

Chapitre 1

Introduction

En 1995, j'effectuais un stage de Biologie Moléculaire dans l'unité du Dr. Jean-Yves Courtois. Avec Jean-Yves Brossas et le Dr. Jacques Tréton, nous tentions alors de caractériser une recircularisation de l'ADN mitochondrial. Pour ce faire, nous séquencions alors quelques centaines de bases "à la main avec du phosphore radioactif". Actuellement, plus d'une trentaine de génomes complets sont disponibles. Les progrès technologiques et méthodologiques nous ont donné une masse d'informations extraordinaire qu'il faut désormais traiter.

Avec cette explosion récente des bases de données aussi bien génomiques que protéiques, se pose de manière cruciale le problème de la *génomique structurale*, du repliement des protéines. Les protéines sont formées par la succession d'acides aminés liés entre eux par des liaisons peptidiques. Cette succession est directement liée à la séquence génomique et contrôle la formation de la structure tridimensionnelle de la protéine. Cette structure contrôle l'ensemble des fonctions biologiques des protéines. La connaissance et la possibilité de prédire leurs structures sont donc fondamentale.

Possédant la séquence d'une protéine, nous devrions pouvoir en déduire directement sa structure tridimensionnelle. Les facteurs physico-chimiques et cinétiques qui agissent lors du repliement d'une protéine sont toutefois encore trop complexes et les approximations trop grandes pour pouvoir simuler un repliement *in silico* à l'identique de celui *in vivo* [119].

Aussi, faut-il obligatoirement connaître une protéine possédant une certaine homologie avec celle qui nous intéresse pour pouvoir travailler. La structure tridimensionnelle des protéines peut être caractérisée en ne tenant compte que de ses parties régulières et répétitives, les feuillets et les hélices. Elles décrivent une géométrie spatiale stabilisée par des liaisons internes. Le reste, plus variable, est dénommé boucles. Les méthodes de prédiction des ces trois états ont

connu ces dernières années grâce à l'utilisation de séquences proches et de réseaux neuronaux une augmentation importante de leur taux de prédiction permettant d'atteindre actuellement environ 75 % de prédiction correcte [166, 168, 26].

Un alphabet à 3 états demeure cependant assez pauvre structurellement. Différentes équipes ont ainsi décidé d'approfondir la spécificité structurale, en analysant les connections qui existent entre deux structures régulières consécutives le long des protéines. Des bibliothèques de boucles plus ou moins longues ont donc été proposées [203, 114, 13, 204]. La composition très spécifiques de certaines boucles ont permis des prédictions intéressantes [189]. Toutefois, ces différentes méthodes sont liées à la définition des structures secondaires.

Pour décrire plus précisément la structure des protéines, différentes équipes ont élaboré des alphabets structuraux comportant un nombre différent de prototypes. Unger et collaborateurs [198] et Schuchhardt et coll. [175] en proposent une centaine, ceux-ci permettant une bonne précision dans l'approximation de la structure, difficilement utilisable néanmoins dans une approche de prédiction. Rooman et coll. [162], Fetrow et coll. [53], Bystroff et Baker [19], et Camproux et coll. [23, 24] en proposent un nombre plus restreint allant de 4 à 13 prototypes. Tous montrent les spécificités en acides aminés selon le prototype observé. Bystroff et Baker [19] propose en plus une méthode de prédiction.

Dans le cadre de cette thèse, nous avons donc décidé de développer une nouvelle méthodologie pour concevoir des séries de blocs protéiques dans un but tant de prédiction locale que d'approximation des structures protéiques. Ayant obtenu différentes séries de prototypes, la série qui sera choisie devra permettre une prédiction avec un taux acceptable et aussi permettre une reproduction correcte de la structure protéique. Il conviendra alors de vérifier et d'analyser la spécificité des blocs protéiques au plan structural, voir leur stabilité et les comparer aux différents alphabets existants.

Notre second objectif est d'utiliser cet "alphabet structural" (ensemble des blocs protéiques) dans une prédiction locale du squelette polypeptidique. Cette prédiction sera effectuée avec une méthode qui permet de comprendre l'importance des acides aminés de manière simple. Pour améliorer cette prédiction, nous nous sommes basés sur deux concepts: (i) *1 repliement local* $\rightarrow$ *n séquences* et (ii) *1 séquence* $\rightarrow$ *n repliements*. Le premier concept signifie que plusieurs types de séquences peuvent être associés à la même structure et le second qu'une séquence peut être associée à plusieurs type de repliements [186]. Ces deux aspects seront développés en se

basant sur la recherche d'un indice de fiabilité lié à la prédiction locale, pour trouver des zones de fortes probabilités.

Ayant défini un alphabet structural, certains mots, i.e. successions de blocs protéiques peuvent apparaître plus fréquemment que d'autres ainsi que l'avaient remarqué Fetrow et collaborateurs [53]. Nous essayerons alors de définir au mieux quelle est l'architecture de ces successions, les liens existants entre ces différents mots.

Les étapes précédentes avaient pour but de caractériser un alphabet structural permettant de décrire au mieux la structure 3D des protéines et les dépendances entre blocs protéiques. Du fait de cette redondance qui peut apparaître dans la structure protéique, il nous a semblé intéressant de mettre au point une méthode de compactage qui permet d'associer des structures proches sur le plan tridimensionnel local. Cette approche appelée "protéine hybride" de conception simple permet de catégoriser en classes "structuralement dépendantes". Cette approche, en plus du compactage, peut être utilisée dans une optique différente, celle de la recherche d'homologie structurale et de la caractérisation des dépendances entre structures et séquences.

Le travail réalisé tourne ainsi autour de la définition d'un alphabet structural, et, de ses potentiels dans le cadre de la prédiction de structure à un niveau local, de la recherche d'homologie structurale et de la caractérisation des dépendances entre structures et séquences. Les différents chapitres détailleront ces différents aspects.