

Chapitre 7

Conclusion et perspectives

a. Conclusion générale Lors de cette thèse, j'ai mis au point une nouvelle méthodologie pour concevoir des petits prototypes protéiques, les **blocs protéiques** (BPs). Après avoir conçu plusieurs séries ayant des nombre de BPs variables, nous avons conservé un **alphabet structural** composé de 16 BPs. Il permet à la fois une approximation correcte de la structure tridimensionnelle des protéines et une prédiction acceptable. La qualité structurale des blocs a été vérifiée, ils sont à la fois bien distincts les uns des autres et permettent une assignation non ambiguë. Les BPs sont, d'un point de vue structural, hautement spécifiques.

Le choix de 16 BPs est adéquat pour analyser la structure tridimensionnelle des protéines. Deux BPs sont spécifiques des parties régulières des protéines, les autres blocs permettent ainsi de bien analyser les extrémités N- et C-terminales de ces structures. Quatre PBs sont même totalement localisés dans les parties "boucles". La correspondance avec les structures secondaires classiques sont bien retrouvées, mais les PBs permettent une analyse beaucoup plus poussée de l'ensemble des structures protéiques.

La **prédiction de la structure 3D** en termes de blocs protéiques à partir de la séquence **par une approche bayésienne** atteint un taux de 30,0 % avec une fenêtre initiale de 5 résidus. L'agrandissement de cette fenêtre à 15 résidus permet un gain de 4,4 %. La mise au point d'une méthode de séparation des séquences (**1 repliement local** $\rightarrow$ **n séquences**) en classes, dites **familles séquentielles**, pour les blocs les plus fréquents permet un taux final de prédiction de 40,7 %.

La spécificité séquentielle importante des blocs a montré, en plus, de ce taux correct de prédiction, que la majorité des blocs réels se trouvait parmi les plus probables. Ce fait a permis, en plus de la méthode bayésienne de prédiction, de mettre en œuvre deux stratégies basées sur

cette information (**1 séquence** $\rightarrow$ **n repliements locaux**): (i) une **stratégie globale** qui permet de déterminer le nombre de blocs protéiques à conserver en chaque site pour atteindre un taux de prédiction donné, (ii) une **stratégie locale** qui permet de délimiter un certain nombre de zones ayant un taux de prédiction donné en prenant un nombre fixe de blocs protéiques.

La recherche des motifs de 5 BPs les plus fréquents dans la base de données a permis de construire un **réseau** qui les connectent. Plus de 83 % des sites protéiques sont contenus dans ce graphe et l'analyse de leur répartition en acides aminés montrent que ce réseau permet de voir des différences significatives selon la position du BP dans le réseau. La vérification de la stabilité structurale du réseau a donné des résultats particulièrement bons avec des *RMSd* faibles (2 Å) pour des fragments ayant des longueurs allant de 8 à 11 C $_{\alpha}$.

Allant plus loin dans l'utilisation de l'alphabet structural, nous avons élaboré la **méthode de la protéine hybride** (MPH). Elle permet la réduction d'une base de données structurales protéiques à une centaine de fragments prototypes d'une longueur de 10 PBs chacun. La grande majorité de ces fragments est bien approximée. Leur répartition en acides aminés montre, en outre, une spécificité importante qui est, de plus, liée dans les 3 cas sur 4 des cas à une information structurale bien déterminée. MPH appliquée aux blocs protéiques permet, de plus, une **recherche d'homologie structurale** rapide et performante.

Enfin, la méthode MPH appliquée à des informations séquences et structures recodées en paramètres physico-chimiques et angulaires a montré son intérêt dans l'analyse des relations entre les deux types d'information. Les blocs protéiques montrent ici aussi leur pertinence dans le domaine de l'analyse en permettant une description bien plus complexe des repliements qu'une simple description en structures secondaires.

b. Perspectives La figure 7.1 montre la logique des recherches effectuées et futures, en traits bleu foncé les travaux déjà réalisés, et, en bleu clair différentes voies possibles pour l'établissement d'un modèle protéique. Différentes évolutions sont possibles:

- o Un premier point concerne l'**homologie structurale**, quand les structures tridimensionnelles sont connues. Déjà, la méthode MPH a démontré ses possibilités. Il conviendrait, en plus, de l'information structurale, de prendre en compte la séquence qui, quoique moins conservée dans l'évolution, possède une information d'intérêt.

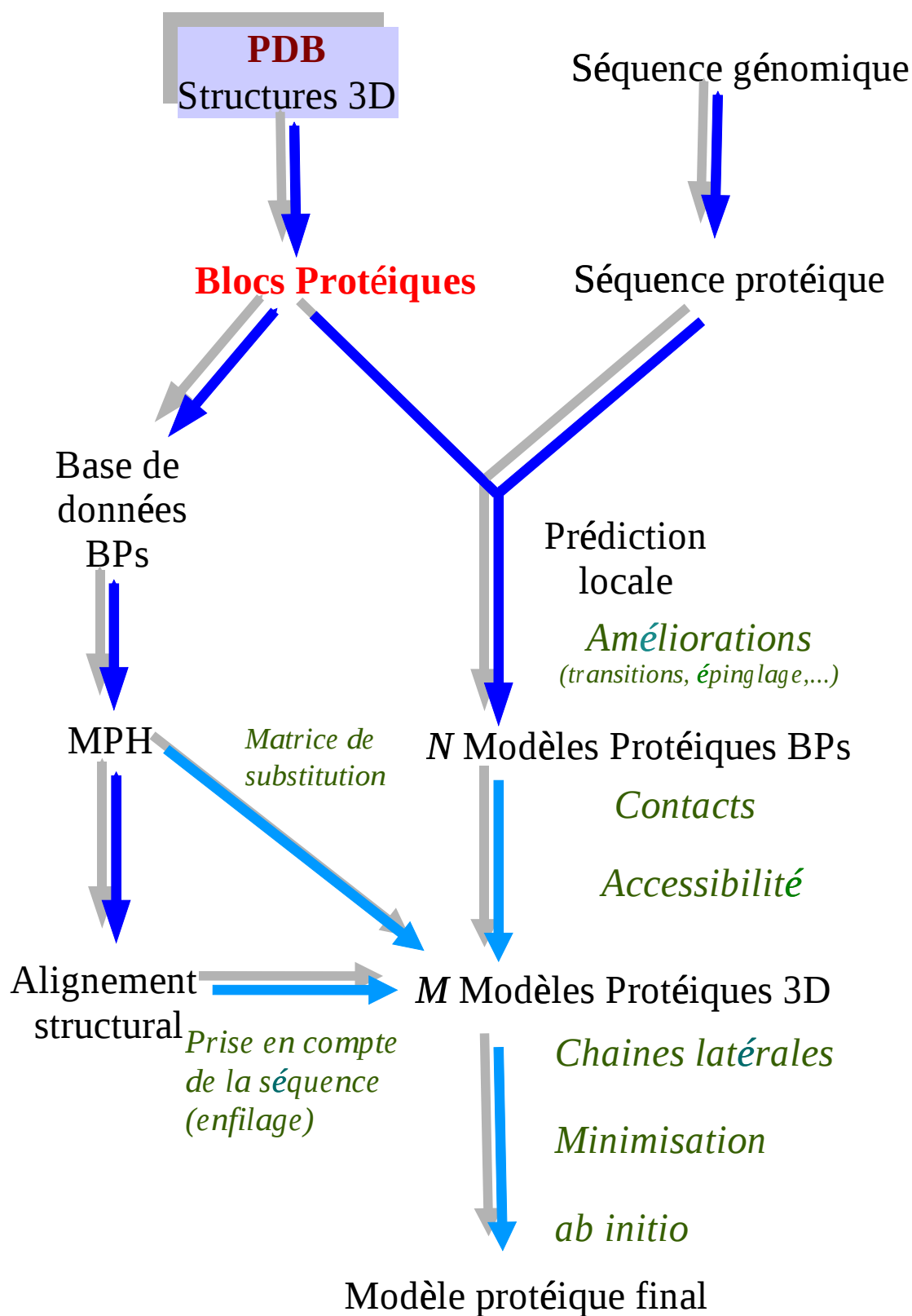


FIG. 7.1 – Schéma récapitulatif des avancées dues aux BPs et leur intégration dans un processus de modélisation moléculaire classique.

- o Dans la même optique, l'utilisation des blocs protéiques pourrait servir, dans une étape de prétraitement, dans une technique d'**enfilage**. En effet, observer les compatibilités entre les résultats de la prédiction locale et des structures connues issues d'une base non redondante permet de diminuer le nombre de repliements à tester.
- o Un travail intéressant, touchant toujours à l'homologie structurale, est la **classification protéique**. Ayant décrit un alphabet structural, la protéine hybride a bien montré la catégorisation de fragments protéiques en un nombre limité de classes. L'utilisation de ces observations devraient être pertinente pour reclasser les structures tridimensionnelles. Une analyse sur les familles obtenues serait bien entendu utile à la compréhension du repliement protéique.
- o Un autre point concerne la **génomique structurale**. Partant d'une séquence codante, nous désirons obtenir un modèle tridimensionnel final le plus proche de la réalité. Pour cela, nous avons développé un alphabet structural et une méthode de **prédiction locale**. Cette dernière peut être améliorée par des approches classiques qui prennent en compte les alignements de séquences ou les transitions existants entre les BPs. En effet, que ce soit par le réseau sur les mots ou l'approche de la protéine hybride avec les blocs protéiques, on a observé qu'une série de BPs est suivie par un nombre limité d'autres séries bien déterminées. Actuellement, une méthode de prédiction basée sur ce principe est en fin de réalisation et montre un intérêt certain.

L'ensemble de ces résultats permet la génération de **différents modèles 3D** qu'il faut travailler. L'usage de la librairie de recodage d'angles est ainsi un premier pas, les angles dièdres étant nettement mieux approximés. De nombreux autres informations devraient être introduits pour aboutir à ce modèle, comme l'accessibilité des résidus, les contacts les plus probables entre résidus, et, le positionnement des chaînes latérales. L'utilisation comme contrainte dans un processus *ab initio* d'une manière similaire aux travaux du groupe de Baker est aussi envisageable [178]. En effet, les contraintes à longues distances ne sont pas prises en compte actuellement.

L'ensemble des travaux effectués donnent des résultats concluants. A partir d'une séquence, il est possible de délimiter des zones hautement probables et ces zones sont décrits par un alphabet structural qui permet une approximation correcte des repliements. Toutefois, des améliorations restent à faire avant d'aboutir à un modèle final.