus fréquentes chez les eucaryotes, tout en étant beaucoup moins sujettes à des variations que les poly-Q. L'analyse de leurs séquences nucléotidiques comparée à celle de leurs pendants protéiques permettrait de mettre à jour de nouvelles perspectives évolutives pour ce type de répétitions.

Finalement, la review du chapitre III a pu rassembler les hypothèses associées aux rôles de ces séquences dans les protéines, et les coupler avec nos observations et graphes en provenance de PRDB. De manière récurrente, les protéines portant des homorepeats ont des fonctions associées à la transcription/traduction, activité catalytique, régulation enzymatique, transport, adhésion cellulaire, signalisation ou des fonctions structurelles. Les eucaryotes, en particulier, présentent une grande diversité de protéines constitutives de grands complexes multiprotéiques souvent impliqués dans la transcription (comme les ribosomes) ou la signalisation (comme les MAP kinases) tandis que les procaryotes concentreraient une grande majorité de protéines associées à des activités enzymatiques. Une des nouveautés mise à disposition par PRDB est l'analyse des homorepeats identifiés chez les virus: les protéines virales porteuses d'homorepeats ont en grande partie un rôle structurel dans les capsides et enveloppes, suivi des facteurs de transcription.

Longtemps considérés comme des régions inutiles à la fonction de la protéine, ou au mieux comme des « linkers » entre des domaines fonctionnels, le rôle des homorepeats dans les protéines a pu être clarifié sur la base de précédents résultats : l'expansion et la contraction de ces régions permettrait de moduler subtilement la fonction des protéines sans en changer la nature, comme par exemple de réguler l'activation transcriptionelle via des variations de longueur des homorepeats. Il a pu être montré que ces modifications pouvaient aboutir à de légers changements phénotypiques, *a contrario* des modifications dans les domaines fonctionnels susceptibles de conduire à des changements drastiques (Fondon and Garner, 2004). Une de nos hypothèses serait que l'homogénéité de ces répétitions autorise une interaction modulable avec les acides nucléiques ou d'autres protéines. Par exemple, les homorepeats positivement chargés sont capables d'interagir avec l'ADN, chargé négativement. Il est probable que les répétitions poly-Q, poly-N, poly-A, poly-P, ou poly-G aient également cette capacité d'interaction via des liaisons hydrogènes. La longueur de la répétition serait intimement corrélée à la force de l'interaction, médiant ainsi une régulation fine de son activité.

Enfin, il serait inconcevable de dire que les homorepeats n'ont aucune fonction dans des protéines où elles jouent un rôle prépondérant, telles que la fibroïne de la soie ou le collagène.

Cette analyse, bien que exhaustive, ne prend en compte qu'une classe spécifique de répétitions en tandem. Il serait envisageable de conduire le même type de travail sur les

répétitions en tandem au sens large, et de comparer les conclusions formulées avec celles des homorepeats.

IV-4 - Analyse de la perfection des répétitions et leur désordre intrinsèque

Le chapitre 4 apporte une nouvelle pierre à l'édifice concernant le désordre intrinsèque des répétitions en tandem. En effet, elles ont été identifiées comme des régions souvent désordonnées (Tompa, 2003), de même que l'évolution des séquences non structurées reposeraient sur des événements de contraction/expansion. Nous avons pu illustrer ces affirmations en implémentant dans PRDB le calcul de l'échelle TOP-IDP(Campen et al., 2008) sur les régions répétées. Il en a été déduit que plus de 73% des répétitions en tandem dans PRDB étaient prédites instructurées. L'étude conjointe des répétitions détectées dans PDB, de leurs structures sur leurs fichiers PDB correspondants, et des prédictions du désordre dans les répétitions issues de Swiss-Prot a permis de préciser et de compléter ces résultats. Ainsi, il a été constaté dans une première phase d'analyse des fichiers PDB que les répétitions les plus parfaitement répétées étaient les plus susceptibles d'être non-structurées. *A contrario*, les plus dégénérées pouvaient présenter des éléments de structure secondaire. Nous avons formulé l'hypothèse que le désordre dans les répétitions en tandem était dépendant de leur perfection.

Dans une seconde phase, nous avons comparé la perfection de la répétition, mesurée par le *Psim*, avec le degré de désordre intrinsèque calculé par plusieurs programmes de prédiction de désordre dans les séquences protéiques. Les résultats ont tous validé notre hypothèse, montrant que le désordre des régions répétées en tandem augmentait avec la perfection des répétitions. De même, une étude des séquences protéiques les contenant a pu indiquer que les répétitions les plus parfaites étaient majoritairement retrouvées dans les séquences hautement désordonnées, et quasi absentes dans les séquences structurées. Enfin, nous avons pu constater que le désordre des répétitions était sujet à des variations selon la localisation cellulaire de la protéine, principalement chez les eucaryotes où 80% des répétitions localisées dans le cytoplasme sont instructurées, pour 60% dans les membranes.

Comme pour les homorepeats, ces préférences de localisation n'ont pu trouver d'explication. Une des hypothèses probables serait que les biais compositionnels des séquences désordonnées, souvent porteuses de résidus chargés (négativement ou positivement), pourraient influer sur le pHi de la protéine et favoriser/interdire une localisation cellulaire

(Schwartz et al., 2001). En général, le pHi (appelé également point isoélectrique) des protéines cellulaires suit une distribution bimodale (Weiller et al., 2004) : les protéines cytoplasmiques sont naturellement enrichies en résidus chargés (acides pour la plupart), avec un pHi aux environs de 5, tandis que les protéines membranaires présentent plutôt des résidus basiques et apolaires, portant le pHi autour de 9 (Weiller et al., 2004). Chez les eucaryotes, les protéines nucléaires engloberaient un mélange des deux types de pHi. L'existence ou l'absence de certains types de répétitions en tandem dans des localisations subcellulaires pourrait s'expliquer par la nécessité de conserver le pHi d'une protéine à une certaine valeur. Les séquences répétées, du fait leur forte dynamique évolutive, seraient susceptibles de faire varier fortement le pHi d'une protéine vers des valeurs qui ne sont pas en accord avec le compartiment cellulaire qui la contient.

Enfin, la perfection des répétitions trouverait une de ses explications dans des raisons évolutives plutôt qu'une conservation de ces séquences basée sur leur importance structurelle et fonctionnelle. En effet, nul besoin de préserver une séquence à un tel degré de perfection si son seul but est de coder pour du désordre intrinsèque, d'autres résidus peuvent jouer ce rôle et substituer ceux d'origine. Dans un même ordre d'idée, les évènements mutationnels dans les régions structurées ont une plus grande incidence sur la structure et/ou fonction de la protéine que dans celles désordonnées. Or, nous avons pu montrer dans le chapitre 4 que les régions les moins parfaitement répétées étaient les plus susceptibles de présenter une structure. On peut ainsi considérer que les répétitions ayant une structure associée seraient des répétitions plus ancestrales et ont pu dégénérer sous certaines contraintes au cours de l'évolution, gardant un certain degré de perfection crucial à leurs structures et/ou fonctions. D'un autre côté, les répétitions parfaites seraient quant à elles des évènements évolutifs récents, n'ayant pas encore subi de mutations. Leur instabilité et rapidité d'évolution combinée à leur absence de contraintes structurelles en font des candidats idéaux pour une réponse rapide à des changements environnementaux.

Cette hypothèse mériterait toutefois d'être étayée par une analyse phylogénétique de ces séquences, afin de suivre la perfection des répétitions au cours de l'évolution.

5- CONCLUSIONS

L'analyse exhaustive des répétitions en tandem dans les séquences protéiques est un challenge tout aussi complexe que dans les séquences génomiques. Les moyens mis en œuvre pour atteindre les objectifs de cette thèse ont apporté de nouveaux concepts dans le domaine.

Ainsi, l'identification exacte des répétitions dans les séquences reste une problématique d'actualité que nous avons tenté de résoudre par le développement d'un outil, fondé sur des méthodes originales, nommé T-REKS. Malgré l'arrivée de nouveaux programmes, T-REKS reste à l'heure actuelle un outil de choix dans la détection à grande échelle des répétitions en tandem protéiques.

Suite à l'exploitation des banques de données protéiques, il nous a été possible de constituer la plus grande ressource en termes de répétitions en tandem protéiques, appelée PRDB. Cette base de données a été étoffée par plusieurs outils de visualisation et de manipulation des données, convertissant PRDB en une véritable plate-forme dédiée à l'étude des séquences répétées en tandem dans les protéines.

Deux études ont pu être déjà menées grâce à PRDB et mettre à jour de nouveaux aspects propres aux répétitions en tandem.

Un grand nombre de fonctions associées aux protéines porteuses de répétitions ont pu être mises en évidence via l'étude des homorepeats. Le fait que ces régions soient dans la plupart des cas prédites comme désordonnées, ne peut suffire à les réduire à de simples régions de liaisons entre des domaines structurés et fonctionnels. A cet égard, nous avons pu proposer plusieurs pistes concernant leurs rôles dans les protéines.

Parallèlement, l'étude des répétitions parfaites a pu prolonger les observations formulées sur le désordre de ces régions. Nos résultats indiquent clairement que le désordre intrinsèque est intimement lié à la perfection des répétitions. Ils constituent de nouveaux éléments à prendre en considération dans la prédiction de structure des séquences protéiques, ou dans la compréhension de l'évolution de ces régions.

Nous n'avons pu réaliser d'analyse systématique aboutissant à une compréhension globale des répétitions en tandem dans les protéines, mais certaines perspectives proposées, comme l'étude des corrélations des variables caractérisant ces répétitions, ont vocation à le faire. De même, une étude fonctionnelle poussée de toutes les répétitions en tandem pourrait mettre en évidence de nouvelles fonctions inhérentes.

6- BIBLIOGRAPHIE

Abraham, A. L., E. P. Rocha and J. Pothier (2008). Swelfe: a detector of internal repeats in sequences and structures. *Bioinformatics*, **24,** 1536-7.

Alba, M. M. and R. Guigo (2004). Comparative analysis of amino acid repeats in rodents and humans. *Genome Res*, **14,** 549-54.

Alba, M. M., M. F. Santibanez-Koref and J. M. Hancock (1999). Conservation of polyglutamine tract size between mice and humans depends on codon interruption. *Mol Biol Evol*, **16,** 1641-4.

Alba, M. M., P. Tompa and R. A. Veitia (2007). Amino acid repeats and the structure and evolution of proteins. *Genome Dyn*, **3,** 119-30.

Albrecht, A. N., U. Kornak, A. Boddrich, K. Suring, P. N. Robinson, A. C. Stiege, R. Lurz, S. Stricker, E. E. Wanker and S. Mundlos (2004). A molecular pathogenesis for transcription factor associated poly-alanine tract expansions. *Hum Mol Genet*, **13,** 2351-9.

Altman, G. H., F. Diaz, C. Jakuba, T. Calabro, R. L. Horan, J. Chen, H. Lu, J. Richmond and D. L. Kaplan (2003). Silk-based biomaterials. *Biomaterials*, **24,** 401-16.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman (1990). Basic local alignment search tool. *J Mol Biol*, **215,** 403-10.

Anan, K., N. Yoshida, Y. Kataoka, M. Sato, H. Ichise, M. Nasu and S. Ueda (2007). Morphological change caused by loss of the taxon-specific polyalanine tract in Hoxd-13. *Mol Biol Evol*, **24,** 281-7.

Andrade, M. A., C. P. Ponting, T. J. Gibson and P. Bork (2000). Homology-based method for identification of protein repeats using statistical significance estimates. *J Mol Biol*, **298,** 521-37.

Armitage, P. (1955). Tests for Linear Trends in Proportions and Frequencies. *Biometrics*, **11,** 375-386.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin and G. Sherlock (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, **25,** 25-9.

Auer, H. E. and P. Doty (1966). The conformational stability of alpha-helical nonpolar polypeptides in solution. *Biochemistry*, **5,** 1716-25.

Bairoch, A. and R. Apweiler (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res*, **28,** 45-8.

Balbirnie, M., R. Grothe and D. S. Eisenberg (2001). An amyloid-forming peptide from the yeast prion Sup35 reveals a dehydrated beta-sheet structure for amyloid. *Proc Natl Acad Sci U S A*, **98,** 2375-80.

Bauer, P. O. and N. Nukina (2009). The pathogenic mechanisms of polyglutamine diseases and current therapeutic strategies. *J Neurochem*, **110,** 1737-65.

Baumann, U., S. Wu, K. M. Flaherty and D. B. McKay (1993). Three-dimensional structure of the alkaline protease of Pseudomonas aeruginosa: a two-domain protein with a calcium binding parallel beta roll motif. *Embo J*, **12,** 3357-64.

Baxa, U., T. Cassese, A. V. Kajava and A. C. Steven (2006). Structure, function, and amyloidogenesis of fungal prions: filament polymorphism and prion variants. *Adv Protein Chem*, **73,** 125-80.

Belay, E. D. (1999). Transmissible spongiform encephalopathies in humans. *Annu Rev Microbiol*, **53,** 283-314.

Bennett, S. T., A. M. Lucassen, S. C. Gough, E. E. Powell, D. E. Undlien, L. E. Pritchard, M. E. Merriman, Y. Kawaguchi, M. J. Dronsfield, F. Pociot and et al. (1995). Susceptibility to human type 1 diabetes at IDDM2 is determined by tandem repeat variation at the insulin gene minisatellite locus. *Nat Genet*, **9,** 284-92.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res*, **27,** 573-80.

Benson, G. and L. Dong (1999). Reconstructing the duplication history of a tandem repeat. *Proc Int Conf Intell Syst Mol Biol*, 44-53.

Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne (2000). The Protein Data Bank. *Nucleic Acids Res*, **28,** 235-42.

Blout, E. R. (1962). Polyamino Acids, Polypeptides and Proteins. *(edit. by Stahmann, M. A.) University of Wisconsin Press, Madison*, 275.

Boby, T., A. M. Patch and S. J. Aves (2005). TRbase: a database relating tandem repeats to disease genes for the human genome. *Bioinformatics*, **21,** 811-6.

Brown, L. Y. and S. A. Brown (2004). Alanine tracts: the expanding story of human illness and trinucleotide repeats. *Trends Genet*, **20,** 51-8.

Buchanan, G., M. Yang, A. Cheong, J. M. Harris, R. A. Irvine, P. F. Lambert, N. L. Moore, M. Raynor, P. J. Neufing, G. A. Coetzee and W. D. Tilley (2004). Structural and functional consequences of glutamine tract variation in the androgen receptor. *Hum Mol Genet*, **13,** 1677-92.

Bucher, P. and A. Bairoch (1994). A generalized profile syntax for biomolecular sequence motifs and its function in automatic sequence interpretation. *Proc Int Conf Intell Syst Mol Biol*, **2,** 53-61.

Bucher, P., K. Karplus, N. Moeri and K. Hofmann (1996). A flexible motif search technique based on generalized profiles. *Comput Chem*, **20,** 3-23.

Burkhardt, S. and J. Karkkainen (2002). One-gapped q-gram filters for Levenshtein distance. *Combinatorial Pattern Matching*, **2373,** 225-234.

Buscaglia, C. A., J. Alfonso, O. Campetella and A. C. Frasch (1999). Tandem amino acid repeats from Trypanosoma cruzi shed antigens increase the half-life of proteins in blood. *Blood*, **93,** 2025-32.

Campen, A., R. M. Williams, C. J. Brown, J. Meng, V. N. Uversky and A. K. Dunker (2008). TOP-IDP-scale: a new amino acid scale measuring propensity for intrinsic disorder. *Protein Pept Lett*, **15,** 956-63.

Cheung, J., X. Estivill, R. Khaja, J. R. MacDonald, K. Lau, L. C. Tsui and S. W. Scherer (2003). Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. *Genome Biol*, **4,** R25.

Cleary, J. D. and C. E. Pearson (2003). The contribution of cis-elements to disease-associated repeat instability: clinical and experimental evidence. *Cytogenet Genome Res*, **100,** 25-55.

Cochran, W. G. (1954). Some Methods for Strengthening the Common X2 Tests. *Biometrics*, **10,** 417-451.

Crapeau, M., C. Marchal, C. Cullin and L. Maillet (2009). The cellular concentration of the yeast Ure2p prion protein affects its propagation as a prion. *Mol Biol Cell*, **20,** 2286-96.

Csink, A. K. and S. Henikoff (1998). Something from nothing: the evolution and utility of satellite repeats. *Trends Genet*, **14,** 200-4.

Eddy, S. R. (2009). A new generation of homology search tools based on probabilistic inference. *Genome Inform*, **23,** 205-11.

Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5,** 113.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, **32,** 1792-7.

El-Agnaf, O. M. and G. B. Irvine (2000). Review: formation and properties of amyloid-like fibrils derived from alpha-synuclein and related proteins. *J Struct Biol*, **130,** 300-9.

Ellegren, H. (2000). Heterogeneous mutation processes in human microsatellite DNA sequences. *Nat Genet*, **24,** 400-2.

Ellegren, H. (2000). Microsatellite mutations in the germline: implications for evolutionary inference. *Trends Genet*, **16,** 551-8.

Falquet, L., M. Pagni, P. Bucher, N. Hulo, C. J. Sigrist, K. Hofmann and A. Bairoch (2002). The PROSITE database, its status in 2002. *Nucleic Acids Res*, **30,** 235-8.

Faux, N. G., S. P. Bottomley, A. M. Lesk, J. A. Irving, J. R. Morrison, M. G. de la Banda and J. C. Whisstock (2005). Functional insights from the distribution and role of homopeptide repeat-containing proteins. *Genome Res*, **15,** 537-51.

Faux, N. G., G. A. Huttley, K. Mahmood, G. I. Webb, M. G. de la Banda and J. C. Whisstock (2007). RCPdb: An evolutionary classification and codon usage database for repeat-containing proteins. *Genome Res*, **17,** 1118-27.

Feschenko, V. V., L. A. Rajman and S. T. Lovett (2003). Stabilization of perfect and imperfect tandem repeats by single-strand DNA exonucleases. *Proc Natl Acad Sci U S A*, **100,** 1134-9.

Finn, R. D., J. Mistry, B. Schuster-Bockler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, R. Durbin, S. R. Eddy, E. L. Sonnhammer and A. Bateman (2006). Pfam: clans, web tools and services. *Nucleic Acids Res*, **34,** D247-51.

Finn, R. D., J. Mistry, J. Tate, P. Coggill, A. Heger, J. E. Pollington, O. L. Gavin, P. Gunasekaran, G. Ceric, K. Forslund, L. Holm, E. L. Sonnhammer, S. R. Eddy and A. Bateman The Pfam protein families database. *Nucleic Acids Res*, **38,** D211-22.

Fischetti, V. A., G. M. Landau, P. H. Sellers and J. P. Schmidt (1993). Identifying periodic occurrences of a template with applications

to protein structure. *Inf. Proc. Lett.*, **45,** 11–18.

Fisher, R. A. (1914). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, **10,** 507-521.

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

Fondon, J. W., 3rd and H. R. Garner (2004). Molecular origins of rapid and continuous morphological evolution. *Proc Natl Acad Sci U S A*, **101,** 18058-63.

Fontenot, J. D., N. Tjandra, D. Bu, C. Ho, R. C. Montelaro and O. J. Finn (1993). Biophysical characterization of one-, two-, and three-tandem repeats of human mucin (muc-1) protein core. *Cancer Res*, **53,** 5386-94.

Fraser, R. D. B. and T. P. MacRae (1973). Conformation in fibrous proteins and related synthetic polypeptides. London and New York, Academic Press.

Fulop, V. and D. T. Jones (1999). Beta propellers: structural rigidity and functional diversity. *Curr Opin Struct Biol*, **9,** 715-21.

Gatchel, J. R. and H. Y. Zoghbi (2005). Diseases of unstable repeat expansion: mechanisms and common principles. *Nat Rev Genet*, **6,** 743-55.

Gelfand, Y., A. Rodriguez and G. Benson (2007). TRDB--the Tandem Repeats Database. *Nucleic Acids Res*, **35,** D80-7.

Giguere, Y., E. Dewailly, J. Brisson, P. Ayotte, N. Laflamme, A. Demers, V. I. Forest, S. Dodin, J. Robert and F. Rousseau (2001). Short polyglutamine tracts in the androgen receptor are protective against breast cancer in the general population. *Cancer Res*, **61,** 5869-74.

Given, H. F., R. Radbourne, H. Oag, S. Merritt, E. Barclay, A. M. Hanby, H. Lamlum, J. McGrath, C. Curran and I. P. Tomlinson (2000). The androgen receptor exon 1 trinucleotide repeat does not act as a modifier of the age of presentation in breast cancer. *Eur J Cancer*, **36,** 533-4.

Glover, J. N. and S. C. Harrison (1995). Crystal structure of the heterodimeric bZIP transcription factor c-Fos-c-Jun bound to DNA. *Nature*, **373,** 257-61.

Goldmann, W. (2008). PrP genetics in ruminant transmissible spongiform encephalopathies. *Vet Res*, **39,** 30.

Goto, Y., D. Carter and S. G. Reed (2008). Immunological dominance of Trypanosoma cruzi tandem repeat proteins. *Infect Immun*, **76,** 3967-74.

Graur, D. and W.-H. Li (2000). Fundamentals of Molecular Evolution. 2nd edn.

Gregory, T. R. (2005). Synergy between sequence and size in large-scale genomics. *Nat Rev Genet*, **6,** 699-708.

Gribskov, M., A. D. McLachlan and D. Eisenberg (1987). Profile analysis: detection of distantly related proteins. *Proc Natl Acad Sci U S A*, **84,** 4355-8.

Groves, M. R. and D. Barford (1999). Topological characteristics of helical repeat proteins. *Curr Opin Struct Biol*, **9,** 383-9.

Gusfield, D. (1993). Efficient methods for multiple sequence alignment with guaranteed error bounds. *Bull Math Biol*, **55,** 141-54.

Gusfield, D. and J. Stoye (2004). Linear time algorithms for finding and representing all the tandem repeats in a string. *Journal of Computer and System Sciences*, **69,** 525-546.

Haber, J. E. and E. J. Louis (1998). Minisatellite origins in yeast and humans. *Genomics*, **48,** 132-5.
Hamming, R. (1950). Error Detecting and Error Correcting Codes. *Bell System Technical Journal*, **29,** 147-160.
Hao, Y., R. Montiel, B. Li, E. Huang, L. Zeng and Y. Huang (2010). Association between androgen receptor gene CAG repeat polymorphism and breast cancer risk: a meta-analysis. *Breast Cancer Res Treat*,
Heger, A. and L. Holm (2000). Rapid automatic detection and alignment of repeats in protein sequences. *Proteins*, **41,** 224-37.
Henikoff, S., K. Ahmad and H. S. Malik (2001). The centromere paradox: stable inheritance with rapidly evolving DNA. *Science*, **293,** 1098-102.
Henikoff, S. and J. G. Henikoff (1992). Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, **89,** 10915-9.
Henikoff, S. and J. G. Henikoff (2000). Amino acid substitution matrices. *Adv Protein Chem*, **54,** 73-97.
Heringa, J. (1998). Detection of internal repeats: how common are they? *Curr Opin Struct Biol*, **8,** 338-45.
Horikawa, Y., N. Oda, N. J. Cox, X. Li, M. Orho-Melander, M. Hara, Y. Hinokio, T. H. Lindner, H. Mashima, P. E. Schwarz, L. del Bosque-Plata, Y. Horikawa, Y. Oda, I. Yoshiuchi, S. Colilla, K. S. Polonsky, S. Wei, P. Concannon, N. Iwasaki, J. Schulze, L. J. Baier, C. Bogardus, L. Groop, E. Boerwinkle, C. L. Hanis and G. I. Bell (2000). Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. *Nat Genet*, **26,** 163-75.
Huang, Q. Y., F. H. Xu, H. Shen, H. Y. Deng, Y. J. Liu, Y. Z. Liu, J. L. Li, R. R. Recker and H. W. Deng (2002). Mutation patterns at dinucleotide microsatellite loci in humans. *Am J Hum Genet*, **70,** 625-34.
Huntley, M. and G. B. Golding (2000). Evolution of simple sequence in proteins. *J Mol Evol*, **51,** 131-40.
Huntley, M. A. and G. B. Golding (2002). Simple sequences are rare in the Protein Data Bank. *Proteins*, **48,** 134-40.
Janke, C., K. Rogowski, D. Wloga, C. Regnard, A. V. Kajava, J. M. Strub, N. Temurak, J. van Dijk, D. Boucher, A. van Dorsselaer, S. Suryavanshi, J. Gaertig and B. Edde (2005). Tubulin polyglutamylase enzymes are members of the TTL domain protein family. *Science*, **308,** 1758-62.
Jeffreys, A. J., V. Wilson, R. Neumann and J. Keyte (1988). Amplification of human minisatellites by the polymerase chain reaction: towards DNA fingerprinting of single cells. *Nucleic Acids Res*, **16,** 10953-71.
Jin, P. and S. T. Warren (2000). Understanding the molecular basis of fragile X syndrome. *Hum Mol Genet*, **9,** 901-8.
Jorda, J. and A. V. Kajava (2009). T-REKS: identification of Tandem REpeats in sequences with a K-meanS based algorithm. *Bioinformatics*, **25,** 2632-8.
Jorda, J. and A. V. Kajava (2010). Protein homorepeats: Sequences, Structures, Evolution, and Functions. *Advances in Protein Chemistry and Structural Biology*, **79,** 59-88.
Jorda, J., B. Xue, V. N. Uversky and A. V. Kajava (2010). Protein tandem repeats - the more perfect, the less structured. *Febs J*, **277,** 2673-82.
Kajava, A. V. (2001). Review: proteins with repeated sequence--structural prediction and modeling. *J Struct Biol*, **134,** 132-44.
Kajava, A. V., U. Baxa and A. C. Steven (2010). Beta arcades: recurring motifs in naturally occurring and disease-related amyloid fibrils. *Faseb J*, **24,** 1311-9.
Kajava, A. V., U. Baxa, R. B. Wickner and A. C. Steven (2004). A model for Ure2p prion filaments and other amyloids: the parallel superpleated beta-structure. *Proc Natl Acad Sci U S A*, **101,** 7885-90.
Kalita, M. K., G. Ramasamy, S. Duraisamy, V. S. Chauhan and D. Gupta (2006). ProtRepeatsDB: a database of amino acid repeats in genomes. *BMC Bioinformatics*, **7,** 336.

Karlin, S., L. Brocchieri, A. Bergman, J. Mrazek and A. J. Gentles (2002). Amino acid runs in eukaryotic proteomes and disease associations. *Proc Natl Acad Sci U S A*, **99,** 333-8.

Karlin, S. and C. Burge (1996). Trinucleotide repeats and long homopeptides in genes and proteins associated with nervous system disease and development. *Proc Natl Acad Sci U S A*, **93,** 1560-5.

Katoh, K., K. Kuma, H. Toh and T. Miyata (2005). MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res*, **33,** 511-8.

Katti, M. V., P. K. Ranjekar and V. S. Gupta (2001). Differential distribution of simple sequence repeats in eukaryotic genome sequences. *Mol Biol Evol*, **18,** 1161-7.

Knight, S. J., M. C. Hirst and K. E. Davies (1992). Molecular analysis of the fragile X syndrome. *Dis Markers*, **10,** 1-5.

Kobe, B. and J. Deisenhofer (1993). Crystal structure of porcine ribonuclease inhibitor, a protein with leucine-rich repeats. *Nature*, **366,** 751-6.

Kobe, B. and A. V. Kajava (2000). When protein folding is simplified to protein coiling: the continuum of solenoid protein structures. *Trends Biochem Sci*, **25,** 509-15.

Kobe, B. and A. V. Kajava (2001). The leucine-rich repeat as a protein recognition motif. *Curr Opin Struct Biol*, **11,** 725-32.

Kolpakov, R., G. Bana and G. Kucherov (2003). mreps: Efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Res*, **31,** 3672-8.

Kolpakov, R. and G. Kucherov (2003). Finding approximate repetitions under Hamming distance. *Theoretical Computer Science*, **303,** 135-156.

Krogh, A., M. Brown, I. S. Mian, K. Sjolander and D. Haussler (1994). Hidden Markov models in computational biology. Applications to protein modeling. *J Mol Biol*, **235,** 1501-31.

Kruskal, W. H. and W. A. Wallis (1952). Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association*, **47,** 583-621.

Kushnirov, V. V., I. M. Alexandrov, O. V. Mitkevich, I. S. Shkundina and M. D. Ter-Avanesyan (2006). Purification and analysis of prion and amyloid aggregates. *Methods*, **39,** 50-5.

Landau, G. M., J. P. Schmidt and D. Sokol (2001). An algorithm for approximate tandem repeats. *J Comput Biol*, **8,** 1-18.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J.

Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blocker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowski, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi and Y. J. Chen (2001). Initial sequencing and analysis of the human genome. *Nature*, **409,** 860-921.

Le Fleche, P., Y. Hauck, L. Onteniente, A. Prieur, F. Denoeud, V. Ramisse, P. Sylvestre, G. Benson, F. Ramisse and G. Vergnaud (2001). A tandem repeats database for bacterial genomes: application to the genotyping of Yersinia pestis and Bacillus anthracis. *BMC Microbiol*, **1,** 2.

Lee, K. A. (2007). Ewings family oncoproteins: drunk, disorderly and in search of partners. *Cell Res*, **17,** 286-8.

Lee, M. S., G. P. Gippert, K. V. Soman, D. A. Case and P. E. Wright (1989). Three-dimensional solution structure of a single zinc finger DNA-binding domain. *Science*, **245,** 635-7.

Levinson, G. and G. A. Gutman (1987). Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol Biol Evol*, **4,** 203-21.

Li, D. and R. Roberts (2001). WD-repeat proteins: structure characteristics, biological function, and their involvement in human diseases. *Cell Mol Life Sci*, **58,** 2085-97.

Li, W., P. Zhang, J. P. Fellers, B. Friebe and B. S. Gill (2004). Sequence composition, organization, and evolution of the core Triticeae genome. *Plant J*, **40,** 500-11.

Lipman, D. J. and W. R. Pearson (1985). Rapid and sensitive protein similarity searches. *Science*, **227,** 1435-41.

Lodish, H., A. Berk, C. Kaiser, M. Krieger, M. Scott, A. Bretscher, H. Ploegh and P. Matsudaira (2000). Molecular Cell Biology (Lodish, Molecular Cell Biology), W. H. Freeman.

MacQueen (1967). Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press.

Malaise, M., H. M. Schatzl and A. Burkle (2008). The octarepeat region of prion protein, but not the TM1 domain, is important for the antioxidant effect of prion protein. *Free Radic Biol Med*, **45,** 1622-30.

Malik, H. S. and S. Henikoff (2001). Adaptive evolution of Cid, a centromere-specific histone in Drosophila. *Genetics*, **157,** 1293-8.

Marcotte, E. M., M. Pellegrini, T. O. Yeates and D. Eisenberg (1999). A census of protein repeats. *J Mol Biol*, **293,** 151-60.

Marguerat, S., B. T. Wilhelm and J. Bahler (2008). Next-generation sequencing: applications beyond genomes. *Biochem Soc Trans*, **36,** 1091-6.

McCutchan, T. F., A. A. Lal, V. do Rosario and A. P. Waters (1992). Two types of sequence polymorphism in the circumsporozoite gene of Plasmodium falciparum. *Mol Biochem Parasitol*, **50,** 37-45.

McDevitt, D. and T. J. Foster (1995). Variation in the size of the repeat region of the fibrinogen receptor (clumping factor) of Staphylococcus aureus strains. *Microbiology*, **141 (Pt 4),** 937-43.

Messier, W., S. H. Li and C. B. Stewart (1996). The birth of microsatellites. *Nature*, **381,** 483.

Metzgar, D., J. Bytof and C. Wills (2000). Selection against frameshift mutations limits microsatellite expansion in coding DNA. *Genome Res*, **10,** 72-80.

Morgante, M., M. Hanafey and W. Powell (2002). Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nat Genet*, **30,** 194-200.

Mularoni, L., R. Guigo and M. M. Alba (2006). Mutation patterns of amino acid tandem repeats in the human proteome. *Genome Biol*, **7,** R33.

Needleman, S. B. and C. D. Wunsch (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, **48,** 443-53.

Neer, E. J., C. J. Schmidt, R. Nambudripad and T. F. Smith (1994). The ancient regulatory-protein family of WD-repeat proteins. *Nature*, **371,** 297-300.

Nelson, K. A. and J. S. Witte (2002). Androgen receptor CAG repeats and prostate cancer. *Am J Epidemiol*, **155,** 883-90.

Ng, K. P., G. Potikyan, R. O. Savene, C. T. Denny, V. N. Uversky and K. A. Lee (2007). Multiple aromatic side chains within a disordered structure are critical for transcription and transforming activity of EWS family oncoproteins. *Proc Natl Acad Sci U S A*, **104,** 479-84.

Ninomiya, H., K. Nomura, Y. Satoh, S. Okumura, K. Nakagawa, M. Fujiwara, E. Tsuchiya and Y. Ishikawa (2006). Genetic instability in lung cancer: concurrent analysis of chromosomal, mini- and microsatellite instability and loss of heterozygosity. *Br J Cancer*, **94,** 1485-91.

Notredame, C. and D. G. Higgins (1996). SAGA: sequence alignment by genetic algorithm. *Nucleic Acids Res*, **24,** 1515-24.

Notredame, C., D. G. Higgins and J. Heringa (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*, **302,** 205-17.

Nowacki, M., B. P. Higgins, G. M. Maquilan, E. C. Swart, T. G. Doak and L. F. Landweber (2009). A functional role for transposases in a large eukaryotic genome. *Science*, **324,** 935-8.

Oesch, B., D. Westaway, M. Walchli, M. P. McKinley, S. B. Kent, R. Aebersold, R. A. Barry, P. Tempst, D. B. Teplow, L. E. Hood and et al. (1985). A cellular gene encodes scrapie PrP 27-30 protein. *Cell*, **40,** 735-46.

Ohno, S. (1972). So much "junk" DNA in our genome. *Brookhaven Symp Biol*, **23,** 366-70.

Orgel, L. E., F. H. Crick and C. Sapienza (1980). Selfish DNA. *Nature*, **288,** 645-6.

Orr, H. T. and H. Y. Zoghbi (2007). Trinucleotide repeat disorders. *Annu Rev Neurosci*, **30,** 575-621.

Osoegawa, K., A. G. Mammoser, C. Wu, E. Frengen, C. Zeng, J. J. Catanese and P. J. de Jong (2001). A bacterial artificial chromosome library for sequencing the complete human genome. *Genome Res*, **11,** 483-96.

Pan, K. M., M. Baldwin, J. Nguyen, M. Gasset, A. Serban, D. Groth, I. Mehlhorn, Z. Huang, R. J. Fletterick, F. E. Cohen and et al. (1993). Conversion of alpha-helices into beta-sheets features in the formation of the scrapie prion proteins. *Proc Natl Acad Sci U S A*, **90,** 10962-6.

Parry, D. A., R. D. Fraser and J. M. Squire (2008). Fifty years of coiled-coils and alpha-helical bundles: a close relationship between sequence and structure. *J Struct Biol*, **163,** 258-69.

Pellegrini, M., M. E. Renda and A. Vecchio (2010). TRStalker: an efficient heuristic for finding fuzzy tandem repeats. *Bioinformatics*, **26,** i358-66.

Perutz, M. F., T. Johnson, M. Suzuki and J. T. Finch (1994). Glutamine repeats as polar zippers: their possible role in inherited neurodegenerative diseases. *Proc Natl Acad Sci U S A*, **91,** 5355-8.

Pritchard, J. K. and M. Przeworski (2001). Linkage disequilibrium in humans: models and data. *Am J Hum Genet*, **69,** 1-14.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly and P. C. Sham (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, **81,** 559-75.

R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria, R Foundation for Statistical Computing.

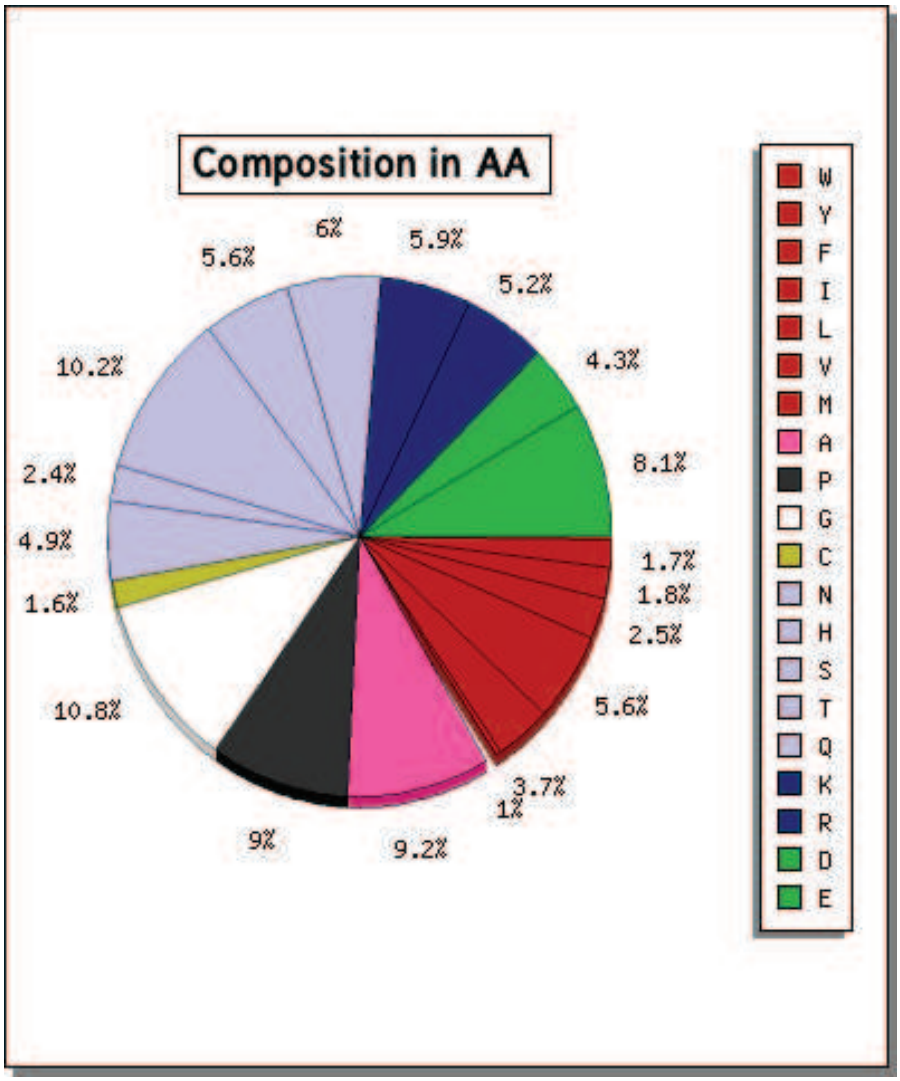
Raetz, C. R. and S. L. Roderick (1995). A left-handed parallel beta helix in the structure of UDP-N-acetylglucosamine acyltransferase. *Science*, **270,** 997-1000.

Rebbeck, T. R., P. W. Kantoff, K. Krithivas, S. Neuhausen, M. A. Blackwood, A. K. Godwin, M. B. Daly, S. A. Narod, J. E. Garber, H. T. Lynch, B. L. Weber and M. Brown (1999). Modification of BRCA1-associated breast cancer risk by the polymorphic androgen-receptor CAG repeat. *Am J Hum Genet*, **64,** 1371-7.

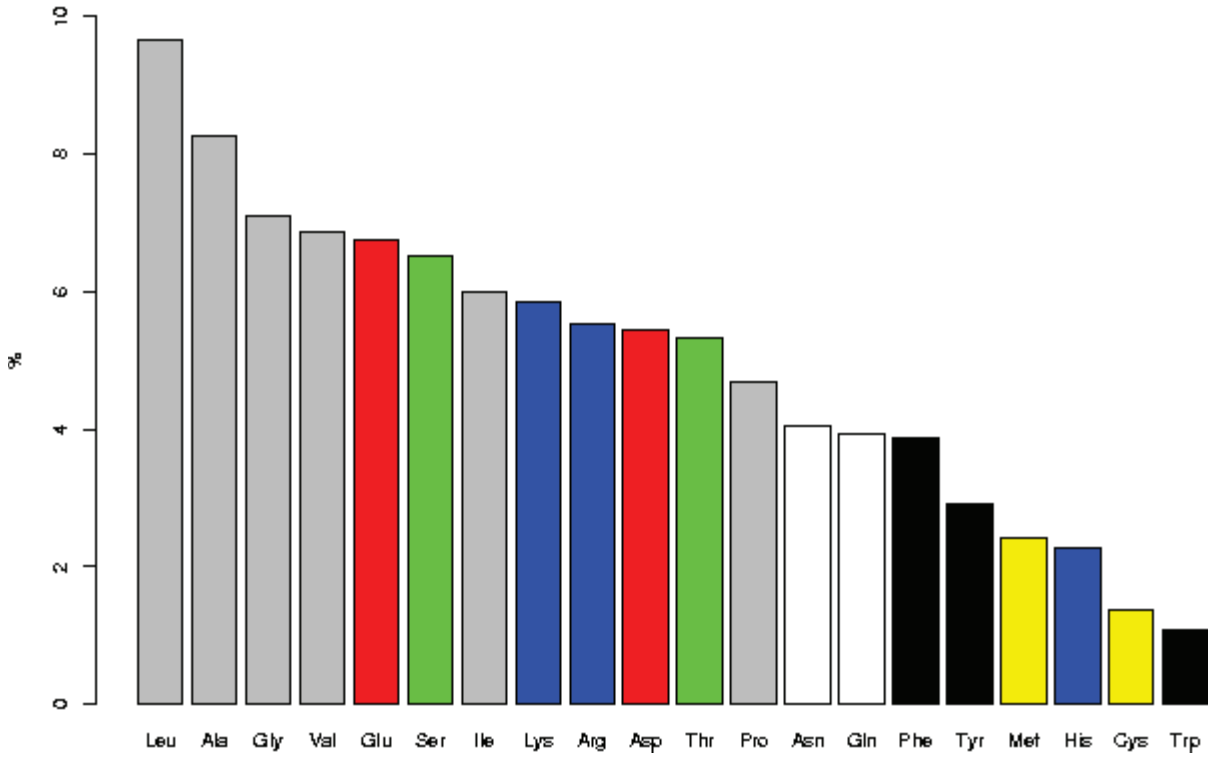
Rich, S. M. and F. J. Ayala (2000). Population structure and recent evolution of Plasmodium falciparum. *Proc Natl Acad Sci U S A*, **97,** 6994-7001.
Richard, G. F. and F. Paques (2000). Mini- and microsatellite expansions: the recombination connection. *EMBO Rep*, **1,** 122-6.
Ronaghi, M., M. Uhlen and P. Nyren (1998). A sequencing method based on real-time pyrophosphate. *Science*, **281,** 363, 365.
Saitou, N. and M. Nei (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, **4,** 406-25.
Sanger, F., G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, C. A. Fiddes, C. A. Hutchison, P. M. Slocombe and M. Smith (1977). Nucleotide sequence of bacteriophage phi X174 DNA. *Nature*, **265,** 687-95.
Schuster, S. C. (2008). Next-generation sequencing transforms today's biology. *Nat Methods*, **5,** 16-8.
Schwartz, R., C. S. Ting and J. King (2001). Whole proteome pI values correlate with subcellular localizations of proteins for organisms within the three domains of life. *Genome Res*, **11,** 703-9.
Selin, I. (1967). Viterbi,Aj - Principles of Coherent Communication. *Ieee Transactions on Information Theory*, **It13,** 527-+.
Shimohata, T., O. Onodera and S. Tsuji (2000). Interaction of expanded polyglutamine stretches with nuclear transcription factors leads to aberrant transcriptional regulation in polyglutamine diseases. *Neuropathology*, **20,** 326-33.
Shkundina, I. S., V. V. Kushnirov, M. F. Tuite and M. D. Ter-Avanesyan (2006). The role of the N-terminal oligopeptide repeats of the yeast Sup35 prion protein in propagation and transmission of prion variants. *Genetics*, **172,** 827-35.
Shopsin, B., M. Gomez, S. O. Montgomery, D. H. Smith, M. Waddington, D. E. Dodge, D. A. Bost, M. Riehman, S. Naidich and B. N. Kreiswirth (1999). Evaluation of protein A gene polymorphic region DNA sequencing for typing of Staphylococcus aureus strains. *J Clin Microbiol*, **37,** 3556-63.
Simon, M. and J. M. Hancock (2009). Tandem and cryptic amino acid repeats accumulate in disordered regions of proteins. *Genome Biol*, **10,** R59.
Singh, G. P., B. R. Chandra, A. Bhattacharya, R. R. Akhouri, S. K. Singh and A. Sharma (2004). Hyper-expansion of asparagines correlates with an abundance of proteins with prion-like domains in Plasmodium falciparum. *Mol Biochem Parasitol*, **137,** 307-19.
Sladek, R., G. Rocheleau, J. Rung, C. Dina, L. Shen, D. Serre, P. Boutin, D. Vincent, A. Belisle, S. Hadjadj, B. Balkau, B. Heude, G. Charpentier, T. J. Hudson, A. Montpetit, A. V. Pshezhetsky, M. Prentki, B. I. Posner, D. J. Balding, D. Meyre, C. Polychronakos and P. Froguel (2007). A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, **445,** 881-5.
Smith, T. F. and M. S. Waterman (1981). Identification of common molecular subsequences. *J Mol Biol*, **147,** 195-7.
Snewin, V. A., M. Herrera, G. Sanchez, A. Scherf, G. Langsley and S. Herrera (1991). Polymorphism of the alleles of the merozoite surface antigens MSA1 and MSA2 in Plasmodium falciparum wild isolates from Colombia. *Mol Biochem Parasitol*, **49,** 265-75.
Sokol, D., G. Benson and J. Tojeira (2007). Tandem repeats over the edit distance. *Bioinformatics*, **23,** e30-5.
Sonnhammer, E. L., S. R. Eddy, E. Birney, A. Bateman and R. Durbin (1998). Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res*, **26,** 320-2.
Subramanian, A. R., J. Weyer-Menkhoff, M. Kaufmann and B. Morgenstern (2005). DIALIGN-T: an improved algorithm for segment-based multiple sequence alignment. *BMC Bioinformatics*, **6,** 66.
Szklarczyk, R. and J. Heringa (2004). Tracking repeats using significance and transitivity. *Bioinformatics*, **20 Suppl 1,** i311-7.

Thakur, A. K. and R. Wetzel (2002). Mutational analysis of the structural organization of polyglutamine aggregates. *Proc Natl Acad Sci U S A*, **99,** 17014-9.
Thompson, J. D., D. G. Higgins and T. J. Gibson (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, **22,** 4673-80.
Tompa, P. (2003). Intrinsically unstructured proteins evolve by repeat expansion. *Bioessays*, **25,** 847-55.
Tormo, J., J. Puiggali, J. Vives, I. Fita, J. Lloveras, J. Bella, J. Aymami and J. A. Subirana (1992). Crystal structure of a helical oligopeptide model of polyglycine II and of other polyamides: acetyl-(glycyl-beta-alanyl)2-NH propyl. *Biopolymers*, **32,** 643-8.
van Belkum, A. (1999). Short sequence repeats in microbial pathogenesis and evolution. *Cell Mol Life Sci*, **56,** 729-34.
van der Woude, M. W. and A. J. Baumler (2004). Phase and antigenic variation in bacteria. *Clin Microbiol Rev*, **17,** 581-611, table of contents.
Venter, J. C., M. D. Adams, G. G. Sutton, A. R. Kerlavage, H. O. Smith and M. Hunkapiller (1998). Shotgun sequencing of the human genome. *Science*, **280,** 1540-2.
Vergnaud, G. (1989). Polymers of random short oligonucleotides detect polymorphic loci in the human genome. *Nucleic Acids Res*, **17,** 7623-30.
Wain, L. V., J. A. Armour and M. D. Tobin (2009). Genomic copy number variation, human health, and disease. *Lancet*, **374,** 340-50.
Wang, V., T. C. Chuang, B. W. Soong, D. E. Shan and M. C. Kao (2009). Octarepeat changes of prion protein in Parkinson's disease. *Parkinsonism Relat Disord*, **15,** 53-8.
Wang, Z., M. Gerstein and M. Snyder (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, **10,** 57-63.
Waterston, R. H., E. S. Lander and J. E. Sulston (2002). On the sequencing of the human genome. *Proc Natl Acad Sci U S A*, **99,** 3712-6.
Weber, J. L. and C. Wong (1993). Mutation of human short tandem repeats. *Hum Mol Genet*, **2,** 1123-8.
Weiller, G. F., G. Caraux and N. Sylvester (2004). The modal distribution of protein isoelectric points reflects amino acid properties rather than sequence evolution. *Proteomics*, **4,** 943-9.
Wheeler, D. A., M. Srinivasan, M. Egholm, Y. Shen, L. Chen, A. McGuire, W. He, Y. J. Chen, V. Makhijani, G. T. Roth, X. Gomes, K. Tartaro, F. Niazi, C. L. Turcotte, G. P. Irzyk, J. R. Lupski, C. Chinault, X. Z. Song, Y. Liu, Y. Yuan, L. Nazareth, X. Qin, D. M. Muzny, M. Margulies, G. M. Weinstock, R. A. Gibbs and J. M. Rothberg (2008). The complete genome of an individual by massively parallel DNA sequencing. *Nature*, **452,** 872-6.
Xu, X., M. Peng and Z. Fang (2000). The direction of microsatellite mutations is dependent upon allele length. *Nat Genet*, **24,** 396-9.
Yam, A. Y., C. M. Gao, X. Wang, P. Wu and D. Peretz The octarepeat region of the prion protein is conformationally altered in PrP(Sc). *PLoS One*, **5,** e9316.
Ybe, J. A., F. M. Brodsky, K. Hofmann, K. Lin, S. H. Liu, L. Chen, T. N. Earnest, R. J. Fletterick and P. K. Hwang (1999). Clathrin self-assembly is mediated by a tandemly repeated superhelix. *Nature*, **399,** 371-5.
Yoder, M. D., S. E. Lietzke and F. Jurnak (1993). Unusual structural features in the parallel beta-helix in pectate lyases. *Structure*, **1,** 241-51.
Yu, H., B. Bharaj, E. J. Vassilikos, M. Giai and E. P. Diamandis (2000). Shorter CAG repeat length in the androgen receptor gene is associated with more aggressive forms of breast cancer. *Breast Cancer Res Treat*, **59,** 153-61.

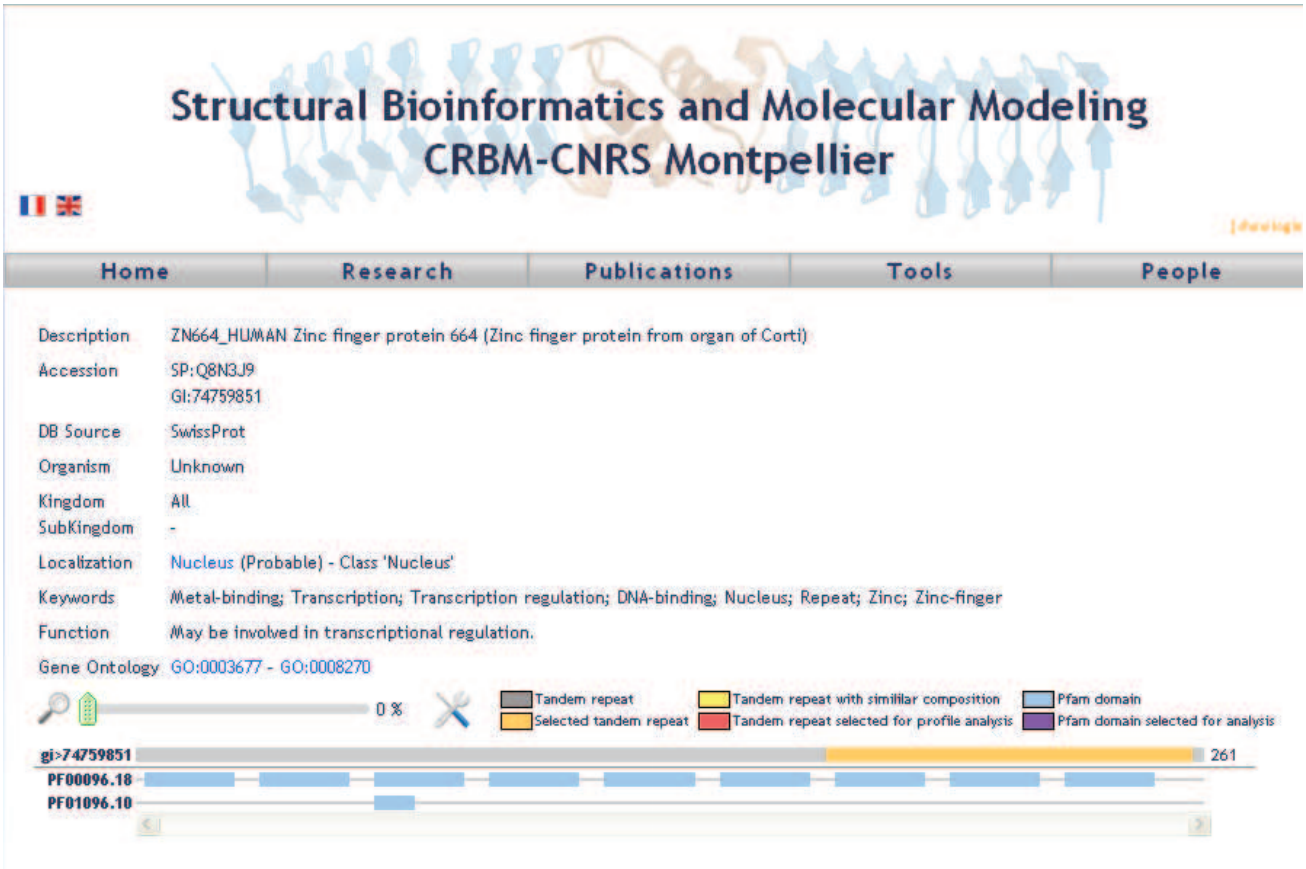
7- ANNEXES



Annexe I. Composition globale en acides aminés des répétitions en tandem de Swiss-Prot.



Annexe II. Distribution des acides aminés dans les séquences protéiques issues de Swiss-Prot (Août 2010).



Annexe III. Interface de Sequence Viewer.