



U.F.R. Sciences de la Vie  
INSERM UMR-S 665, Dynamique des  
Structures et Interactions des  
Macromolécules Biologiques (DSIMB)  
INTS, 6 rue Alexandre Cabanel, 75739, Paris



MEDIT SA,  
2 rue du Belvédère  
91120, Palaiseau



Thèse de Doctorat de l'Université Paris 7 – Denis Diderot

Ecole Doctorale B3MI – Biochimie, Biothérapies, Biologie Moléculaire et  
Infectiologie

Spécialité

**Analyse du Génome et Modélisation Moléculaire**

Présentée par

**Olivia Doppelt-Azeroual**

Pour obtenir le titre de Docteur de l'Université Paris Diderot – Paris 7

**Développement d'une nouvelle méthode performante de classification  
des surfaces protéiques d'interaction.**

**Optimisations et extensions du logiciel MED-SuMo.**

Soutenue le Lundi 30 Mars 2009, devant le jury composé de :

Pr Gilbert Deléage, Professeur, Université Claude Bernard, Lyon 1	Rapporteur
Dr Anne Imberty, Directeur de Recherche, CNRS, Grenoble	Rapporteur
Pr Alessandra Carbone, Professeur, Université Pierre et Marie Curie, Paris 6	Examineur
Dr Cyril Daveu, Modélisateur, SANOFI-AVENTIS	Examineur
Pr Fernando Rodrigues-Lima, Professeur, Université Paris-Diderot, Paris 7	Examineur
Dr Alexandre de Brevern, Chargé de Recherche, INSERM	Directeur de thèse
Mr François Delfaud, PDG de MEDIT-SA	

# Remerciements

## **Abréviations**

GOLD: Genome On Line Database  
PDB: Protein Data Bank  
3D: tri-Dimensionnel  
ARN: Acide RiboNucléique  
ADN: Acide DésoxyriboNucléique  
RMN: Résonance Magnétique Nucléaire  
ATP: Adénosine Tri-Phosphate  
AFP: Aligned Fragment Pair  
SSE: Structural Secondary Element  
RMSD: Root Mean Square Deviation  
SCOP: Structural Classification of Proteins  
SCF: Surface Chemical Feature  
MED-sumo-clui: MED-SuMo Command Line User Interface  
MED-SuMo GUI: MED-SuMo Graphical User Interface  
MED-SMA: MED-SuMo Multi Approach  
MCL: Markov CLustering  
 $N_{eq}$ : Nombre équivalent d'états  
HSP: Heat Shock Protein  
GHKL: Gyrase Hsp90 histidine Kinase mutL  
NAD: Nicotinamide adenine dinucleotide  
GMP: Guanosine Mono-Phosphate  
GDP: Guanosine Di-Phosphate  
GNP: Guanylyl Imidodiphosphate  
GTP: Guanosine Tri-Phosphate  
ADP: Adénosine Di-Phosphate  
ANP: Adenyl Imidodiphosphate  
AMP: Adénosine Mono-Phosphate  
CDK2: Cyclin Dependant Kinase  
GSK3B: Glycogen synthase kinase 3 beta  
VEGFR-2: Vascular Endothelial Cell Growth Factor Receptor 2

---

# TABLE DES MATIÈRES

<b>REMERCIEMENTS</b> .....	<b>2</b>
<b>ABRÉVIATIONS</b> .....	<b>3</b>
<b>TABLE DES MATIÈRES</b> .....	<b>4</b>
<b>TABLES DES FIGURES</b> .....	<b>7</b>
<b>LISTES DES TABLEAUX</b> .....	<b>10</b>
<b>INTRODUCTION GÉNÉRALE</b> .....	<b>12</b>
<b>I. LES STRUCTURES DES PROTÉINES</b> .....	<b>16</b>
A.    STRUCTURES PROTEIQUES (3D).....	16
1. <i>État de l'art</i> .....	16
2. <i>Méthodes de résolution</i> .....	18
i.    La cristallographie aux rayons X .....	18
ii.   La spectroscopie RMN .....	21
iii.   La CryoMicroscopie Électronique en Transmission .....	22
iv.   Consortiums de génomique structurale .....	24
3. <i>Notion de qualité</i> .....	25
i.    La résolution .....	25
ii.   Fiabilité de l'affinement cristallographique .....	26
iii.   Vibration et désordre.....	28
B.    CLASSIFICATION DES STRUCTURES.....	29
1. <i>Méthodes de comparaison de structures</i> .....	30
i.    Généralités .....	30
ii.   Quelques méthodes de comparaison structurales.....	31
2. <i>Les bases de données existantes</i> .....	39
i.    SCOP .....	39
ii.   CATH .....	40
iii.   FSSP.....	42
3. <i>Applications et limites</i> .....	43
C.    NOTION DE SURFACE D'INTERACTION .....	44
1. <i>Propriétés des différents types d'interfaces</i> .....	44
i.    Interactions protéine-protéine .....	44
ii.   Interactions protéine-nucléotide.....	47
iii.   Interactions protéine–ligand.....	48
2. <i>Méthodes de localisation de sites de liaisons et annotation fonctionnelle</i> .....	50
i.    Méthodes basées sur la séquence .....	50
ii.   Méthodes basées sur la structure.....	54

<b>II. PRÉSENTATION DE MED-SUMO .....</b>	<b>66</b>
A.    HEURISTIQUE DE LA METHODE .....	66
B.    UTILISATION DU LOGICIEL.....	69
C.    INTERFACE GRAPHIQUE : MED-SuMo GUI .....	71
D.    DETAILS SUR LES COMPOSANTS DE MED-SUMO .....	73
1. <i>Différents types de requêtes</i> .....	73
2. <i>Différents types de bases de données</i> .....	73
i.    L'index de la PDB .....	74
ii.   La base de sites .....	74
iii.   La base de surface entière : la base Full.....	75
iv.   Plusieurs modes d'utilisation .....	75
3. <i>Le score MED-SuMo</i> .....	76
<b>III. MED-SUMO : OPTIMISATIONS ET NOUVEAUX DÉVELOPPEMENTS.....</b>	<b>80</b>
A.    MED-SUMO SERVEUR .....	80
1. <i>Nouvelles fonctionnalités</i> .....	80
i.    La base de sites .....	80
ii.   Communication client-serveur.....	81
2. <i>Annotation fonctionnelle avec MED-SuMo</i> .....	84
B.    UNE NOUVELLE METHODE PERFORMANTE DE CLASSIFICATION DES SURFACES	
PROTEIQUES D'INTERACTION: MED-SMA.....	91
1. <i>Description globale</i> .....	91
2. <i>Détails d'implémentation</i> .....	94
i.    Comparaison deux à deux.....	94
ii.   Score MED-SuMo entre multipatches.....	95
iii.   Matrice de Similarité.....	96
iv.   Lancement du programme .....	97
v.    Analyse des résultats.....	98
3. <i>Deux applications de MED-SMA</i> .....	101
i.    GHKL fold.....	101
ii.   Sites de liaison aux purines.....	109
4. <i>Adaptation de la méthode pour la classification de gros jeux de données</i> .....	132
i.    MED-Distribute .....	133
ii.   Stockage des résultats .....	135
C.    UNE NOUVELLE METHODE DE CONCEPTION DE NOVO DE MOLECULES ACTIVES .....	138
1. <i>Idée Générale</i> .....	138
2. <i>Mise en place du protocole</i> .....	141
i.    Construction de la base fragments .....	141
ii.   Cartographie des surfaces d'interactions avec des MED-Portions .....	145
iii.   Génération de molécules réelles avec MED-Hybridise .....	146

---

3. Application de l'approche fragmentale sur le récepteur protéine kinase VEGFR-2.....	148
<b>IV. CONCLUSION GÉNÉRALE .....</b>	<b>155</b>
<b>BIBLIOGRAPHIE.....</b>	<b>160</b>
<b>ARTICLES.....</b>	<b>183</b>

---

## TABLES DES FIGURES

<b>Figure 1</b> : Évolution du nombre de structures stockées dans la PDB .....	18
<b>Figure 2</b> : Représentation des différentes étapes de la cristallographie à rayon X. ....	20
<b>Figure 3</b> : Exemple de combinaison de méthodes de résolution de structures de protéine .....	24
<b>Figure 4</b> : Carte de densité électronique en fonction de la résolution des structures.....	26
<b>Figure 5</b> : A. Localisation des angles $\Phi$ (N-C $\alpha$ -C) et $\psi$ (C $\beta$ -C $\alpha$ -C) et B. diagramme de Ramachandran.....	27
<b>Figure 6</b> : Principe du codage d'une protéine à l'aide d'un alphabet structural. ....	38
<b>Figure 7</b> : Exemple de classification hiérarchique présente dans CATH. ....	42
<b>Figure 8</b> : Représentation des atomes enfouis et accessibles des interfaces. ....	47
<b>Figure 9</b> : Trois exemples de complexes protéiques décrits. ....	50
<b>Figure 10</b> : Représentation d'un résultat de Ligsite. ....	55
<b>Figure 11</b> : Exemple de site de liaison détecté par la méthode PocketFinder sur la structure 2IZI.....	56
<b>Figure 12</b> : Sept étapes de l'algorithme du logiciel PASS.....	57
<b>Figure 13</b> : Exemple d'analyse de la protéine kinase 1H1S avec les champs de force de GRID-MIFs.....	60
<b>Figure 14</b> : Procédure de comparaison disponible dans MED-SuMo.....	68
<b>Figure 15</b> : Lancement du même <i>run</i> MED-SuMo de deux manières différentes.....	70
<b>Figure 16</b> : L'interface graphique de MED-SuMo: MED-SuMo GUI. ....	72
<b>Figure 17</b> : Liste et couleurs des SCFs utilisés dans MED-SuMo GUI.....	72
<b>Figure 18</b> : Différents types de requêtes possibles pour MED-SuMo sur 1B54.....	73
<b>Figure 19</b> : Les différents modes d'utilisation du logiciel MED-SuMo .....	76
<b>Figure 20</b> : Architecture globale de MED-SuMo .....	82
<b>Figure 21</b> : Exemples de résultats obtenus par MED-SuMo sur la protéine TM1012....	87
<b>Figure 22</b> : Superposition de la protéine YBL036C avec d'autres alanines racémases détectées par MED-SuMo. ....	89
<b>Figure 23</b> : Procédure globales de MED-SMA.....	92
<b>Figure 24</b> : Les six étapes de la méthode MED-SMA. ....	94
<b>Figure 25</b> : Exemple simplifié illustrant le calcul du score entre deux <i>multipatches</i> . ....	96

---

<b>Figure 26</b> : Exemple des transformations appliquées à la matrice de similarité entre cinq <i>multipatches</i> . .....	97
<b>Figure 27</b> : Vue 2D de la molécule radicicol. ....	102
<b>Figure 28</b> : Représentation de deux HSP90. ....	102
<b>Figure 29</b> : Protéines membres de la superfamille SCOP GHKL.....	103
<b>Figure 30</b> : Superposition de deux topoisomérases VI séparées par MED-SMA.....	106
<b>Figure 31</b> : Superposition de quatre protéines issues de trois familles SCOP rassemblés par MED-SMA.....	107
<b>Figure 32</b> : Vue rapprochée autour du ligand radicicol. ....	108
<b>Figure 33</b> : Distribution de la taille des groupes. ....	111
<b>Figure 34</b> : Structures 2D des ligands puriques sélectionnés pour le jeu de données..	112
<b>Figure 35</b> : Distribution des motifs PROSITE au sein des MED-clusters. ....	114
<b>Figure 36</b> : Distribution des valeurs de $N_{eq}$ .....	116
<b>Figure 37</b> : Distribution des valeurs de $N_{eq}$ dans une représentation 2D des MED-clusters.....	117
<b>Figure 38</b> : Exemples de superposition de sites de liaison en 3D par MED-SuMo.....	118
<b>Figure 39</b> : Superposition de ligands du cluster 33.....	119
<b>Figure 40</b> : Représentation d'un réseau de groupes dans la classification.....	122
<b>Figure 41</b> : Illustration de deux liens inter-groupes .....	124
<b>Figure 42</b> : Résultats d'une comparaison du site de liaison d'une cycline dépendante kinase 2 avec tous les sites de liaison du jeu de données avec MED-SuMo.....	126
<b>Figure 43</b> : Superposition 3D d'une biotine carboxylase et d'une CDK2. ....	127
<b>Figure 44</b> : Superposition des ligands GTP et ANP dans deux sites de liaison très similaires. ....	129
<b>Figure 45</b> : Exemple de découpage du jeu de données pour optimiser l'étape de comparaison multiple de MED-SMA. ....	134
<b>Figure 46</b> : Fonctionnement du programme <i>MED-Distribute</i> . ....	135
<b>Figure 47</b> : Architecture de la base <i>multi.db</i> . ....	136
<b>Figure 48</b> : Représentation d'un exemple de MED-Portions.....	140
<b>Figure 49</b> : Protocole de l'approche fragmentale pour la construction de la base de MED-Portions. ....	142
<b>Figure 50</b> : Description du ligand GIG par 10 MED-Portions.....	144



---

<b>Figure 51</b> : Exemple d'enchaînement des programmes pour l'approche fragmentale avec le <i>framework</i> Scitegic Pipeline Pilot <sup>TM</sup> 7.0 [143]. .....	147
<b>Figure 52</b> : Sous-structure intéressante pour la protéine kinase VEGFR-2 .....	148
<b>Figure 53</b> : Représentation de la requête soumise à la base de MED-Portions. ....	149
<b>Figure 54</b> : MED-Portions détectés par MED-SuMo. ....	150
<b>Figure 55</b> : Exemple d'hybrides obtenus avec le programme MED-Hybridize. ....	151
<b>Figure 56</b> : Molécule D de la figure 55 représentée 3D dans la structure 2OH4.....	152
<b>Figure 57</b> : Exemples de surface de Connolly. ....	177

---

## LISTES DES TABLEAUX

<b>Tableau 1</b> : Comparaison de 13 méthodes récentes de comparaison et détermination de structure de protéines. ....	39
<b>Tableau 2</b> : Comparaison des méthodes de détection de sites basées sur des critères géométriques sur un jeu de 210 structures. ....	58
<b>Tableau 3</b> : Description des MED-clusters obtenus à l'issu de la classification par MED-SMA par rapport aux familles SCOP.....	104
<b>Tableau 4</b> : Matrice distribution des ligands dans les MED-clusters. ....	113

## Introduction générale

---

## Introduction générale

L'amélioration des techniques de séquençage à haut débit est une conséquence directe et un des succès des projets Génome. Depuis l'an 2000, l'ensemble du génome humain, ainsi que de nombreux autres génomes sont disponibles [1]. Le nombre de séquences accessibles a augmenté de manière exponentielle [2] mais la quantité de séquences n'ayant aucune annotation fonctionnelle grandit aussi très rapidement. Á titre d'exemple, 57% des gènes du vecteur du paludisme *Plasmodium falciparum* sont non annotés. La fonction d'un gène s'exprime au travers de son produit, la protéine, dont le rôle est directement lié à sa structure tridimensionnelle (3D). La résolution d'une structure de protéine issue d'un gène de fonction inconnue est un moyen de comprendre les mécanismes biochimiques dans lesquels elle est impliquée. De ce fait et dans le but d'augmenter le nombre de structures de protéines résolues, des travaux ont permis une amélioration nette des méthodes de résolution des structures 3D des protéines, notamment grâce au rayonnement synchrotron [3]. Les trois méthodes les plus communément utilisées sont la cristallographie à rayons X, la spectroscopie à résonance magnétique nucléaire et la microscopie électronique à transmission.

Les effets positifs des projets Génome ont incité à la mise en place de la « Protein Structure Initiative » (PSI) [4] qui favorise la création de consortiums de génomique structurale à travers le monde. Leurs objectifs premiers sont de résoudre le plus grand nombre de structures de protéines, le plus rapidement possible, et si possible des structures ayant des repliements inconnus. Ces structures résolues sont régulièrement mises à la disposition de la communauté scientifique. Les consortiums ont ainsi largement contribué à l'évolution quadratique de la taille de la banque de données publique Protein Data Bank (PDB) [5].

Le type des structures résolues ces dernières années a aussi évolué. Leur qualité s'est affinée avec des résolutions toujours plus précises et de nouveaux types de structures sont apparus. En effet, de plus en plus de gros complexes tels les virus ou les ribosomes sont décrits précisément, alors que des structures de très fine résolution co-cristallisées avec différents petits ligands sont déterminées [6]. Certains d'entre eux sont naturels tel l'ATP, alors que d'autres sont des molécules actives d'origine synthétique. La PDB est devenue une source majeure d'informations structurales expérimentales sur les interactions entre les ligands et les protéines. Toutefois, elle contient aussi de nombreuses protéines purifiées et séquencées, mais dont la fonction biochimique reste inconnue. Aujourd'hui plus de 3000 structures ne sont pas fonctionnellement annotées.

---

La fonction d'une protéine ne s'exprime qu'à partir du moment où elle interagit avec un (ou plusieurs) partenaire(s) spécifique(s). Ces interactions sont la base de tous les mécanismes biologiques. Les interactions protéine-ligand ont des rôles capitaux par exemple dans les fonctions de transport et de transmission de signaux cellulaires. La détection et la comparaison de surfaces d'interaction sont ainsi des étapes essentielles pour annoter fonctionnellement les protéines, et pour l'élaboration de nouveaux médicaments.

Le logiciel SuMo et son produit dérivé à vocation industrielle MED-SuMo représentent une des premières approches utilisant la position relative des groupements chimiques fonctionnels élémentaires disponibles à la surface des protéines, pour comparer les surfaces d'interaction [7, 8, 9, 10]. Elle possède des analogies avec d'autres méthodes [11,12]. Toutefois, son approche est particulièrement innovante et son originalité réside dans ses descripteurs, les « *Surface Chemical Features* » (SCFs) qui, rassemblés en triplets, forment un graphe constituant la clé des performances élevées de l'heuristique de comparaison. Cette procédure se base sur la détection de sous-graphes communs. Particulièrement rapide, l'approche permet la comparaison d'un site de liaison à tous les sites de liaisons de la PDB en quelques minutes. SuMo/MED-SuMo intègre également en sortie, une étape de superposition 3D basée sur la complémentarité des surfaces d'interaction identifiées. Cette superposition s'applique non seulement à la protéine identifiée dans la PDB mais aussi à l'ensemble de ses ligands co-cristallisés. La comparaison des surfaces protéiques offre un grand nombre d'applications nouvelles de *drug design*, basées sur l'hypothèse selon laquelle les surfaces similaires interagissent avec des molécules de même type.

Mon travail de thèse repose sur trois points:

Tout d'abord, j'ai pu participer activement à l'optimisation du logiciel MED-SuMo, notamment dans le cadre de la mise en place des communications avec la nouvelle interface graphique dédiée, MED-SuMo GUI. J'ai également contribué à l'amélioration de la méthode de détection des ligands ou encore mis en évidence l'intérêt de MED-SuMo pour l'annotation fonctionnelle de protéines hypothétiques à travers deux applications. La première porte sur la protéine Tm1012 [13] et la seconde sur la protéine YBL036C [14].

En second lieu, le développement et l'implémentation d'une nouvelle méthode de classification des sites de liaison protéique MED-SMA, ainsi que son adaptation à l'analyse de grands jeux de données ont constitué la partie essentielle de mon travail de recherche. Cette approche a été validée sur la famille SCOP comprenant les HSP90; le repliement

---

Bergerat [15], puis sur l'ensemble des sites du purinôme ([16] en révision). Une classification de tous les sites de la PDB (> 90000 sites) est en cours.

Enfin, j'ai participé au développement d'une nouvelle méthode de conception de molécules actives *de novo* combinant la détection MED-SuMo de similitudes locales des surfaces des protéines avec une approche fragmentale [17,18].

# PARTIE I : Les structures des protéines

---

## I. Les structures des protéines

### A. Structures protéiques (3D)

#### 1. État de l'art

L'ensemble du génome humain est accessible à la communauté scientifique depuis plusieurs années [19]. De nombreux autres projets de séquençage de génomes concernant divers organismes tel que *Helicobacter pylori* ont aussi abouti, et d'autres sont en cours comme celui du chien, *Canis canis* [1]. La banque de données GOLD en répertorie 4613 au 1 mars 2009 [1]. Ces différents projets ont généré une quantité considérable d'informations trop souvent difficilement exploitables. En effet, après séquençage et localisation d'un gène, il reste à déterminer sa fonction, première étape pour caractériser son rôle dans un processus biologique [20]. Sans cette information, l'utilité d'une séquence est négligeable. La manière la plus commune d'attribuer une fonction à un nouveau gène est d'identifier des séquences proches. En fonction du degré d'identité des séquences, il est possible d'inférer ou non une fonction au nouveau gène. La méthode de recherche de séquences proches la plus couramment utilisée est PSI-BLAST [21] ; d'autres approches performantes sont basées sur les chaînes de Markov cachées (HMM) [22]. Ces méthodes sont rapides et utilisées dès le séquençage d'un gène. Malgré l'efficacité de ces méthodes, de plus en plus de gènes orphelins sont répertoriés: les séquences « *ORFans* » [23]. Ces gènes putatifs ne sont associés à aucune information structurale et fonctionnelle. L'étude de la fonction des gènes se fait par des approches biologiques qui se basent sur le produit du gène: la protéine. Au niveau sub-cellulaire, la protéine va participer aux différents équilibres biochimiques en exerçant une fonction particulière. Des techniques de biologie moléculaire et de biochimie (northern blot, hybridation *in situ* et western blot, immunofluorescence ...) permettent d'observer l'expression d'un gène (ARN) ou d'étudier son produit (protéine). Il est aussi possible d'identifier expérimentalement ses partenaires protéiques (double hybride, immuno-précipitation ...), résultats pouvant donner des informations importantes sur la fonction du gène étudié. D'autres stratégies basées sur des modifications du niveau d'expression d'un gène (les puces à ADN...) ou sur une modification de l'activité de la protéine, permettent d'approfondir les différentes études engagées. Ces techniques expérimentales sont parfois difficiles à mettre en œuvre et coûteuses.

*In vivo* la chaîne polypeptidique des protéines se replie pour adopter une architecture tridimensionnelle (3D). Leur fonction est directement liée à ce repliement (en anglais, *fold*).



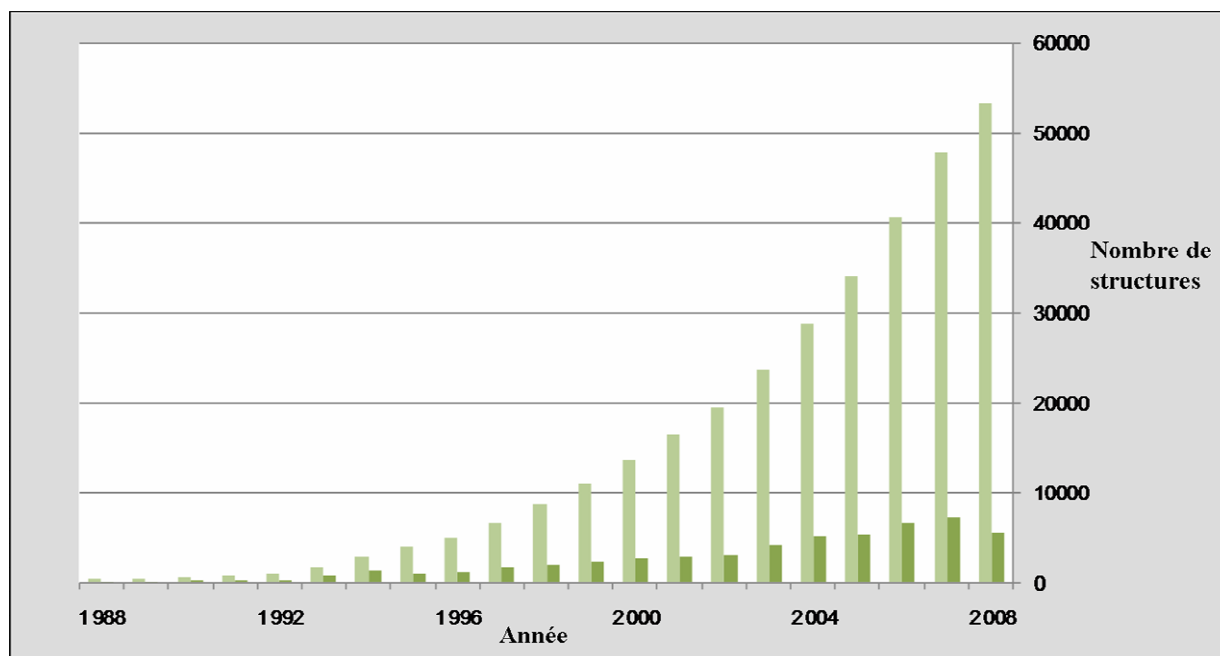
---

Les interactions mises en jeu sont des interactions non-covalentes de faible énergie: liaisons hydrogènes, interactions électrostatiques, contacts Van der Waals et effets hydrophobes. Le gain d'énergie libre de ces types d'interaction est environ dix à cent fois plus faible que celui d'une liaison covalente. La stabilité structurale 3D est aussi assurée la complémentarité des interactions avec le solvant (molécules d'eau et ions). La structure d'une protéine repose donc sur des interactions faibles. Sa conformation est stable dans des conditions physico-chimiques données mais s'adapte à l'environnement ; la protéine n'est pas une macromolécule rigide. Cette propriété est à la base de nombreuses caractéristiques fonctionnelles essentielles. Les modifications de conformation peuvent résulter, par exemple, d'un changement de pH, d'une interaction avec une petite molécule ou encore avec d'autres macromolécules (protéines, membranes, ADN...). Cette flexibilité conformationnelle est capitale pour appréhender les fonctions biologiques.

La biologie structurale joue un rôle majeur dans l'étude des macromolécules et des complexes associés (protéines, acides nucléiques, lipides et petits ligands). Son but est de déterminer leurs structures, de comprendre les propriétés dynamiques de ces assemblages, de suivre les changements de conformations propres à l'exploration de l'espace réactionnel et d'associer ces informations à des données fonctionnelles afin de comprendre les mécanismes moléculaires. La détermination des mécanismes biochimiques contribue à élucider le fonctionnement de la machinerie cellulaire. Elle joue aussi un rôle majeur dans le cadre de la conception de nouveaux composés spécifiques de certaines cibles pharmaceutiques. La détermination de la structure des protéines constitue donc un pas essentiel dans la création rationnelle de nouveaux médicaments [24]. La biologie structurale se situe à l'interface de la biologie, de la chimie et de la physique, alors que le problème à résoudre concerne la biologie, et que les mécanismes réactionnels de catalyse enzymatique sont des processus chimiques. La physique, quant à elle, contribue de façon importante aux méthodes expérimentales de résolution de structures, aux approches de modélisation moléculaire, et aussi à la compréhension de certains processus (par exemple la contraction musculaire ou le fonctionnement de moteurs moléculaires).

La Protein Data Bank [5] (PDB) est le dépôt public privilégié des structures 3D de macromolécules biologiques; elle contient principalement des protéines (~50000 structures) mais aussi des acides nucléiques (ADN, ARN) (~2000 structures). Ces structures sont essentiellement déterminées par 3 méthodes : la cristallographie aux rayons X, la spectroscopie à Résonance Magnétique Nucléaire (RMN) ainsi que la cryo microscopie électronique à transmission (cryoMET). Ces données expérimentales sont déposées dans la

PDB par des biophysiciens du monde entier. Leur consultation est gratuite et peut se faire directement depuis le site web de la banque ([www.rcsb.org](http://www.rcsb.org)). La PDB est la principale source de données expérimentales de biologie structurale. La figure 1 montre l'évolution quadratique du nombre de structures dans la PDB depuis 1988 à nos jours.



**Figure 1 : Évolution du nombre de structures stockées dans la PDB**

En vert foncé, le nombre de structures déposées durant l'année, en vert clair, le nombre cumulé de structures dans la PDB.

## 2. Méthodes de résolution

Une protéine est composée de plusieurs milliers d'atomes qui, engagés dans des fonctions chimiques, vont interagir entre eux pour générer son repliement 3D, qui sera aussi le siège de son interaction avec d'autres molécules. En biologie structurale, la notion de structure d'une protéine concerne son repliement mais tend surtout à sa résolution à l'échelle atomique. La structure nous révèle alors l'arrangement détaillé de chaque atome dans chaque partie de la protéine. La résolution d'une structure est un processus complexe, coûteux et long. Les trois méthodes les plus utilisées pour l'obtention de structures sont la cristallographie aux rayons X, la RMN et la cryo-microscopie électronique à transmission.

### i. *La cristallographie aux rayons X*

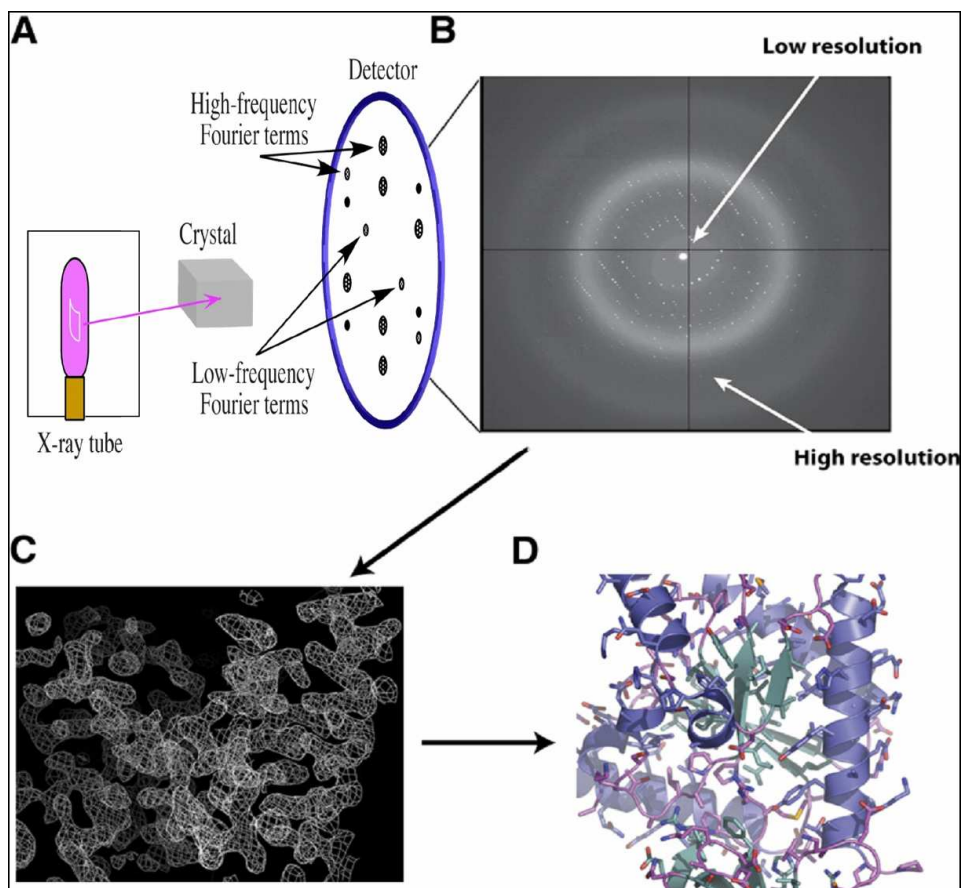
La cristallogénèse nécessite de disposer d'une quantité importante et purifiée du complexe macromoléculaire. Il est donc nécessaire de la sur-exprimer dans des cellules

---

transfectées avec un plasmide contenant sa séquence codante. L'obtention de cristaux homogènes ou monocristaux est un processus ardu qui n'est pas toujours couronné de succès.

Ensuite, pour étudier la structure des macromolécules avec un système optique, il convient de disposer d'une source de rayonnement compatible avec la résolution à atteindre. Afin de visualiser la manière dont les atomes ou groupes d'atomes sont arrangés dans l'espace, le rayonnement doit avoir une longueur d'onde du même ordre de grandeur que les distances inter atomiques (~ de l'ordre de l'Angstrom). Les rayons X conviennent pour étudier l'arrangement des atomes dans les protéines car leurs longueurs d'onde varient entre 0,01 Å et 10 Å. Comme toutes les ondes électromagnétiques, les rayons X provoquent un déplacement du nuage électronique par rapport au noyau des atomes. Les oscillations induites provoquent une réémission d'ondes électromagnétiques de même fréquence. Les interférences des rayons diffusés vont être alternativement constructives ou destructives. Selon la direction de l'espace, le flux de photons X sera important, ou au contraire très faible. Les variations ainsi créées constituent le phénomène de diffraction. Ce phénomène a été découvert par Max Von Laue (Prix Nobel en 1914).

Pour les protéines cristallisées, les plans dans lesquels se situent les atomes se comportent comme des surfaces réfléchissant le rayonnement X, les centres diffractant étant les atomes constitutifs de la molécule. La diffraction aux rayons X consiste à envoyer un faisceau fin de rayons X sur le cristal sous un certain angle (cf. Figure 2). Une partie des rayons est diffractée et ensuite détectée. Les réflexions sont enregistrées en tournant le cristal. Un diagramme est ainsi obtenu, il contient un grand nombre de taches arrangées régulièrement. Une fois le problème de phase résolu, il est possible de calculer par transformée de Fourier, la densité électronique de chaque molécule contenue dans le cristal (à partir des taches de diffraction).



**Figure 2 : Représentation des différentes étapes de la cristallographie à rayon X.**

A) Schéma de la diffraction. B) Diagramme de diffraction obtenu après le rayonnement. C) Carte de densité électronique de la protéine. D) Structure de la protéine (figure extraite du livre de Gail Rhodes [25]).

L'augmentation massive du nombre de structures déterminées par cristallographie à rayon X est due en partie à l'utilisation répandue des rayonnements synchrotron et des différentes techniques qui en découlent. Ces sources de rayonnement permettent la production de rayons X intenses, stables, de longueurs d'ondes variables et de grande qualité optique. Ils offrent des données exploitables sur les cristaux de macromolécules que des méthodes classiques ne permettaient pas d'étudier : mailles plus grandes (et donc molécules ou complexes plus grands), cristaux plus petits voire imparfaits. En outre, le rayonnement synchrotron a considérablement accéléré la vitesse d'acquisition des données, réduisant celle-ci à quelques heures au lieu de quelques semaines. Ainsi, en 2 ou 3 jours un cristallographe peut récolter des données qui n'auraient pu être accessibles qu'après plusieurs mois avec un rayonnement conventionnel [25]. La très grande majorité des études cristallographiques récentes reposent sur l'utilisation de données collectées à partir d'un rayonnement synchrotron. Une conséquence directe concerne l'évolution quadratique observée de la PDB. Au 21 janvier 2009, celle-ci contient ~47 000 structures issues de la cristallographie à rayon X sur les ~55 000 enregistrements totaux de cette base. L'augmentation massive découle aussi

---

d'une volonté politique d'obtention de structures à haut débit (point détaillé dans le paragraphe I.iv).

## ii. *La spectroscopie RMN*

Le principe de la RMN consiste à plonger une protéine en solution dans un champ magnétique intense (entre 10 et 20 Tesla) [26]. Cette technique d'analyse possède un avantage non négligeable : elle est non dommageable pour les échantillons protéiques. Elle se base sur une propriété quantique particulière des particules, le spin, qui caractérise le comportement d'une particule sous l'effet de la symétrie de rotation de l'espace. Certains noyaux possèdent un spin nul (ceux dont le nombre de protons et de neutrons sont tous les deux pairs) mais d'autres ont un spin nucléaire différent de zéro, ce qui implique qu'il est possible de leur associer un moment magnétique de spin se comportant comme un moment magnétique (telle une sorte de petit aimant). Les atomes de carbone 12 et d'oxygène 16 sont très répandus mais leur spin nucléaire est nul. En revanche l'hydrogène n'a qu'un proton, l'azote 14 en possède sept et le carbone 13 est aussi utilisé. Les moments magnétiques nucléaires de ces types d'atomes sont ainsi non nuls. Sous l'impulsion du champ magnétique, les spins des noyaux atomiques composant la protéine s'orientent le long de son axe principal. Lorsque l'état d'équilibre de ces spins est perturbé, il est possible de mesurer le courant induit lors du retour à l'équilibre à l'aide d'une bobine de réception située à proximité de l'échantillon. Le signal RMN enregistré contient la combinaison de l'ensemble des contributions des différents atomes en solution. L'application de la transformée de Fourier fournit les différentes fréquences de résonance des spins observés pour chaque atome. La valeur de la fréquence de résonance rapportée à une fréquence de référence et exprimée en ppm est appelée « déplacement chimique ». Ce dernier est fortement sensible au type d'atome lié au proton (carbone, oxygène, azote), et à son implication dans des liaisons non covalentes comme les liaisons hydrogènes ou encore à la proximité de noyaux aromatiques. Ainsi, chaque proton résonne à travers un déplacement chimique qui le caractérise. Afin de pouvoir séparer les résonances, des spectres à multiple dimension ont été développés. Ces comportements permettent de corrélater un proton avec le carbone ou l'azote sur lequel il est lié ainsi que ses voisins dans l'espace ou à travers les liaisons.

La détermination de la structure des protéines à partir des données RMN passe par deux étapes. La première consiste à établir la liste des déplacements chimiques de tous les noyaux observables de la protéine,  $^1\text{H}$ ,  $^{15}\text{N}$ ,  $^{13}\text{C}$ , en utilisant les spectres à trois dimensions corrélant les atomes au travers des liaisons chimiques. La seconde étape comprend l'analyse des

---

expériences *Nuclear Overhauser Effect* (nOe) qui listent les corrélations résultant de la proximité spatiale des différents noyaux. Les corrélations sont attribuées à des couples de noyaux et sont introduites sous forme de contraintes de distance dans une procédure transformation d'un espace de distance en coordonnées cartésiennes atomiques, tel que ARIA [27]. L'ensemble de ces contraintes de distance représente un maillage avec lequel le modèle structural de la protéine doit être compatible. D'autres données expérimentales issues de mesures RMN (déplacements chimiques, constantes de couplage), donnent accès à des informations angulaires caractéristiques de la structure de la protéine et sont également exploitées lors de la proposition des modèles structuraux. De récents développements ont permis d'élargir la méthode à la résolution de structures de complexes entre macromolécules ou de complexe de protéines en interaction avec des petites molécules [28]. De même des études complexes permettent désormais d'obtenir des informations sur des protéines transmembranaires contenues dans des bicouches [29]. A l'heure actuelle, les résultats de la RMN correspondent à 14% des 55 000 structures de la PDB.

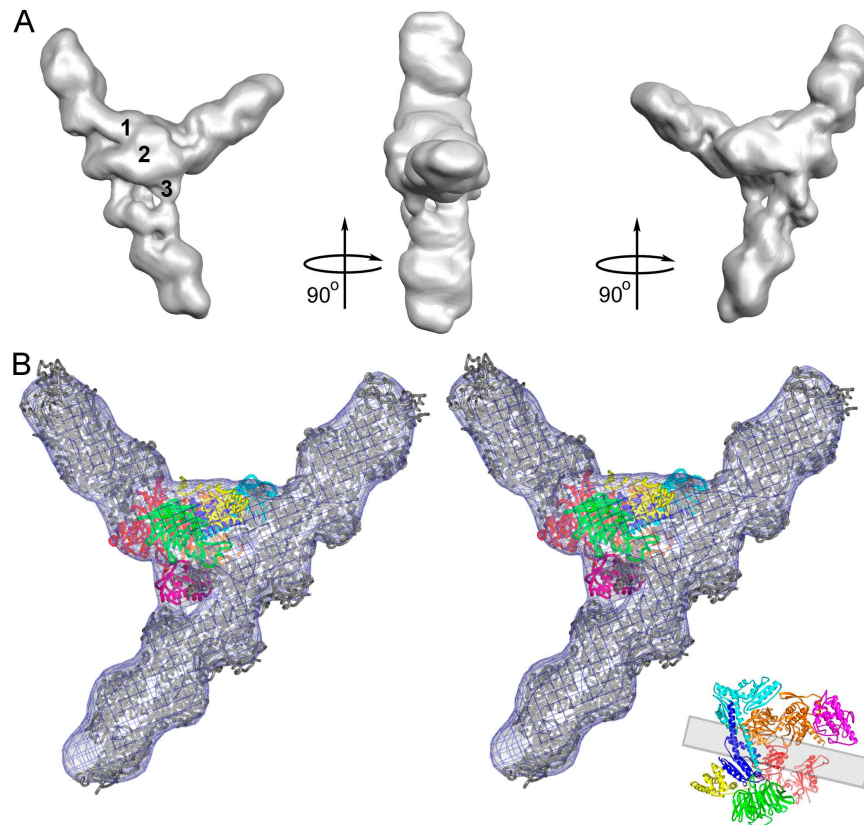
### iii. *La CryoMicroscopie Électronique en Transmission*

La Cryomicroscopie Électronique à Transmission (Cryo-MET) de macromolécules biologiques est employée pour déterminer l'architecture de complexes de grandes tailles tels les ribosomes, les protéines chaperons ou les virus. Des progrès technologiques ont permis d'étudier les structures et les interactions moléculaires à des résolutions inférieures au nanomètre. La particularité de la Cryo-MET 3D est qu'elle fournit une structure à haute résolution globale des assemblages macromoléculaires, préservés dans leurs conditions physiologiques d'activité. Pour ce faire, une fois les échantillons constitués, ils sont instantanément congelés dans une mince pellicule conservant leurs environnements habituels (hydratation, pH, sels, détergents, cofacteurs, analogues de l'ATP, ARN, ADN, *etc.*). Dans de telles conditions, la formation de cristaux de glace qui endommageraient l'intégrité des assemblages est évitée. Les échantillons sont ainsi préservés dans une fine pellicule de glace vitreuse. La cryomicroscopie –proprement dite- consiste à bombarder l'échantillon avec un faisceau d'électrons. Les électrons primaires, issus du canon à électrons, frappent la surface de l'échantillon; ils sont diffusés de manière élastique ou inélastique, la zone influencée prenant la forme d'une poire. Les images des structures résultent de l'interférence entre la partie du faisceau d'électrons diffusée par l'objet et la partie non diffusée.

Des techniques d'analyse d'images et de calculs intensifs sur réseaux d'ordinateurs sont utilisées pour extraire et traiter le signal provenant des images bruitées de Cryo-MET. Elles

---

servent à créer une carte de Cryo-MET 3D qui sera combinée avec d'autres approches (rayon X, RMN, etc.) afin de relier l'aspect dynamique de ces structures à leurs activités biologiques. Aujourd'hui, 199 structures issues de Cryo-MET sont disponibles dans la PDB. L'approche de Cryo-MET 3D combinée aux autres méthodes est désormais employée pour résoudre les structures de protéines, et pour fournir une vision dynamique multi-échelles des assemblages macromoléculaires en action. Les travaux de Rouiller et collaborateurs ont permis la détermination de la structure de la jonction de branche d'actine contenant le complexe Arp2/3 à une résolution de 26 Å [30]. Cette jonction est un élément clé pour la compréhension de différents mécanismes physiques des cellules comme leur motilité ou celle de vésicules intracellulaires en initiant la formation des filaments d'actine. La résolution par cristallographe à rayon X de la forme inactive du complexe Arp2/3 a révélé l'architecture des sous-unités mais pas la manière dont les branchements des filaments sont formés. La compréhension du mécanisme nécessite la connaissance de la structure 3D de la jonction. La Figure 3 illustre la structure de la jonction et la manière dont le complexe Arp2/3 est placé. Cette reconstruction fournit une armature structurale pour comprendre le mécanisme de formation des banches. Elle montre également l'importance de la combinaison de méthodes de résolution de structures.



**Figure 3 : Exemple de combinaison de méthodes de résolution de structures de protéine.**

A. Trois vues de la structure de la jonction du filament d'actine déterminée avec une résolution de 26 Å. B. Vue du placement du complexe Arp2/3 résolu par cristallographie dans la structure 3D de la jonction, figure extraite de [30].

#### iv. Consortiums de génomique structurale

Les descriptions des différentes méthodes de détermination de structures protéiques montrent clairement que la résolution d'une structure n'est pas un processus trivial. Elle demande un grand savoir technique, du temps, de la puissance de calcul et des moyens financiers importants. Le projet Génome a été une initiative de la communauté scientifique internationale pour séquencer l'ensemble du génome humain. Ce projet a permis des avancées technologiques et scientifiques sans précédent notamment pour les méthodes de séquençage [3]. De manière équivalente, la *Protein Structure Initiative* (PSI) a été fondée par le *National Institute of Health* (NIH) aux États-Unis d'Amérique en 2000 afin de financer différents consortiums de génomique structurale présents dans le monde entier. Le but de ces consortiums est d'étendre les connaissances sur les structures de macromolécules tout en réduisant le coût moyen de la détermination des structures [31]. Chaque consortium s'engage à rendre publique les structures résolues, chacun d'eux ayant par ailleurs des objectifs différents, principalement en ce qui concerne la résolution de famille entière de protéine ou le



---

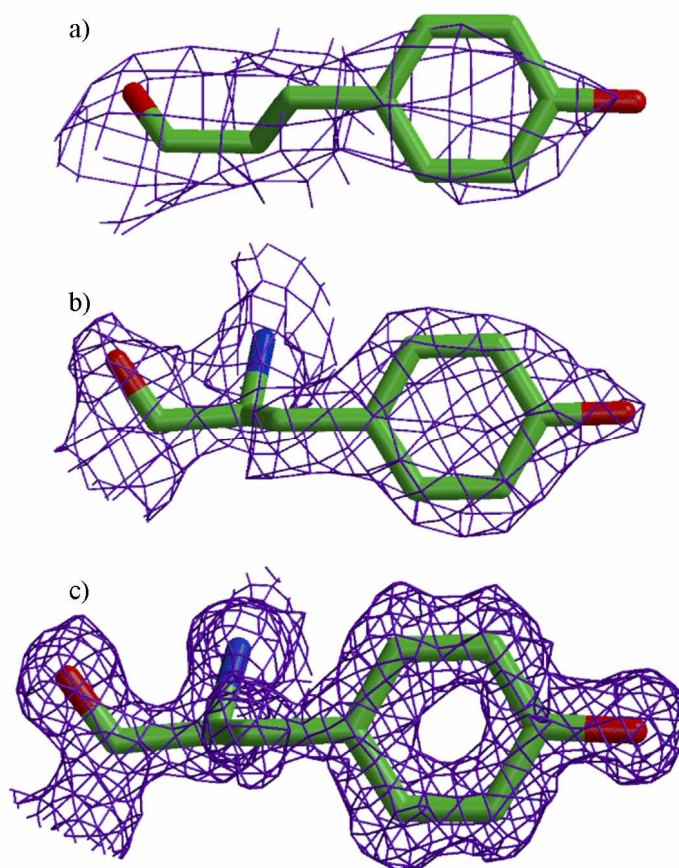
nombre de structures à rendre publique (exemple issue de [32]). Le but principal de la PSI est toutefois plus large. En effet, son objectif est le développement d'un protocole automatique permettant la détermination à haut débit de structures protéiques afin d'aboutir à la connaissance d'au moins une structure par type de repliement existant. Cette recherche est ardue et les sommes allouées importantes. Actuellement, l'avenir de la prochaine vague de financement du PSI n'est pas assuré [4].

### 3. Notion de qualité

Depuis la création de la PDB [33], un nombre important de méthodes utilisant la structure des protéines ont été développées, aussi bien pour visualiser une protéine, pour analyser un nouveau *fold*, ou pour y rechercher un site de liaison particulier. Ces méthodes utilisent les structures protéiques le plus souvent comme un 'objet' optimal. Toutefois, des erreurs ont cependant déjà été détectées dans la PDB [34]. Comment est-il possible de s'assurer de la qualité d'une structure ? Différents paramètres sont à prendre en considération.

#### i. La résolution

La résolution d'une structure donne une mesure de la quantité de détails pouvant être « discernés » sur une carte de densité électronique calculée à partir d'un cristal. Elle dépend d'une part de la qualité de ce cristal (s'il est bien ordonné), mais également de l'orientation des faisceaux qui le traversent. En général, les chaînes latérales sont difficilement visibles à basse résolution (supérieure ou égale à 4 Å). Ces structures ne conviennent donc pas pour étudier les chaînes latérales ou pour détecter des interactions possibles avec la protéine étudiée. La Figure 4 illustre les effets de la résolution sur la qualité des cartes de densité électronique. À 3 Å, le squelette de la protéine s'esquisse au travers de la densité. Ce n'est qu'à partir de 2 Å que les chaînes latérales peuvent être placées avec une certaine confiance. Les structures les plus précises sont celles ayant une résolution atomique de l'ordre de 1,2 Å ou 0,9 Å, qualité permettant le positionnement des atomes d'hydrogène.



**Figure 4 : Carte de densité électronique en fonction de la résolution de la structure protéique.**  
a) Résolution à 3 Å, b) résolution à 2 Å, c) résolution à 1,2 Å (figure extraite de [35], pages 281)

Les valeurs de résolution des structures de la PDB sont très variées, elles commencent à 0.5 Å, avec un pic autour de 2 Å et peuvent aller jusqu'à 70 Å pour des structures de très grande taille résolues par cryo-MET (par exemple, la structure du rhinovirus code PDB: 1D3I a une résolution de 33 Å).

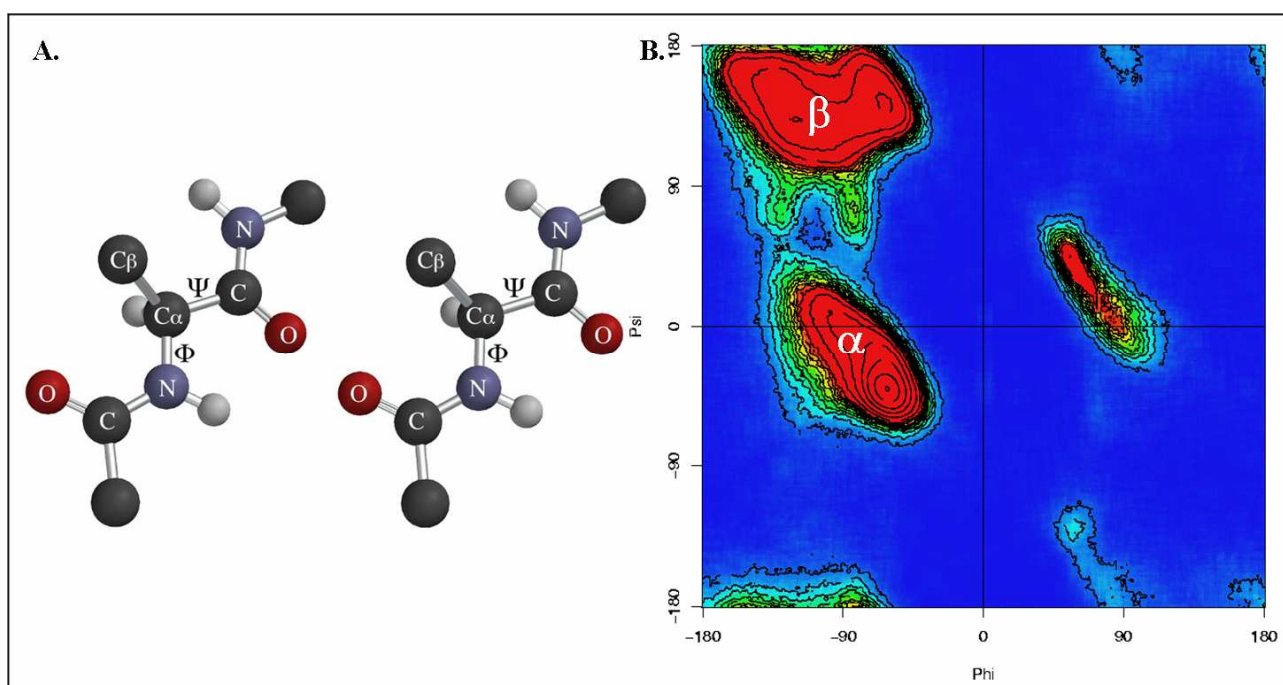
La résolution est probablement la mesure la mieux adaptée pour définir la qualité globale d'une structure de protéines. Toutefois, elle ne peut être calculée sur les structures issues de la RMN qui se base sur la présence ou l'absence d'atomes. Il convient donc de considérer d'autres paramètres.

### ii. *Fiabilité de l'affinement cristallographique*

Les dernières étapes de la détermination d'une structure sont une alternance de différentes phases d'affinement. Elles ont pour objectif la convergence vers une structure de mieux en mieux définie. Les différents paramètres suivants sont à surveiller.

## Les paramètres structuraux

Comme les contraintes sur les longueurs de liaison et sur les angles des résidus sont levées, leurs valeurs doivent être surveillées. Un écart quadratique moyen (RMSD (Root Mean Square Deviation)) sur les longueurs de liaison, et sur les déviations angulaires  $\Phi$  (N-C $\alpha$ -C) et  $\psi$  (C $\beta$ -C $\alpha$ -C) (cf. Figure 5) sont calculés et ne doivent pas excéder respectivement 0,02 Å et 4°. Traditionnellement, les angles  $\Phi$  et  $\psi$  sont représentés les uns par rapport aux autres dans un diagramme de Ramachandran [36]. La répulsion stérique au sein de chaque acide aminé a un effet restrictif sur ces angles qui ne peuvent adopter qu'un nombre limité de valeurs. La figure 5b représente les angles possibles. Les  $\alpha$  et  $\beta$  sont mis sur les valeurs d'angles caractéristiques des structures secondaires, hélice  $\alpha$  et feuillet  $\beta$ . Pour chaque étape d'affinement, le diagramme de Ramachandran est calculé afin de vérifier la cohérence du nouveau modèle généré.



**Figure 5 : A. Localisation des angles  $\Phi$  (N-C $\alpha$ -C) et  $\psi$  (C $\beta$ -C $\alpha$ -C) et B. diagramme de Ramachandran.**

Le diagramme indique une distribution des valeurs statistiques que peuvent adopter ces angles dans les structures protéiques [36]. Les lettres indiquent les valeurs des angles dans les structures secondaires hélice  $\alpha$  et feuillet  $\beta$ .

## Les paramètres de convergence

Le modèle construit dans la carte de densité électronique ne rend que partiellement compte des données expérimentales même s'il est proche de la structure réelle. L'affinement cristallographique vise donc à réduire l'écart entre les facteurs de structure observés et ceux calculés en affinant les paramètres du modèle. Le facteur de structure est calculé à partir de la transformée de Fourier du modèle atomique. Il est nommé facteur cristallographique R et

---

permet de mesurer la similitude entre la carte de densité atomique calculée à partir du modèle créé et celle issue directement des données expérimentales. Les valeurs de R vont de 0 quand un accord parfait entre données théoriques et les données calculées est trouvé, à 0,6, valeur obtenue pour un facteur entre des données expérimentales et une carte de densité électronique aléatoire (de résidus disposés aléatoirement). Pour une structure de protéine résolue à 2,5 Å, le facteur R idéal est 0,2. Cependant, malgré l'importance du facteur R, des valeurs faussement positives peuvent lui être attribuées dans certaines structures, pouvant induire les cristallographes en erreur [37]. En effet, son calcul est fortement influencé par la modification des paramètres pendant l'affinement de la structure.

De ce fait, un autre critère, le facteur  $R_{\text{libre}}$  ( $R_{\text{free}}$ ) a été créé [38]. Il est calculé de la même manière que le facteur cristallographique mais il est basé sur un jeu de réflexions exclues de l'affinement. Il est donc neutre et permet une réelle évaluation de la qualité du modèle créé. Le facteur  $R_{\text{libre}}$  permet de suivre d'une manière objective la convergence vers une structure de meilleure qualité.

### *iii. Vibration et désordre*

Les paramètres détaillés dans le paragraphe précédent permettent une évaluation statistique de la qualité d'une structure résolue. Les raisons physiques justifiant une incertitude sur la position de certains atomes sont les mouvements thermiques et le facteur d'ordre. Les mouvements thermiques font référence à la vibration d'un atome autour de sa position moyenne. Le facteur d'ordre traduit le fait que certains atomes ou groupes d'atomes ne sont pas positionnés au même endroit dans toutes les mailles du cristal. Ces deux valeurs sont représentées par le facteur de température B qui est calculé pendant la phase d'affinement du cristal pour chaque atome de la structure. Le facteur B apporte des indices importants sur la dynamique des protéines. En effet, la distribution des valeurs le long de la séquence protéique est considérée comme un indicateur important de la structure de la protéine, reflétant sa flexibilité et sa dynamique. Des valeurs élevées indiquent une forte mobilité des atomes ainsi que des chaînes latérales [39]. Les résidus catalytiques ont tendance à avoir des facteurs B inférieurs à ceux des autres résidus suggérant qu'ils sont plus stables que les autres. Les résidus catalytiques des enzymes non liés à un ligand ou à un co-facteur ont les facteurs B légèrement supérieurs aux résidus évoqués précédemment laissant penser que ces résidus sont plus stables lorsqu'ils sont liés [40]. Ce genre d'observation a mené au développement de nombreuses applications basées sur l'étude des facteurs B, telles que la prédiction de la flexibilité des protéines [41], l'étude de la stabilité thermique des protéines,

---

l'analyse des sites actifs ou encore l'étude de la corrélation entre la mobilité des chaînes latérales et la conformation de la protéine. Il convient de noter que des artefacts expérimentaux peuvent biaiser ces valeurs comme l'effet du *packing* cristallin, ou l'existence de molécules tel le glycérol qui rigidifie des régions flexibles [42].

Ces indicateurs globaux ou atomiques sur la qualité des structures des protéines ne doivent pas servir à mettre en doute l'ensemble des données de la PDB, mais ils doivent permettre d'interpréter la précision des coordonnées 3D disponibles. Ces différents paragraphes illustrent simplement pourquoi il est indispensable de prendre du recul par rapport aux données et donc indispensable de vérifier la qualité d'une structure avant de commencer son étude, d'analyser avec précision ses chaînes latérales pour éventuellement comprendre les mécanismes auxquels elles participent... Au travers des différents critères décrits mais aussi en visualisant la structure protéique pour observer la cohérence de la structure, il est aussi fréquent que certaines parties de la protéine soient incomplètes. La majorité des outils de visualisation, tel PyMol [43], permettent une coloration des atomes en fonction de leur facteur B.

## **B. Classification des structures**

Le repliement 3D d'une protéine est propriété remarquable qui permet à la macromolécule d'adopter les différentes conformations 3D accessibles au sein d'un espace conformationnel de grande dimension. En structurant ainsi la protéine, le repliement 3D va conditionner la fonction de la protéine. Cette remarque peut laisser supposer que si les repliements de deux protéines se ressemblent, elles peuvent exercer les mêmes fonctions. Cependant il existe plusieurs degrés de similitude entre les structures des protéines. En effet, la quasi-totalité des protéines partagent des similitudes de repliements structuraux avec d'autres protéines [44]. Elles ont souvent une origine commune qui n'implique pas nécessairement une analogie de fonction (cf. partie I.B.3. « Applications et limites »). Hiérarchiser les structures existantes à un impact fort sur la biologie structurale. Savoir si une architecture protéique apparaît plus fréquemment qu'une autre permet aussi d'aider les cristallographes dans l'analyse des nouvelles structures. De plus, cette organisation permet d'ordonner les structures des protéines existantes stockées dans la PDB. Les méthodes de classification semi-automatique ou automatique de structures de protéines ont d'ailleurs permis d'organiser la PDB dont la taille devenait trop importante pour avoir une vision

---

d'ensemble de toutes les structures répertoriées [45]. Il en existe plusieurs et chacune utilise une méthode de comparaison particulière. Avant de s'intéresser aux classifications existantes, nous allons répertorier quelques méthodes de comparaison de structures protéiques.

## 1. Méthodes de comparaison de structures

### i. Généralités

Les objectifs des méthodes de comparaisons structurales sont le calcul d'une mesure quantitative de similitudes entre deux structures protéiques. Les quatre points importants sont [46]:

- **La représentation des structures** : La représentation peut s'effectuer à plusieurs niveaux, du plus fin au plus grossier : les atomes, les résidus (souvent représentés uniquement par les  $C\alpha$ ), les structures secondaires (comme les hélices  $\alpha$  ou les feuillettes  $\beta$  nommés *Structural Secondary Element* (SSE)) ou d'autres fragments structuraux, et la forme générale de la molécule. La comparaison de structure de protéines est toujours initiée d'un point de vue géométrique, par exemple par comparaison d'un ensemble de points ou de vecteurs. La représentation géométrique ou spatiale de ces points ou vecteurs est assez constante mais d'autres informations peuvent être ajoutées: l'ordre de ces éléments dans la séquence protéique, les propriétés physico-chimiques des acides aminés, leur type, leur charge, leur accessibilité au solvant, les liaisons hydrogène ou autres... Les liaisons hydrogène, ou toute autre propriété impliquant deux acides aminés, sont dites *relationnelles*. Au lieu de classer les informations en géométriques et non géométriques, il est possible de les classer en fonction de leurs propriétés individuelles (coordonnées cartésiennes, charge...), et leurs propriétés relationnelles (distances internes, liaisons chimiques...)
- **La mesure de similitude ou de dissimilitude** : Elles sont totalement dépendantes de la représentation.
- **L'algorithme de comparaison** : Les méthodes de comparaison de structures sont essentiellement géométriques. Ainsi comparer deux structures peut se

---

résumer à la comparaison de plusieurs éléments géométriques (points, vecteurs, courbes...) ordonnés ou non. Si la correspondance n'est pas donnée, le problème est le plus souvent NP-complet [47]. Il convient donc de choisir les correspondances avec la méthode et/ou avec la mesure de similarité. Des algorithmes de comparaison d'ensembles de points non-dédiés ont été adaptés pour la comparaison structurale comme la programmation dynamique, certains algorithmes de la théorie des graphes (recherche de cliques maximales, recherche de graphes isomorphes), les algorithmes d'optimisation/stochastiques (Monte Carlo, recuit simulé, algorithmes génétiques) ou certains algorithmes de classification.

- **Les post-traitements** : Un score final est calculé pour mesurer la similarité des deux structures comparées. Ce score peut servir à quantifier les similitudes entre les deux structures étudiées. Généralement, les méthodes de recherche de similarités structurales se font entre une structure dite « requête » et toutes les structures présentes dans une banque de structures. Ces dernières sont le plus souvent pré-compilées dans un format propre à la méthode utilisée. Le score sert alors au classement final des structures cibles de la banque selon leur degré de similarité avec la requête.

## ii. *Quelques méthodes de comparaison structurales*

Pour comparer deux structures à l'aide d'une description choisie, il faut effectuer une transformation rigide d'une structure sur l'autre. Ce type de comparaison est dit « externe » par rapport aux descriptions en coordonnées « internes » où les deux structures peuvent être comparées directement. Différentes méthodes de comparaison structurale basées sur divers niveaux de représentation de la protéine existent. Certaines utilisent des modèles « tout atomes » mais sont limitées à des petites sous-structures [48]. D'autres se basent sur les caractéristiques du squelette protéique décrites par les coordonnées cartésiennes des C $\alpha$  comme les distances internes ou les angles dièdres ( $\Phi$  et  $\psi$ ). La description des protéines en structures secondaires est aussi utilisée dans certaines méthodes.

---

### Méthodes basées sur les distances internes

Ces méthodes permettent d'éviter l'étape de superposition. Les structures sont comparées au niveau peptidique en ne prenant en compte que les C $\alpha$ . Pour décrire ces atomes, les distances internes, distances entre les atomes d'une même structure sont calculées. Il y a alors  $N^2/2$  descripteurs pour chaque structure qui sont souvent présentés sous la forme d'une matrice (symétrique) dite de distances internes,  $N$  étant le nombre d'atomes. DALI [49] et CE [50] utilisent le même principe de recherche de petits fragments similaires (AFP : *Aligned Fragment Pair*), puis de la meilleure série d'AFP. Les AFP sont ensuite assemblées particulièrement pour chaque méthode.

**DALI** utilise des « hexa matrices » issues de la division des matrices de distances des deux structures comparées en plus petites sous-matrices de taille 8 x 8. Les protéines sont chacune représentée par  $N^2$  « hexa matrices » qui sont filtrées afin d'éliminer celles représentant des SSEs redondantes. Des matrices similaires sont recherchées dans les deux structures et les 40 000 meilleures paires sont conservées. La deuxième étape est une étape d'assemblage. Elle est effectuée à l'aide de la technique de Monte Carlo et permet d'obtenir le meilleur alignement global. La correspondance finale entre résidus est ensuite affinée. DALI est utilisé dans la classification FSSP [51] qui sera abordée plus loin.

**CE** pour *Combinatorial Extension* [50] est un programme utilisant principalement les distances internes entre C $\alpha$ . La première étape consiste à rechercher de petits fragments de 8 résidus similaires dont la moyenne des différences des distances internes doit être inférieure à 3 Å. Les paires de fragments trouvées sont des AFP. L'étape d'assemblage consiste à ajouter une AFP à la série déjà construite en suivant certaines règles.

- (1) La nouvelle AFP n'est chevauchante avec aucune de celles déjà présentes.
- (2) La nouvelle AFP est contiguë avec une des AFPs de la série sur au moins une des deux protéines.
- (3) S'il y a un *gap*, il doit être de longueur inférieure à 30 résidus.
- (4) La distance moyenne entre toutes les paires d'AFP (nouvelle AFP et AFP de la série) doit être inférieure à 4 Å. Si toutes ces conditions sont respectées pour plusieurs AFP, seule la meilleure est ajoutée. Plusieurs alignements sont initiés avec toutes les AFP dont les distances internes sont en moyenne inférieures à 3 Å. Parmi les 20 meilleurs alignements finaux obtenus, seul celui ayant le meilleur RMSD est conservé. Il est ensuite affiné par une



---

procédure classique de superposition/alignement utilisant la programmation dynamique NWS (Needleman-Wunsch-Sellers) [52].

SSAP [53] combine la programmation dynamique et les distances internes. Ce programme est utilisé pour établir la classification CATH [45] qui est décrite dans la partie I.ii.

#### Méthodes basées sur les angles internes

Les angles utilisés pour la comparaison de structures protéiques sont les angles dièdres ( $\Phi$ ,  $\psi$ ) [54] définis Figure 5 et les angles ( $\alpha$ ,  $\tau$ ) [55]. Un angle  $\alpha$  est l'angle dièdre entre quatre C $\alpha$  consécutifs alors que l'angle  $\tau$  est l'angle de la liaison de trois C $\alpha$  consécutifs.

**YAKUSA** [56] utilise les angles  $\alpha$  pour décrire les protéines. Les angles  $\alpha$  sont discrétisés en classes selon une maille de  $10^\circ$ , 36 classes existent, chacune représentée par un symbole (un nombre entier). Une structure est donc représentée par une liste de symboles sur laquelle les algorithmes de recherche de motifs classiques peuvent être appliqués. YAKUSA permet de comparer une structure requête à une base de structures protéiques. Le principe général de la méthode est d'abord de rechercher tous les petits motifs de taille fixe et communs à la structure requête et aux autres structures de la banque. La suite de symboles de la structure requête est découpée en motifs chevauchants de taille fixe qui sont rangés dans un automate, avec leur position dans la structure (numéro du résidu). Cet automate contient non seulement tous les motifs présents dans la requête, mais aussi tous ceux non présents sur la requête qui leur sont proches, tels que ceux dont les symboles représentent des angles voisins. Ensuite, pour chaque structure de la banque, la recherche de similarités structurales locales se déroule en trois étapes:

- (1) Recherche de motifs communs entre les deux structures (graines).
- (2) Sélection des graines et extension en segments structuraux les plus long possibles : les SHSP (*Structural High Scoring Pairs*).
- (3) Sélection des SHSP et calcul d'un score global.

#### Méthodes basées sur les SSE

Les structures secondaires sont des structures régulières très fréquentes dans les protéines. Elles sont considérées comme les fragments structuraux les mieux conservés dans les structures similaires ou homologues. Ce niveau de représentation engendre moins

---

d'éléments (de l'ordre d'une dizaine de SSE pour une protéine de 300 résidus), les comparaisons de structures sont plus rapides que les précédentes. Cependant, les résultats sont aussi moins précis compte tenu du fait que de nombreuses régions des protéines sont ignorées. Les structures secondaires prises en compte par ces méthodes sont généralement exclusivement les hélices  $\alpha$  et les feuillets  $\beta$ . Plusieurs méthodes d'attribution des SSE existent comme DSSP [57]. Le mode de représentation le plus fréquent pour une SSE est un vecteur.

**SSM** [58] utilise les graphes pour représenter les structures des protéines. Cette méthode fonctionne en quatre étapes.

(1) Détection des structures secondaires par le programme PROMOTIF [59] (qui est basée sur un assignement similaire à DSSP).

(2) Construction d'un graphe pour chaque structure ayant pour sommets les SSE. Les arêtes sont définies par les paramètres d'angle et de distance entre les vecteurs définissant les SSE dans une même structure.

(3) Comparaison des graphes en utilisant la recherche des sous-graphes isomorphes.

(4) Retour au niveau peptidique et alignement résidu par résidu.

Les programmes VAST [60] et GRATH [61] utilisent aussi des graphes de SSEs pour comparer les structures des protéines.

**LOCK** [62,63] utilise la programmation dynamique adaptée à la comparaison de structures selon leur SSE. Les structures secondaires sont détectées par DSSP et sept paramètres sont calculés : cinq entre les paires de SSE d'une même structure (angles entre deux vecteurs d'une même protéine, distances internes...) et deux entre deux vecteurs de protéines différentes (distance et angle entre les vecteurs des deux protéines). Une superposition préalable est donc nécessaire. La première étape de l'alignement est de trouver les paires de SSE qui sont compatibles, par exemple celles dont la somme des 5 scores indépendants de l'orientation est supérieure à un seuil et dont au moins deux SSE sont contigus sur une des deux protéines (sans *gap*). À partir de ces quatre SSE, les deux structures sont superposées et les deux paramètres manquants sont calculés pour toutes les paires de SSE (y compris les paires de SSE de structures différentes). La programmation dynamique est alors utilisée pour trouver le meilleur alignement, le score dépendant des deux paramètres calculés après superposition. Pour le meilleur alignement, le RMSD est minimisé et l'alignement est affiné selon la méthode classique de superposition-alignement. Le

---

programme FOLDMINER permet de construire des « profils » à partir de ces alignements de paires de structures, profils similaires à ceux portant sur les séquences [63,64]. Ces profils décrivent la conservation des éléments de structure secondaire dans plusieurs structures similaires.

**TOP** [65] ressemble plus à une méthode itérative de superposition/alignement. Les éléments de structures secondaires sont détectés par DSSP et deux points, situés aux extrémités N et C terminales, sont calculés. Les paires de SSE similaires entre les deux protéines (deux SSE par protéine) sont recherchées en superposant les quatre points définis. Deux paires de SSE sont retenues si les angles entre SSE sont similaires et si la distance minimale entre les SSE est faible. Suit alors une étape de superposition/alignement où sont itérativement ajoutés les SSE proches. L'alignement au niveau des C $\alpha$  est ensuite effectué grâce aussi à une méthode itérative de superposition/alignement par ajout des C $\alpha$  proches (et respectant certaines contraintes, par exemple concernant la direction de la liaison C $\alpha$ -C $\beta$ ). Le score de similarité final est  $RMSD/(N_{match}/N_0)$  où  $N_{match}$  est le nombre de résidus en correspondance et  $N_0$  la moyenne des longueurs des protéines, la plus petite longueur, ou encore la longueur de la structure requête dans le cas d'une comparaison avec une banque.

#### Méthodes basées sur les coordonnées du squelette protéique

Ces approches se basent sur les coordonnées cartésiennes des structures protéiques et utilisent le RMSDc. Il s'agit de la racine carrée de la moyenne des distances entre les atomes mis en correspondance dans les deux structures décrites par leurs coordonnées cartésiennes. Elles sont généralement restreintes à la comparaison des C $\alpha$ . Dans ces méthodes de comparaison globale, l'objectif est d'avoir le plus de C $\alpha$  en correspondance avec un RMSD le plus faible possible.

**WHATIF** [66] utilise une méthode qui cherche des AFP de taille fixe (entre 10 et 15 résidus) dont le RMSDc (décrit par leurs coordonnées cartésiennes) n'est calculé que si les distances internes entre le premier C $\alpha$  et les 5 derniers C $\alpha$  sont les mêmes dans les deux fragments. Si le RMSDc est assez faible, l'AFP est étendue aux C $\alpha$  voisins tant que le RMSDc reste en dessous d'un certain seuil. Les ensembles d'AFPs sont ensuite générés de la manière suivante : les deux structures sont superposées selon l'ensemble d'AFPs courant et une AFP n'est ajoutée que si les centres de gravité de ses deux fragments ne sont pas trop éloignés. La notion de séquence des AFPs n'est pas toujours observée, une correspondance

---

entre AFPs est donc obtenue. La superposition globale des deux structures est alors recalculée pour cet ensemble d'AFPs en considérant les nouvelles correspondances entre C $\alpha$ . Finalement, seule la correspondance la plus longue est conservée et affinée. Elle permet de calculer une distance entre les structures comparées.

### Méthodes de comparaison flexibles

Ces méthodes prennent en compte la flexibilité des protéines lors de l'étape d'assemblage. Leur concept est que l'alignement de deux protéines se décompose en plusieurs segments rigides joints par des zones flexibles.

**FLEXPROT** [67,68] détermine les AFP par rapport au RMSDc, ainsi que plusieurs étapes itératives de superposition/ajout des C $\alpha$  contigus. Le but est d'étendre au maximum les AFPs dont le RMSDc ne doit pas dépasser 3 Å. La méthode d'assemblage utilisée est inspirée du logiciel FASTA [69] et nécessite la construction d'un graphe orienté acyclique. Toutes les AFPs sont des nœuds du graphe et un arc est défini entre deux AFPs si celles-ci peuvent se succéder dans l'alignement final. La notion de séquence est donc conservée. Un poids qui pénalise les gaps est ensuite associé aux arcs. Enfin, un processus de classification des AFPs est utilisé pour rechercher tous les chemins de ce graphe allant d'un bout à l'autre des structures comparées. Les alignements disjoints sont regroupés en un seul alignement si la topologie de la protéine le permet.

**FATCAT** [70] utilise à la fois le RMSDc et les coordonnées internes. Les paires d'AFPs considérées sont de longueur 8 et leur RMSDc doit être inférieur à 3 Å. La méthode consiste à combiner les AFPs en prenant en compte les *gaps* et les mésappariements. La meilleure série d'AFPs est recherchée par programmation dynamique. Pour que deux paires d'AFPs soient compatibles, le RMSD entre les matrices de distances des résidus dans les fragments de chaque protéine formant les AFPs connectés. Si la similitude est grande, les AFPs sont compatibles, une torsion est introduite afin de connecter les paires d'AFPs. Le nombre maximal de torsions est fixé avant l'alignement. S'il vaut zéro, la méthode se comporte comme une approche par comparaison rigide. Le score d'une AFP dépend de son RMSD et de sa taille. Le meilleur enchaînement d'AFPs est recherché, puis différents post-traitements sont appliqués. Des torsions peuvent être introduites si elles ont pour conséquence une baisse du RMSD global, d'autres peuvent être enlevées.

---

### Méthodes basées sur les alphabets structuraux

Les structures secondaires (SSE) considérées jusqu'à présent ne décrivent qu'une partie des structures protéiques. En effet, seuls les hélices  $\alpha$  et les feuillets  $\beta$  sont assignés. Les boucles constituent un groupe non homogène de structures plus ou moins irrégulières, il est donc difficile de les définir systématiquement. Les alphabets structuraux permettent de décrire simplement et avec une grande précision l'ensemble des structures tridimensionnelles [71,72]. Les blocs structuraux sont le plus souvent de taille fixe (quelques résidus) et sont chevauchants. Ils sont classés en comparant entre elles toutes les sous-structures d'une taille donnée d'une banque non redondante de structures. Les classes de blocs représentant le mieux l'espace des structures sont choisies pour constituer un alphabet. Un symbole est très souvent affecté à chaque classe de blocs. Certains programmes utilisent les alphabets structuraux pour comparer les structures 3D de protéines, comme par exemple, 3D-Blast [73] ou PB-ALIGN [72,74,75 ].

**PB-ALIGN** [72] permet de transcrire une structure protéique en une combinaison de 16 blocs protéiques (BPs) représentés par des lettres ( $a \rightarrow p$ ) (cf. Figure 6). Ces séquences peuvent ensuite être comparées en utilisant une matrice de substitution de blocs protéiques couplée à une programmation dynamique. Cette méthode est simple et rapide, elle est applicable à d'importants jeux de données. Le Tableau 1 est une synthèse de treize méthodes récentes de comparaison structurale de protéines sur un jeu de données type comprenant des protéines particulièrement difficiles à superposer et donc à retrouver. Pour chaque méthode, les protéines des différents ensembles servent de requête et sont recherchées dans la PDB. Ce tableau est extrait de [72]. Il montre que les méthodes récentes (PB-ALIGN et YAKUSA) fonctionnent particulièrement bien, et que PB-ALIGN, malgré sa simplicité fonctionne au moins aussi bien si ce n'est mieux que toutes les autres méthodes disponibles actuellement. En effet, 96.6 % des protéines requêtes utilisées sont retrouvées dans SCOP.



TOPSCAN	15	12	9	7	70
TOPS	2	15	14	7	62
PRIDE	14	14	7	3	62
LOCK	0	14	11	8	54
SSM	5	13	10	5	54

**Tableau 1 : Comparaison de 13 méthodes récentes de comparaison et détermination de structure de protéines.** Chaque méthode est testée sur quatre ensembles de protéines de classes SCOP différentes. Une protéine requête est soumise pour chaque méthode. Ces quantités correspondent au nombre de fois où la protéine requête est retrouvée.

## 2. Les bases de données existantes

La PDB [5] est la base de données de structures protéiques la plus utilisée au monde. Elle répertorie la majorité des structures résolues et rendues publiques à ce jour et met à la disposition de la communauté scientifique leurs coordonnées ainsi que des annotations sur leur structure, leur fonction... Ces informations sont extraites d'autres bases de données qui servent à hiérarchiser les structures des protéines. Les structures sont généralement catégorisées en quatre classes principales en fonction de leur composition en structure secondaire: « tout  $\alpha$  », « tout  $\beta$  », «  $\alpha / \beta$  » ( $\alpha$  et  $\beta$  mélangé) ou «  $\alpha + \beta$  » ( $\alpha$  d'un côté et  $\beta$  de l'autre côté). Les classifications les plus utilisées sont SCOP (Structural Classification Of Proteins) [44] et CATH (Class Architecture Topology Homology) [45]. Ces classifications sont des classifications hiérarchiques des domaines structuraux protéiques. Les structures sont découpées en domaines, puis regroupées selon leurs similarités de séquences, puis sur leurs similarités de structures, et enfin sur leur similarité en composition et organisation des structures secondaires. Des classifications non hiérarchiques existent aussi telles que FSSP [51] construite avec le programme DALI [49] ou CE Database, construite avec le programme CE [50].

### i. SCOP

La classification **SCOP** [44] est construite manuellement d'après des informations structurales et des connaissances plus générales sur chaque protéine. Les outils automatiques de comparaison structurale ne sont utilisés que pour accélérer la classification par inspection visuelle. Dans un premier temps, les structures protéiques sont découpées en domaines, avant d'être classées. Les quatre niveaux de classification du niveau le plus général au plus fin, sont:

- *class* : la composition en structures secondaires est similaire. Il y a quatre classes principales définies par M. Levitt et C. Chothia [77]: « tout  $\alpha$  », « tout  $\beta$  », «  $\alpha / \beta$  » ( $\alpha$  et  $\beta$  mélangé) ou «  $\alpha + \beta$  » ( $\alpha$  d'un côté et  $\beta$  de l'autre côté) ainsi que les classes des protéines multidomaines, des protéines membranaires et des petites protéines.
- *fold* : la composition en structures secondaires (hélices  $\alpha$  et feuilletts  $\beta$ ), leur arrangement spatial et leurs connexions sont similaires.
- *superfamily* : ensemble de structures au sein desquelles l'identité de séquence peut être faible mais les structures et les fonctions suggèrent une origine évolutive commune.
- *family* : les structures protéiques ont au moins 30% d'identité de séquence, ou alors possèdent des fonctions et des structures très similaires.

La banque SCOP est une classification « manuelle » de domaines. Elle n'est pas souvent mise à jour. Elle comprend actuellement 68% des structures de la PDB. Cette banque comprend par exemple la classe *Bergerat fold* [78] qui contient des protéines de fonction différentes telles que des gyrases, des histidines kinases, des mutL, ayant très peu d'identité de séquences entre elles mais ayant un *fold* commun; nous retrouverons ce repliement plus tard dans le manuscrit. La dernière mise à jour de SCOP date de novembre 2007 et comprenait 1086 *folds*, 1777 superfamilles et 3464 familles.

## ii. CATH

La classification **CATH** [45,79] est effectuée à la fois automatiquement et manuellement. Comme SCOP, elle est hiérarchique et subdivisée en quatre niveaux principaux. Elle possède trois niveaux supplémentaires de classification établis sur la similarité des séquences protéiques (cf. Figure 7). Les niveaux de classification sont :

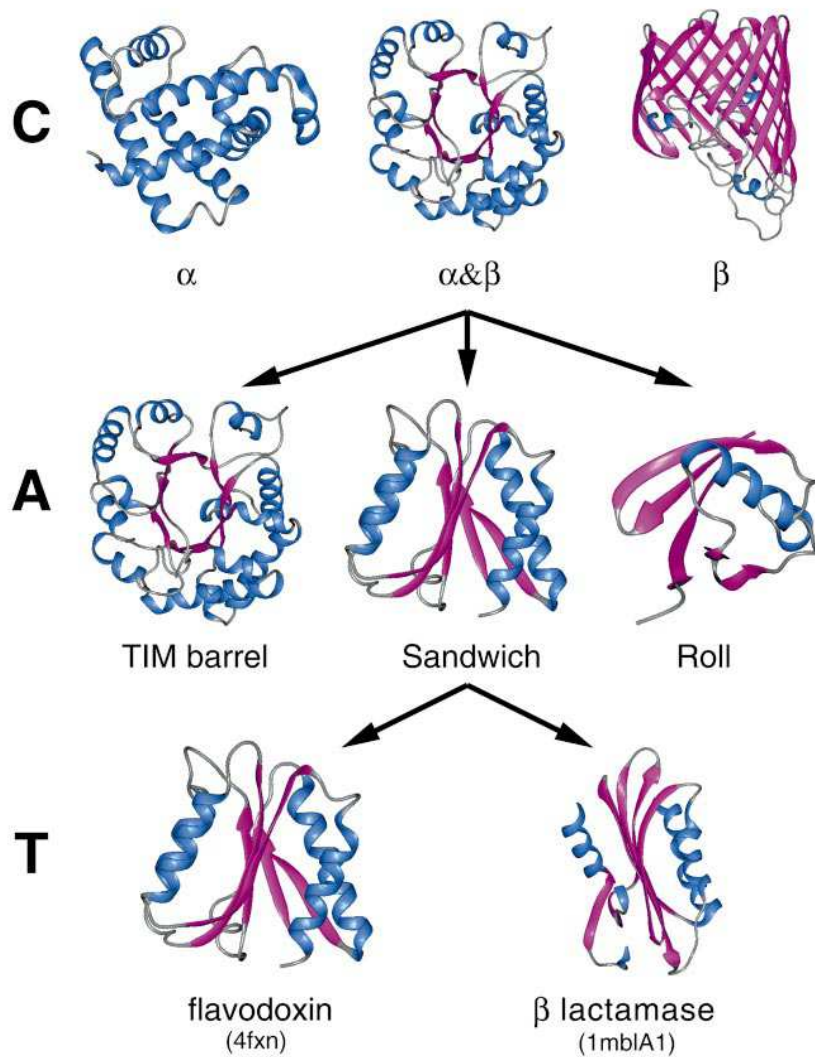
- *class* : les structures sont regroupées selon leur composition en structures secondaires et les contacts entre celles-ci. Il y a quatre classes : « mainly  $\alpha$  » et « mainly  $\beta$  » qui sont similaires aux deux classes « tout  $\alpha$  » et « tout  $\beta$  » de SCOP, « mixed  $\alpha$ - $\beta$  » et « Few secondary structures ». L'assignation d'une structure à l'une de ces quatre classes est automatique dans 90% des cas (les 10% restant sont assignés manuellement) [80].



- 
- *architecture* : les structures sont regroupées selon l'organisation générale de leurs structures secondaires. Dans une même classe, l'architecture est la même (cf. Figure 7).
  - *topology* : les structures ayant un même repliement en termes de nombre, d'ordre et de connexions de structures secondaires sont regroupées. La méthode de comparaison utilisée est SSAP [53] et implique une contrainte sur la longueur de l'alignement et le score obtenu.
  - *homologous surperfamily*: les structures et les fonctions d'un même groupe sont très similaires, suggérant un ancêtre commun. SSAP est aussi utilisé.

Des niveaux supplémentaires S, N, I regroupent les structures ayant une identité de séquence respectivement > 35%, > 95% et de 100% (ce dernier niveau regroupe les protéines qui ont été résolues plusieurs fois, par exemple complexées ou non avec leur ligand). L'algorithme d'alignement des séquences est celui de Needleman et Wunsch [52].

Contrairement à SCOP, CATH est régulièrement mise à jour [79,81]. Les algorithmes de comparaison de détection des domaines sont aussi régulièrement améliorés. Par exemple, la méthode CATHEDRAL [82] a été introduite en 2007, elle se base sur un protocole itératif qui localise les repliements protéiques déjà connus dans des nouvelles protéines multi-domaines. Des informations sur le degré de diversité des structures présentes dans CATH sont accessibles, ainsi que d'autres informations concernant les chevauchements structuraux possibles entre superfamilles peuvent aussi être localisées.



**Figure 7 : exemple de classification hiérarchique présente dans les différents niveaux de CATH.**

Figure extraite de l'article d'Orengo et collaborateurs [45]. Les deux structures au bas du schéma sont de même classe ( $\alpha$  et  $\beta$ ), même architecture (sandwich  $\beta$ : des hélices  $\alpha$  sont présentes de chaque côté de plusieurs feuillets  $\beta$ ), mais de topologies différentes (Rossmann fold et Beta-lactamase).

### iii. FSSP

La classification « *Families of Structurally Similar Proteins* » (FSSP) est une classification non hiérarchique [46,51]. Elle a été construite à partir des alignements de DALI [49]. La méthode consiste en un découpage des structures en domaines puis en une comparaison exhaustive deux à deux avec l'algorithme DALI. Un algorithme de classification ascendante hiérarchique fondé sur les scores des alignements de DALI « *average linkage clustering* » permet ensuite de regrouper les domaines en familles. N'étant plus mise à jour régulièrement et ne disposant pas de liens avec d'autres banques de données ni de moyen de

---

visualiser les alignements en 3D (seuls les alignements des paires de séquences sont fournis), cette base de données est de moins en moins utilisée.

### 3. Applications et limites

De nombreuses méthodes de comparaison structurale ont été développées depuis une vingtaine d'années. Ces méthodes permettent de classer les structures protéiques de la PDB suivant leur repliement. Des serveurs web sont désormais disponibles pour permettre de caractériser rapidement une nouvelle protéine par rapport aux structures existantes. Ainsi, la plupart des structures issues des consortiums de génomique structurale sont annotées. Leur classe CATH est définie alors que leur fonction peut ne pas l'être. Par exemple, les structures 2Q78 et 2NR4 de la PDB sont issues des deux consortiums : le « New York Structural Genomic Research Consortium » (NYSGRG, [www.nysgrg.org](http://www.nysgrg.org)) et le « Joint Center for Structural Genomics » (JCSG, [www.jcsg.org](http://www.jcsg.org)) et sont annotées comme protéine hypothétique. Leurs caractéristiques CATH sont:

- pour 2Q78: Class : « Alpha Beta » ; architecture : « Roll » ; topology : « Thiol Ester Dehydrase ».
- pour 2NR4: Class : « Mainly Alpha » ; architecture : « Up-down Bundle » ; topology : « Methane Monooxygenase Hydroxylase » ; Homology: « Hypothetical membrane protein ta0354\_69\_121 ».

L'assertion « le repliement d'une protéine est fortement corrélé à sa fonction » a longtemps laissé supposer qu'une similarité structurale impliquait une similarité de fonction. Cependant, il est complexe de prévoir l'évolution des fonctions des protéines au sein d'une même superfamille car des protéines dites homologues ont parfois des fonctions différentes [83]. Par exemple, les protéines de la superfamille *HUP-domain* illustrent la diversité de fonctions et de structures de protéines homologues et regroupent les fonctions *electron transfer flavoprotein* ou *deoxyribodi-pyrimidine photo-lyases* qui sont très différentes [84]. Le repliement TIM-Barrel est aussi reconnu pour sa versatilité en terme de fonction [85]. Il est composé d'une alternance de huit hélices alpha et huit feuilletts beta. Un minimum de 200 résidus est nécessaire pour former le repliement TIM-BARREL dont 160 sont considérés équivalent structurellement pour toutes les protéines de cette classe. Pourtant, ces enzymes sont impliqués dans au moins 15 fonctions différentes, par exemple la xylose isomérase (code PDB : 1DXI), l'aldose réductase (oxydoréductase, code PDB 2ACS) ou l'adénosine

---

déaminase (hydrolase, code PDB 1FKW) [86]. Les divergences entre les séquences de cette famille se trouvent dans les boucles qui relient les structures secondaires, ainsi que dans la région C terminale qui contient la plupart du temps le site actif de la protéine.

Les méthodes de comparaison structurale ne sont pas suffisantes pour émettre de solides hypothèses quant aux fonctions potentielles de protéines. Pourtant, la structure d'une protéine n'en est pas moins importante. En effet, le repliement global d'une protéine a pour rôle de stabiliser sa structure alors que sa fonction dépend d'un petit ensemble de résidus conservés souvent situés à la surface de la protéine [40]. La notion de surface d'interaction est importante et la nécessité de pouvoir comparer des surfaces devient donc cruciale pour l'annotation fonctionnelle structurale.

### **C. Notion de surface d'interaction**

La fonction d'une protéine ne s'exprime qu'à partir de moment où elle interagit avec un ou plusieurs partenaires spécifiques. Ces interactions sont la base de tous les mécanismes biologiques. Elles n'impliquent généralement qu'une partie des résidus de la protéine. Cette zone est nommée la surface d'interaction. Les protéines interagissent avec des petites molécules et des macromolécules (d'autres protéines ou des acides nucléiques). En fonction du type de leur partenaire, les caractéristiques des interfaces se divisent en deux catégories: (i) les propriétés structurales: taille, forme et complémentarité géométrique, et (ii) les propriétés chimiques: potentiel de solvation, hydrophobicité, potentiel électrostatique, liaisons hydrogène et ponts salins. Les interactions entre la protéine et son partenaire sont généralement non covalentes et de faibles énergies. Elles sont de même type que celles impliquées dans le repliement des protéines (liaisons hydrogène, interactions électrostatiques, contacts de type Van der Waals et effets hydrophobes). En fonction du partenaire, une interface a des propriétés particulières. Les trois types d'interfaces sont décrits dans les parties suivantes.

## **1. Propriétés des différents types d'interfaces**

### *i. Interactions protéine-protéine*

Les interactions protéine-protéine sont au coeur de la nano-machinerie cellulaire. Elles sont essentielles par exemple dans la reconnaissance antigène-anticorps, ainsi que dans les voies de transduction du signal dans les cellules saines et tumorales, ce qui en fait des cibles

---

intéressantes pour le développement de molécules thérapeutiques. La taille des interfaces des complexes protéine-protéine est une des caractéristiques de ces interfaces. Elle est généralement mesurée par la différence entre l'aire de la surface accessible (ASA) du complexe et celle des composés séparés :

$$B = \text{ASA}_{\text{prot.1}} + \text{ASA}_{\text{prot.2}} - \text{ASA}_{\text{complexe}} \quad (1)$$

La mesure B donne une indication sur la force de liaison entre les deux protéines. L'aire de la surface enfouie est en relation directe avec l'énergie de l'interaction hydrophobe de désolvatation [87,88]. Pour caractériser les interfaces protéines-protéine, Lo Conte et ses collaborateurs ont analysé la structure atomique des sites de reconnaissance dans 75 complexes protéine-protéine: 24 complexes protéase/inhibiteur, 19 complexes anticorps/antigène et 32 autres complexes, dont 9 complexes enzyme/inhibiteur et 11 complexes impliqués dans la transduction du signal [89]. Sur les 75 complexes, 52 ont une interface de « taille standard » égale à  $1600 (\pm 400) \text{ \AA}^2$ . Ces complexes sont principalement des complexes anticorps/antigène et protéase/inhibiteur. La plus petite des interfaces est de  $1150 \text{ \AA}^2$  (le cytochrome peroxydase et son substrat, le cytochrome C). Les petites interfaces caractérisent le plus souvent des complexes temporaires et de basse stabilité. Vingt complexes ont des interfaces dont la taille s'étend de 2000 à  $4660 \text{ \AA}^2$ . Les complexes étudiés sont principalement de protéases complexées avec une classe particulière d'inhibiteurs, des complexes permanents, par exemple le complexe trombine/rhodniine. Par leur taille, les interfaces de ces complexes ressemblent à celles des sous-unités protéiques des assemblages oligomériques [90]. Toutefois, elles sont moins hydrophobes et contiennent une proportion plus importante de groupes chargés. Dans ces complexes, l'association implique de grands changements conformationnels, principalement un remodelage de la chaîne principale à l'interface.

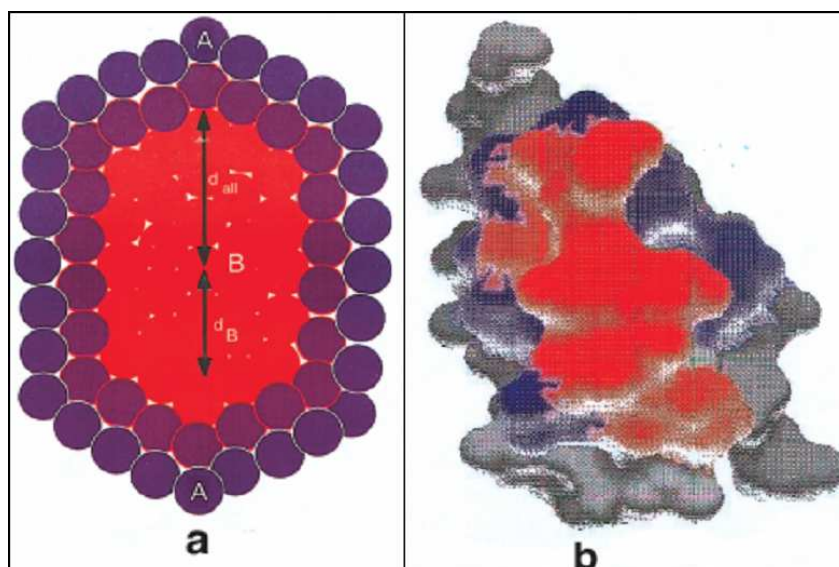
Structuralement, le site de liaison est plus souvent composé de feuillet  $\beta$  et pour les chaînes relativement longues non structurées. Les hélices  $\alpha$  sont moins présentes [91]. Les complexes anticorps/antigène et enzyme/inhibiteur présentent un compactage étroit («*close packing*») [88,92]. Les complexes homodimériques sont plus compacts que les complexes hétérodimériques. Parmi les atomes qui se retrouvent à l'interface après l'association, environ la moitié sont en contact avec la protéine partenaire, dont un tiers deviennent complètement inaccessible au solvant [89]. Les interfaces des complexes anticorps/antigène sont plus planes que celles des complexes enzyme/inhibiteur pour lesquels les résidus catalytiques sont

---

généralement localisés dans des crevasses de la surface. Pour les autres types de complexes, cette caractéristique est mixte [90].

Le degré de compactage et le nombre de liaisons hydrogènes sont largement proportionnels aux tailles des surfaces en interaction [93]. Les molécules d'eau liées ont surtout été observées aux interfaces des complexes anticorps/antigène [94]. Il est logique d'en trouver moins dans les interfaces enzyme-inhibiteur, car la complémentarité de surfaces de ces derniers complexes est meilleure, laissant peu de place aux molécules d'eau. Les interfaces des cristaux de protéines sont plus hydratées que celles des assemblages spécifiques, ce qui reflète une composition plus hydrophile des interfaces non spécifiques [95].

Une analyse détaillée de 70 complexes protéine-protéine a mis en évidence la présence de zones contiguës typiques pour l'interaction (« *patch* ») [96]. Ces patches sont caractérisés par une aire d'au moins  $800 \text{ \AA}^2$ , impliquant approximativement un peu plus de 20 résidus et un peu moins de 100 atomes. Pour les interfaces enfouies dont l'aire est supérieure à  $2000 \text{ \AA}^2$ , plusieurs parcelles peuvent être observées. Ces parcelles sont composées d'un cœur (« *core* ») et d'un bord (« *rim* ») (cf. Figure 8); le rôle du bord serait principalement d'isoler les résidus du cœur de la parcelle du solvant alors que le cœur assure le compactage du complexe. La composition en résidus du cœur et du bord des interfaces protéine-protéine est différente. En effet, le tryptophane, la tyrosine, l'arginine, ainsi que la leucine et la valine sont plus présents dans le cœur des interfaces alors que les résidus sérine et thréonine le sont moins. Ces observations ont été faites par Chakrabarti et ses collaborateurs [96]. Bogan et ses collaborateurs ont observés que le tryptophane, la tyrosine, l'arginine sont les résidus plus présents alors que les résidus sérine et thréonine sont moins présent et les résidus leucine et valine sont quasiment absent [97]. En conclusion, nous pouvons affirmer que la caractérisation des interfaces protéine-protéine est un travail extrêmement complexe [98].



**Figure 8 : Représentation des atomes enfouis et accessibles des interfaces.**

a) Schéma d'une interface protéine-protéine divisée en deux parties : le cœur (rouge) composé d'atomes enfouis et le bord (bleu) composé d'atomes exposés au solvant. b) Site de reconnaissance de l'inhibiteur du dichlore en complexe avec la subtilisine (code PDB:1SNI). La surface en rouge représente les atomes enfouis et la surface en bleu, les atomes accessibles. Figure extraite de [96].

## ii. *Interactions protéine-nucléotide*

La première caractéristique des complexes ADN-protéine est la complémentarité de surface entre ces macromolécules. Lors de la complexation, les surfaces complémentaires interagissent de façon intime pour devenir inaccessibles au solvant. Seules quelques molécules d'eau sont piégées dans le cas des surfaces protéiques polaires [99]. À l'interface, il est possible d'observer des interactions de type Van der Waals, des liaisons hydrogène, et des ponts salins dont les molécules d'eau sont souvent les intermédiaires. Les surfaces de contact sont étendues et varient entre 1120 et 5000 Å<sup>2</sup> [100]. L'ADN est un polyélectrolyte chargé et, par conséquent, ses interactions avec les protéines dépendent fortement de la concentration en sels dans le milieu. L'aspect thermodynamique qui renseigne sur la nature des énergies conduisant à l'interaction (enthalpie d'interaction et perte d'entropie associée) est difficile à relier aux observations issues de la cristallographie. De plus, les études tant théoriques qu'expérimentales portant sur la thermodynamique et sur l'électrostatique de l'association ne sont pas aisées [101,102].

Deux modes de reconnaissance ADN-protéine existent: (i) la « lecture directe » des bases de l'ADN par les chaînes latérales de la protéine et (ii) la « lecture indirecte » dépendant des propriétés structurales de l'ADN. La « lecture directe » implique la formation de liaisons hydrogène dites spécifiques entre les chaînes latérales et les sites donneurs ou accepteurs des bases de l'ADN, ceci le plus souvent via le grand sillon. A surface d'interaction égale, les

---

liaisons hydrogène sont beaucoup plus fréquentes dans les interactions ADN-protéine que dans les interactions protéine-protéine. Toutefois, la plupart des ces contacts ont lieu entre les acides aminés et le squelette phosphodiester [103] et représentent la « lecture indirecte » de l'ADN. Dans ce type de reconnaissance, les propriétés mécaniques de cet oligonucléotide jouent un rôle dans la spécificité de la complexation. De nombreuses structures cristallographiques présentent des déformations mineures ou majeures de l'ADN. La flexibilité intrinsèque de l'ADN, qui dépend de sa séquence, peut servir comme un code "additionnel" qui dirige les protéines vers un site de liaison particulier. Un exemple caractéristique est le complexe ADN-TBP [104] dans lequel l'ADN est courbé de 90° dans la direction du grand sillon, et où quasiment aucun contact spécifique entre l'ADN et la protéine ne permet de rendre compte de la préférence pour les sites "TATA". Les principales déformations de l'ADN rencontrées dans les complexes sont la sur-tension ou sous-torsion (*over/under twisting*), l'étirement (*stretching*) et la courbure (*bending*).

Les protéines se liant à l'ADN sont reconnaissables par quelques indices, comme les motifs structuraux ou la composition spécifique en résidus, mais surtout par leurs propriétés électrostatiques globales et locales (charges positives abondantes) [105-107]. Les interactions électrostatiques jouent un rôle très important dans la « lecture indirecte » de l'ADN. Dans le complexe impliquant le répresseur de l'opéron tryptophane, deux hélices du motif « hélice tour hélice » ont un potentiel électropositif, l'autre face de la protéine étant très fortement électronégative (la protéine possède une charge totale négative de -6 à pH 7). Cette différence de potentiel permettrait l'orientation de ce motif sur la protéine vers l'ADN avant la complexation [108].

### iii. *Interactions protéine–ligand*

Dans ce type d'interactions, un ligand est une molécule de petite taille (inférieure à 100 atomes) ou un petit peptide (chaîne protéique d'une dizaine de résidus). Les interactions protéine-ligand ont des rôles clé dans les fonctions de transport, de transmission de signaux cellulaires, ainsi que dans la capacité des scientifiques à moduler la fonction des protéines en élaborant des médicaments inhibiteurs compétitifs des ligands naturels (« *drug design* »). De plus en plus de structures de la PDB sont résolues avec des molécules avec lesquelles elles interagissent avec plus ou moins d'affinité. Ces complexes permettent l'extraction des caractéristiques structurales de ces interfaces et permettent d'établir des règles sur les modes de fixation des ligands aux protéines. Les sites de liaisons dépendent de la forme de la surface

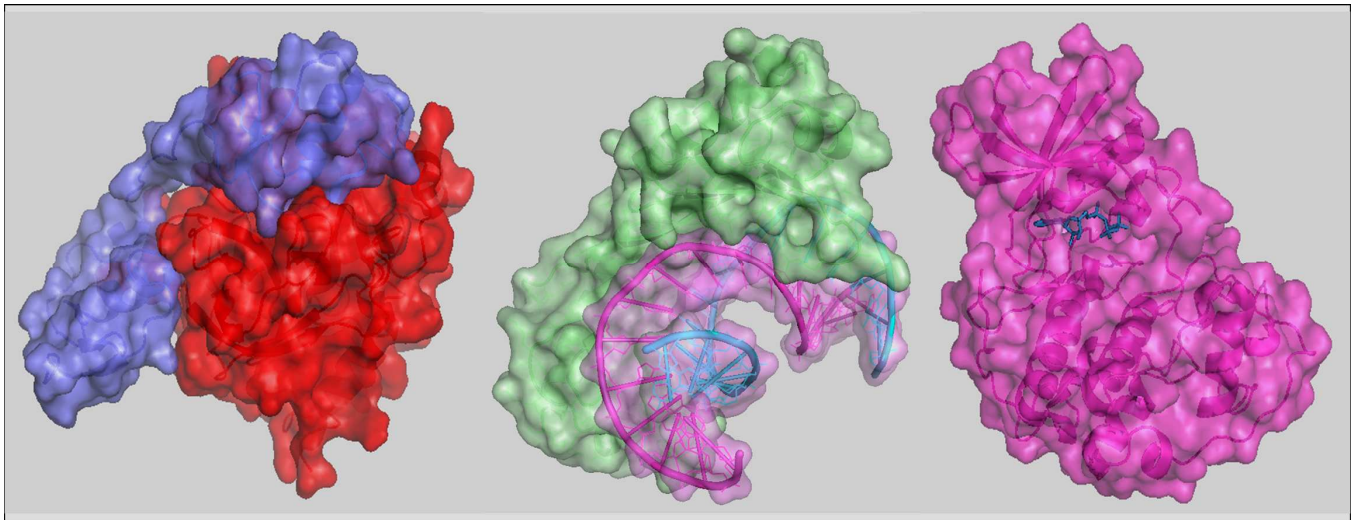


---

des protéines [109]. La surface protéique est le plus souvent irrégulière composée de plusieurs cavités et de sillons de tailles différentes. Ces cavités sont impliquées dans différents types d'interactions. Une cavité de grande taille offre une surface plus importante, il est postulé que cette propriété augmenterait les possibilités à la protéine d'interagir avec d'autres molécules et en particulier avec de petits ligands. Laskowski et ses collaborateurs ont montré que les ligands se fixent en général dans la plus grande cavité de la protéine étudiée [109]. A l'inverse, lorsque plusieurs cavités de tailles similaires sont détectées sur une surface, le ligand peut se fixer dans l'une ou l'autre des cavités, cette cavité n'étant pas forcément la plus grande. Les ligands allostériques induisent un changement de conformation de la protéine et se fixent le plus souvent dans les petites cavités [109].

Cette corrélation observée avec les dépressions surfaciques (cavités) est la conséquence des principes physiques qui gouvernent la reconnaissance moléculaire. Tant que les contributions entropiques de la flexibilité du ligand ne deviennent pas trop pénalisantes et dans les cas où la contribution des termes de solvation est constante. Il est possible d'admettre que l'affinité est principalement une contribution additive de termes enthalpiques entre les atomes du ligand et de la protéine. Elle est donc relativement liée à la taille de la surface d'interaction. Une spécificité correcte est plus aisément obtenue si l'environnement des protéines impose des contraintes géométriques et physicochimiques.

Bien qu'essentiels, les critères géométriques ne sont pas suffisants pour caractériser un site de liaison. En effet, ces régions de la surface des protéines doivent pouvoir interagir avec une molécule. Les interactions protéine-ligand sont soit covalentes pour les inhibiteurs irréversibles ou non covalentes pour les inhibiteurs compétitifs. Certaines propriétés physicochimiques à la surface des protéines sont donc détectables et servent à la localisation de sites de liaison potentiels. Des approches exploitent ces propriétés en calculant des cartes électrostatiques d'interaction entre différentes sondes (des groupements méthyles, hydroxyles, carbonyles...) et les atomes de la surface de la protéine (cf. partie I.C.2) [110]. Dans la partie suivante, je détaillerai ces propriétés et quelques unes de ces méthodes. Ces cartes permettent de localiser des sites de liaisons potentiels appelés « *hot-spots* ». D'autres approches se basent sur le fait que les résidus présents dans les sites de liaison sont mieux conservés que les autres [111]. L'analyse de l'évolution phylogénétique des séquences en acides aminés est aussi utilisée pour prédire la position des sites d'interaction. La Figure 9 représente les trois différents types de complexe.



**Figure 9 : Trois exemples de complexes protéiques décrits.**

A gauche, le complexe Thrombine-Rhodniine (code PDB : 1TBR), la thrombine est en rouge et la rhidniine en bleu.

Deux interfaces distinctes sont présentes. Au milieu, le complexe ADN-TBP [104] (code PDB : 1CDW), plus communément appelé boîte TATA. A droite, le complexe CDK2-ATP (code PDB : 1B38), le ligand ATP se trouve dans la cavité la plus grande de la surface.

## 2. Méthodes de localisation de sites de liaisons et annotation fonctionnelle

De plus en plus de structures sont résolues avec très peu d'information sur leur fonction. Aujourd'hui plus de 3000 structures de la PDB sont classées comme protéines hypothétiques (« unknown function »). Il faut aussi noter qu'un site unique de liaison par protéine est souvent caractérisé alors qu'une protéine interagit en moyenne avec trois à neuf partenaires [112]. Des méthodes d'annotations fonctionnelles de protéines permettent de localiser les résidus impliqués dans les mécanismes de liaison. Les descripteurs utilisés peuvent dépendre de la séquence (notion de signature ou de profil), d'autres de la structure (notion de surface d'interaction) et d'autres des groupements chimiques à la surface des protéines (notion de signature chimique 3D)

### i. Méthodes basées sur la séquence

L'approche classique d'annotation d'une séquence protéique est la recherche de séquences annotées ayant un taux d'identité élevé. La plupart des protéines hypothétiques le sont car les méthodes de comparaison de séquence comme PSI-BLAST [113] ou FASTA [114] ne détecte aucune similitudes avec des protéines annotées des bases de données de séquences tel UNIPROT [2]. Cependant, des relations plus lointaines entre séquences existent

---

et sont détectées à partir d'alignements multiples qui servent à la génération de séquences consensus des régions conservées nommées signatures ou motifs. Une signature n'est pas obligatoirement composée de résidus contigus. Elle est importante car les régions spécifiques impliquées dans la liaison avec d'autres molécules ou dans l'activité enzymatique de la protéine sont souvent des motifs conservés. Certaines familles ne peuvent cependant pas être détectées par recherche de motifs. En effet, cette recherche se base sur la présence stricte des acides aminés de la signature. Elle peut s'avérer infructueuse si certains résidus sont absent d'une séquence requête. Un profil est un tableau de poids et de pénalités sur la position d'acides aminés dans une séquence. Une recherche de profils permet de détecter des similarités plus éloignées qu'une recherche de signature, et cela s'avère être plus efficace [115]. Ces poids permettent le calcul d'un score de similarité entre un profil et une séquence. Un score seuil délimite l'appartenance d'une séquence à une famille caractérisée par le profile. L'augmentation de la taille des bases de données de séquences a incité à l'émergence de bases de données de motifs et de profils, ainsi que des méthodes automatique pour détecter leur présence. Leurs objectifs sont de catégoriser les banques de séquences en familles fonctionnelles de séquences bien définies et aussi de pouvoir attribuer une fonction aux séquences de protéines issues de consortium de séquençage à haut débit. Prosite [116], Pfam [117] ou Catalytic Site Atlas [118] sont des bases de données de domaines fonctionnels publiques: elles sont utilisables pour comparer une séquence requête à l'ensemble de leurs entrées.

**PROSITE** [116] est une base de données de motifs et de profils protéiques qui ont tous une signification biologique particulière. Les motifs sont représentés sous forme d'expressions régulières globales qui peuvent être associées à plusieurs signatures; par exemple l'entrée PDOC00100 est associé à l'annotation "*eukaryotic protein kinase*" et il rassemble quatre signatures *PROTEIN\_KINASE\_DOM* - PS50011, *PROTEIN\_KINASE\_ATP* - PS00107, *PROTEIN\_KINASE\_ST* - PS00108 et *PROTEIN\_KINASE\_TYR* - PS00109. À chaque entrée est jointe une documentation complète sur son rôle biologique. La plupart ont été construites en utilisant des alignements multiples basés sur l'approche de Gribskov (chaînes de Markov cachées (HMM)) [115,119]. Les cinq lignes de conduite de PROSITE sont :

(1) Son exhaustivité: la détermination d'annotations fonctionnelles requiert la présence d'une très grande quantité de motifs ayant une signification biologique.

---

(2) La spécificité des motifs permet de limiter au maximum le nombre de faux positifs et d'augmenter celui des vrais positifs.

(3) Une documentation exhaustive pour chaque entrée expliquant les raisons d'existence du motif concerné.

(4) Une révision périodique surveillant la validité des motifs.

(5) Un lien étroit avec la SWISSPROT [2] qui est la base de séquences la plus utilisée. Elle est en permanence mise à jour avec les séquences issues des différents projets Génome et la version 56.5 datant du 25 novembre 2008 contient 402 482 séquences extraites de plus de 250 organismes. Afin de faciliter l'accès aux annotations fonctionnelles, une autre base de données a été intégrée à PROSITE: ProRule [120]. Les annotations dépendent de la présence d'acides aminés précis à certaines positions, de l'occurrence d'autres domaines ou encore de spécificités de taxonomie. Chaque motif peut être retrouvé dans un certain nombre de séquences dont le statut peut être rangé en plusieurs catégories, FP : faux positif, TP : vrai positif, FN : faux négatif, TN : vrai négatifs. ProRule a permis l'incorporation de ScanProsite <http://www.expasy.org/tools/scanprosite/> [121], un outil permettant la détection de motifs PROSITE dans n'importe quelle séquence protéique. La version 20.40 datant du 26 novembre 2008 contient 1539 entrées de documentation, 1315 patterns, 819 profiles et 819 ProRule.

**PFAM** [117] est une base de données de famille de protéines chacune représentée par des alignements multiples de séquences ou des chaînes de Markov Cachées (HMM). Les protéines sont généralement composées de une ou plusieurs régions fonctionnelles, communément appelées domaines. Des combinaisons diverses de domaines forment les variétés de protéines existantes. PFAM se base sur l'assertion que la localisation de domaine dans les protéines peut fournir des éléments pour comprendre leur fonction. Chaque famille est vérifiée manuellement et est représentée par deux alignements différents. Un alignement local ou de graine («*seed*») qui contient les représentants caractéristiques et sûrs de chaque famille. L'alignement local d'une famille définit un profil qui sert de requête pour détecter tous les membres potentiels de la famille concernée. Le programme HMMER2, basé sur les modèles de chaînes de Markov cachées est utilisé sur les séquences de la SWISSPROT [2] pour créer un alignement global pour chaque famille. PFAM est composée de deux bases de données distinctes : PFAM-A où sont stockés les alignements locaux pour chaque famille et PFAM-B qui contient les alignements globaux correspondants. Comme la SWISSPROT est quotidiennement mise à jour avec de nouvelles séquences résolues, PFAM est aussi dotée d'un astucieux système de mise à jour. Le second rôle de PFAM-B est de constituer un espace

---

tampon pour les familles de protéines pas encore caractérisées et donc incluses dans PFAM-A. Lorsque la taille d'une famille devient assez grande, un alignement local est généré et est inséré dans PFAM-A. La taille de PFAM-A est donc quasiment constante excepté lorsque de nouvelles familles de protéines sont détectées et ajoutées. Comme PROSITE, des annotations spécifiques et claires sont disponibles pour chaque famille. Des informations structurales sont aussi ajoutées, notamment les classes SCOP (lorsqu'elles sont disponibles) auxquelles appartiennent les structures protéiques correspondant aux séquences. PFAM permet aussi de rechercher à quelle famille appartient une séquence via le site <http://pfam.sanger.ac.uk/search>. La version 23.0, datant de juillet 2008 contient 10 340 familles.

**Catalytic Site Atlas (CSA)** [118] contient des descriptions de sites actifs enzymatiques extraites de la littérature. La sélection des enzymes se fait en fonction d'un *E.C. number* (*Enzyme Class Number*) choisi. Un *E.C. number* est une classification des enzymes basée sur la réaction qu'ils catalysent [122]. Pour un *E.C. number* donné, si une structure de protéine contenant assez d'informations sur son site actif existe dans la PDB et que suffisamment des données sur la réaction qu'il catalyse sont disponibles dans la littérature, il est ajouté à la base de données CSA. D'autres recherches croisées sont ensuite effectuées afin d'être sûr que les données utilisées sont les plus à jour. Les résidus catalytiques sont annotés pour chaque entrée. Ensuite, les entrées de la base servent de requête pour PSI-BLAST [21] qui va rechercher des séquences homologues dans une base de séquences non redondantes des structures de la PDB. Chaque alignement est ensuite analysé. Les résidus correspondant aux résidus catalytiques de la requête sont extraits et si au maximum, un résidu diffère, la structure est stockée dans la base. La première version de CSA comptait 177 entrées originales pour rassembler 2608 entrées « homologues ». Aujourd'hui, la version 2.2.10 compte 968 entrées originales pour 23 265 homologues.

Les méthodes de comparaison de profils ou de motifs utilisées par ces bases de données ont une sensibilité plus importante que les méthodes de comparaison de séquences. En effet, les profils contiennent implicitement des informations sur les résidus qui sont les mieux conservés mais aussi les plus variables de la séquence en acides aminés. La fonction d'une protéine peut donc être déterminée si sa séquence contient un des profils ou motifs stockés dans une des bases de données décrites. Dans ce cas, les résidus impliqués dans la fonction sont identifiés. Leur localisation sur la surface d'une protéine peut aider à la compréhension des mécanismes dans lesquels la protéine est impliquée.

---

## ii. Méthodes basées sur la structure

Lorsque la structure 3D d'une protéine est disponible, la stratégie à adopter pour annoter fonctionnellement cette protéine dépend du type d'information disponible. Ici, la notion de surface d'interaction (ou de site de liaison) de la structure protéique est capitale. Elle se définit par une liste de résidu ou d'atomes qui vont interagir avec un partenaire biologique. En règle générale, trois possibilités s'offre au chercheur [123]:

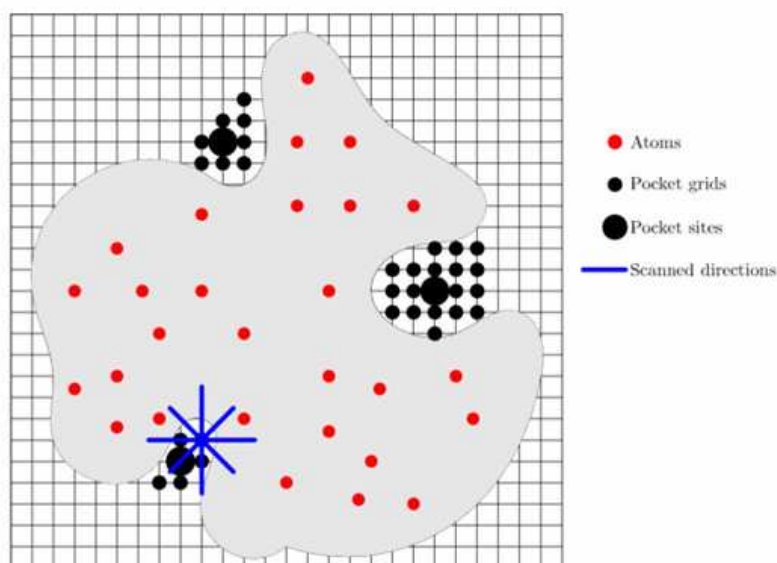
1. **Le site de liaison n'est pas connu.** La fonction générale de la protéine cible peut être connue sans que les détails structuraux du mécanisme soient bien caractérisés. Il convient donc d'utiliser des méthodes de localisation de sites de liaison. Une fois l'emplacement du site actif localisé, il est possible d'analyser la composition de la surface pour appréhender le mécanisme d'action. Il faut noter que de plus en plus de structures sans ligand sont résolues.
2. **La surface d'interaction est approximativement localisée sur la surface globale de la protéine sans qu'aucune information sur ses caractéristiques ou sur le type de ligand qui peut interagir ne soit disponible.** Cette situation est la plus commune et nécessite une analyse rigoureuse de la région concernée, afin de localiser les régions où certains groupes fonctionnels chimiques pourraient interagir. Une carte d'interaction fonctionnelle peut ainsi être générée permettant de guider le positionnement d'un ligand potentiel. Les méthodes de *docking* et de criblage virtuel positionnent ou construisent des ligands ciblant le site de liaison. D'autres types de méthodes permettent de construire un pharmacophore pour chercher des ligands spécifiques dans une base de données de molécules.
3. **Le site de liaison est connu et la structure est co-cristallisée avec son ligand.** Ces structures sont très utiles pour aider à la compréhension des mécanismes d'action de la protéine étudiée. Sa fonction n'est cependant pas toujours connue. Le site de liaison localisé peut être comparé à d'autres sites dont la fonction est connue. Les méthodes permettant cette comparaison utilisent le plus souvent des descripteurs mixtes pour

comparer les surfaces d'interactions, par exemple des descripteurs unissant des propriétés structurales et fonctionnelles.

### Critères géométriques

L'emplacement des sites de liaisons dépend directement de la forme de la surface des protéines [109]. Cette dernière est le plus souvent composée de cavités où se fixent les petites molécules. Aussi, de nombreuses méthodes de recherches ont été développées pour détecter les cavités sur les surfaces de protéines. Ces méthodes se basent principalement sur des critères géométriques. Les trois principales méthodes, LIGSITE [124], PASS [125] et PocketFinder [126] sont décrites ici.

**LIGSITE** [124] utilise une grille 3D de maille régulière qui contient la protéine étudiée. Les cellules de la grille en intersection avec la sphère de Van der Waals d'un atome sont éliminées et les autres intersections sont évaluées en fonction de leur degré d'enfouissement. Ce dernier est déterminé en parcourant les lignes de la grille selon les trois axes cartésiens, ainsi que les quatre diagonales cubiques de la grille (cf. Figure 10). Si une maille de la grille est entre deux atomes de la protéine, sa valeur est incrémentée. Les valeurs des scores varient entre zéro pour une maille à l'extérieure de la protéine, et sept pour une maille complètement enfouie dans la protéine. Les mailles adjacentes avec des scores supérieurs à 1 sont regroupées et forment les cavités détectées par cette approche. Seules les cavités d'une certaine taille sont conservées.



**Figure 10 : Représentation d'un résultat de Ligsite.**

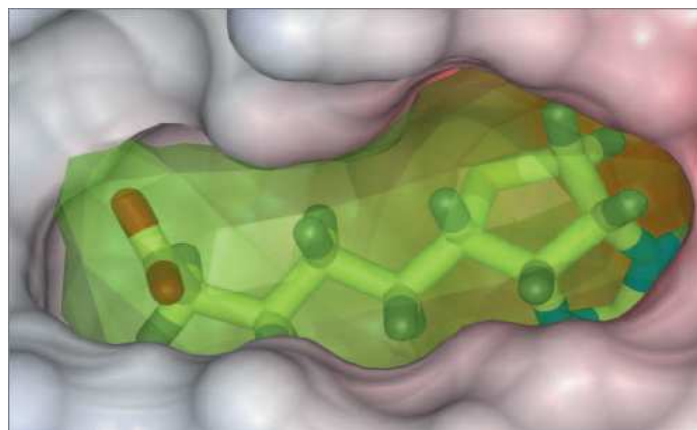
Trois cavités sont détectées par la méthode, les lignes bleues représentent les directions parcourues par la sonde (figure extraite de [127]).

**PocketFinder** [126] se base sur une grille 3D de taille de maille de 1 Å où la protéine est insérée. Dans cette méthode, le potentiel Lennard-Jones est calculé entre une sonde aliphatique qui parcourt les maillages de la grille et les atomes de la protéine (les molécules d'eau et les ligands co-cristallisés sont exclus), avec la formule suivante :

$$P_p^0 = \sum_{i=1}^N \left( \frac{A_{X_i C}}{r_{PI}^{12}} - \frac{B_{X_i C}}{r_{PI}^6} \right) \quad (2)$$

Où  $r_{PI}$  est la distance entre la sonde  $p$  et l'atome  $X_i$ . Les paramètres  $A_{XC}$  et  $B_{XC}$  sont extraits du champs de force (ECEPP)/3 (*Empirical Conformational Energy Program for Peptides* [128]).

Si un point de la grille est proche d'un nombre élevé d'atomes, la valeur du potentiel pour ce point est élevée. Si au contraire un point de la grille est proche de très peu d'atomes, la sonde interagit très peu et le potentiel est faible. Ce calcul permet aux régions enfouies d'être caractérisées par un haut potentiel. Une carte énergétique 3D est construite permettant de facilement localiser les régions à potentiel élevé. L'étape suivante consiste à créer des enveloppes électrostatiques des cavités potentielles en contournant ces régions. Finalement les sites de liaison détectés sont filtrés en fonction de la taille des enveloppes, seules celles ayant une taille supérieure à 100 Å<sup>3</sup> sont conservées. La Figure 11 illustre une des enveloppes détectées qui se forme autour d'un ligand co-cristallisé.



**Figure 11 : Exemple de site de liaison détecté par la méthode PocketFinder sur la structure 2IZI.**

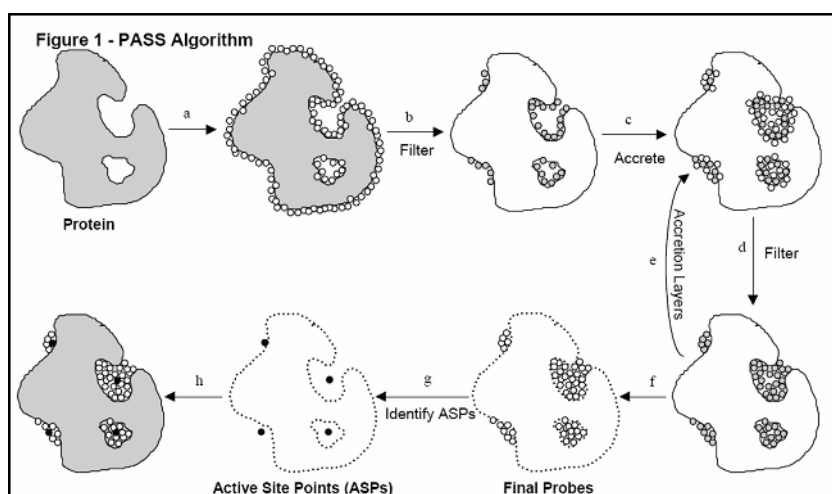
L'enveloppe prédite est colorée en vert autour du ligand biotine représenté en *stick*.

**PASS** (Putative Active Sites with Sphere) [125] est un outil qui utilise des sphères pour remplir les cavités de la surface des protéines afin d'identifier celles qui ont le plus de chance



d'être les centres de sites de liaison. Ces sphères sont appelées « active site points » (ASP). L'algorithme se décrit en sept étapes (cf. Figure 12):

- (1) Lecture des coordonnées dans le PDB et attribution des rayons atomiques.
- (2) Première couche de sphères tout autour de la protéine en utilisant la méthode des trois points de Connolly [129] (cf. Annexe 1).
- (3) Premier filtre appliqué sur les sphères: élimination des sphères ayant un clash stérique avec la protéine, de celles qui ne sont pas assez enfouies et de celles qui sont à une distance inférieure à 1Å d'une sphère plus enfouie.
- (4) Ajout d'une nouvelle couche de sphères.
- (5) Second filtre pour éliminer les sphères s'encastant avec d'autres (clash stérique).
- (6) Renouvellement des étapes 4 et 5 jusqu'à ce qu'aucune nouvelle sphère ne soit ajoutée.
- (7) Calcul des ASP, situés au centre de gravité des amas de sphères.



**Figure 12 : Sept étapes de l'algorithme du logiciel PASS.**

Figure extraite de l'article [125]

Ces méthodes sont efficaces pour localiser un grand nombre de cavités à la surface des protéines et dans certains cas cela peut suffire à situer le site fonctionnel d'une protéine. Huang et ses collaborateurs présentent une méthode basée sur LIGSITE, LIGSITEcsc et la compare à d'autres approches géométriques dont certaines sont décrites précédemment [127] sur un jeu de 210 structures de la PDB liées à des ligands (cf. **Tableau 2**).

Method	Top1	Top3
LIGSITE <sup>csc</sup>	75%	
LIGSITE <sup>cs</sup>	67%	87%
LIGSITE	65%	85%
PASS	54%	79%
SURFNET	42%	56%

**Tableau 2 : Comparaison des méthodes de détection de sites basées sur des critères géométriques sur un jeu de 210 structures.** La colonne « TOP 1 » contient le pourcentage de sites détectés en premier et la colonne « TOP 3 » contient le pourcentage de sites détectés dans les trois premiers. Ce tableau est extrait de l'article [127].

Malgré l'efficacité de ces méthodes, quand plusieurs cavités sont détectées sur une même protéine, il faut pouvoir définir celle qui devrait être fonctionnelle. La question de savoir ce qui distingue un site de liaison des autres cavités est fondamentale. Il est difficile d'y répondre simplement.

L'utilisation d'une fonction d'énergie ou la caractérisation physico-chimique de la surface des protéines permet de classer des sites d'interactions déjà détectés par une recherche de cavités. Ces critères permettent aussi de prévoir directement l'emplacement de sites d'interaction potentiels.

### Critères énergétiques

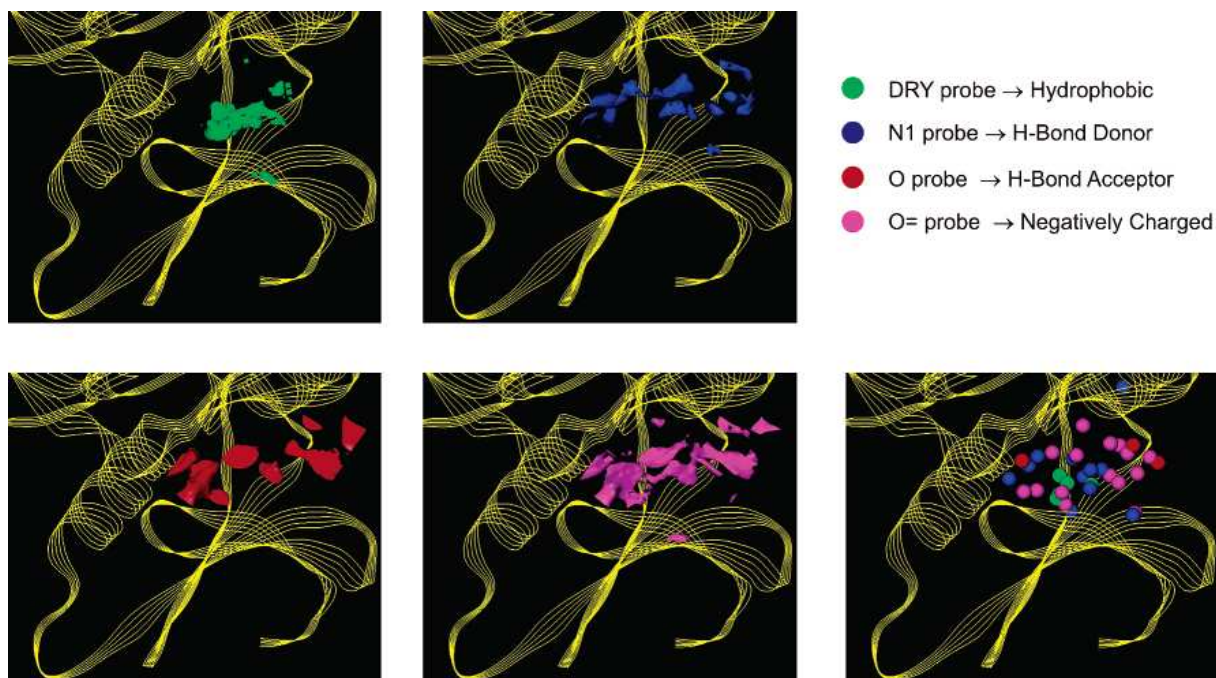
D'un point de vue énergétique, la structure d'une protéine est particulièrement complexe [130]. Les interactions hydrophobes ou hydrophiles modifiées durant la complexation de la protéine avec un ligand doivent rester très localisées afin de ne pas perturber la stabilité de la protéine. Pour rester accessible, une surface d'interaction doit exposer des régions hydrophobes et des charges au solvant. Elle doit aussi fournir des accepteurs ou des donneurs de liaison hydrogène non liés sans engendrer de repliements locaux. Ainsi, les sites de liaisons semblent être des régions plus instables que le reste de la surface. Certaines méthodes exploitent cette propriété et se basent sur différents potentiels énergétiques pour localiser des sites de liaison.

Roterman et collaborateurs [131] ont mis au point un modèle pour la reconnaissance de site de liaison (le modèle « *Fuzzy Oil Drop* », FOD) basé sur l'analyse de la distribution non aléatoire de l'hydrophobicité sur la structure des protéines. Ce modèle avait initialement été utilisé pour simuler le repliement partiel des protéines dû aux cœurs hydrophobes des protéines [132]. Des études de la distribution hydrophobe sur des structures de protéines

---

naturelles ont permis de détecter un comportement particulier au niveau des sites de liaisons. Ces études ont mené au développement d'une méthode de détection de sites fonctionnels utilisant un champ de force hydrophobe externe. Elle se base sur la détection d'irrégularités au sein de la distribution hydrophobe à la surface des protéines et compare les sites détectés aux données expérimentales de la base CSA. Sa comparaison avec Profunc [133] et SuMo [8,9] (méthodes détaillées dans la suite du manuscrit) sur un jeu de protéines enzymatiques fonctionnellement caractérisées, montre des résultats satisfaisants [131].

**FLAP** (“*Fingerprints for Ligands And Proteins*”) [134] analyse les cavités détectées à la surface des protéines en utilisant différents types de sondes qui représentent chacune une fonction chimique qui pourrait interagir avec la protéine étudiée. En évaluant systématiquement au sein d'une grille 3D l'énergie calculée à partir des champs de force d'interaction moléculaire GRID (GRID-MIFS) [110]. Les sondes vont permettre de créer une carte 3D d'arrangements énergétiques favorables et défavorables à l'interaction d'une molécule avec la protéine. Cinq cartes du site de liaison de la structure 1H1S sont représentées Figure 13. Les couleurs représentent les endroits du site où une interaction avec le type de sonde donnée serait probable. La carte qui contient différentes couleurs est une synthèse des quatre autres. La carte énergétique d'un site peut être comparée à celles d'autres sites afin de les comparer. Elle peut aussi être considérée comme un négatif de pharmacophore ayant pour utilité le criblage de molécules afin d'identifier des molécules potentiellement actives sur le site étudié. Cette méthode produit donc une armature commune pour analyser et comparer les ligands ensembles. Les cartes énergétiques peuvent aussi être représentées sous forme d'empreintes compactes (« *fingerprints* ») et plus facilement exploitables pour de comparer les sites entre eux. Une validation de la méthode sur un jeu de 23 protéines kinases de différentes familles est effectuée et elle montre que la méthode permet de séparer les protéines par familles de protéines kinases [134].



**Figure 13 : Exemple d'analyse de la protéine kinase 1H1S avec les champs de force de GRID-MIFs.** Chaque couleur est spécifique d'un type de sonde : hydrophobe (vert), donneur de liaison hydrogène (bleu), accepteur de liaison hydrogène (rouge) et charge négative (rose). La carte en bas à droite est la synthèse des quatre sondes. Elle peut être utilisée pour le criblage virtuel de ligands spécifique de ce site de liaison ou pour la comparaison avec d'autres sites de liaisons. Figure extraite de [134].

La base de données **eF-site** (« *electrostatic surface of functional-site in proteins* ») [135] contient des surfaces moléculaires de sites fonctionnels de protéines calculées avec la méthode de Connolly [129] (cf. Annexe 1). La surface est représentée sous forme de maille triangulaire et pour chaque sommet, sa position, un potentiel électrostatique et ces propriétés hydrophobes sont stockées dans la base de données. Ces propriétés servent de descripteurs. Le but de cette approche est d'attribuer la fonction d'une protéine en comparant sa structure aux éléments de la base. Une solution est retenue si la surface requête ainsi que ses descripteurs électrostatiques ressemblent à un élément de la base. Pour valider leur approche, les auteurs détectent des similitudes fonctionnelles entre des phosphoenolpyruvate carboxy kinases et des protéines liant des mono-nucléotides de différents repliements. Ensuite ils tentent d'attribuer une fonction à la protéine hypothétique MJ0226 (code PDB 1B78) en comparant sa surface aux éléments de la base. Un serveur web est disponible. Il est donc possible de comparer une structure à l'ensemble des surfaces contenues dans eF-site.

### Critères physico-chimiques

La reconnaissance moléculaire engendre la modification des réseaux d'interactions entre groupements chimiques présents à la surface d'une protéine avec ceux de son partenaire

---

et de l'environnement. Ces groupements sont des donneurs et des accepteurs de liaisons hydrogène, des groupements chimiques hydrophobes, des charges positives ou négatives, *etc.* Ils sont tous facilement caractérisables sur les acides aminés. La prise en compte des positions et des types de groupements chimiques impliqués dans les liaisons protéine-ligand permet la création de **motifs fonctionnels tridimensionnels**. La conservation de ces motifs dans des sites actifs issus de protéines différentes implique une relation fonctionnelle [123]. Quelques méthodes de comparaison structurale utilisent ces motifs pour comparer des sites de liaison ou pour détecter la présence de surfaces d'interactions spécifiques sur les structures de protéines hypothétiques. Ces motifs permettent aussi la comparaison des différents modes de liaison de certains ligands. Par exemple, deux protéines peuvent interagir avec le ligand ATP et avoir des sites actifs très différents. Cavbase [12], SiteEngine [11,136-138] et SuMo/MED-SuMo [7,9,10] sont trois méthodes de reconnaissance de sites fonctionnels sur la surface des protéines. Chacune se base sur la recherche de ces motifs fonctionnels tridimensionnels pour comparer et annoter les protéines. Elles se distinguent au niveau de leur algorithme de comparaison, ou des types de groupements chimiques inclus dans les motifs, de leur rapidité et aussi leur accessibilité.

**Cavbase** [12] est une base de données de cavités des structures issues de la PDB détectées par LIGSITE [124]. Ces cavités sont caractérisées par les coordonnées atomiques des résidus qui les composent, réduits à un ensemble de descripteurs appelés des *pseudocentres*, regroupés en 5 propriétés chimiques : donneur de liaison hydrogène, accepteur de liaison hydrogène, objet mixte donneur/accepteur, groupe hydrophobe et groupe aromatique. Une fois ces *pseudocentres* détectés, ils sont examinés pour vérifier s'ils sont effectivement situés à la surface de la cavité. Chaque objet est ensuite placé à une position définie par LIGSITE sur la surface. La forme de la cavité, l'ensemble des descripteurs et leur correspondance sur la surface de la protéine sont stockés dans la base de données. L'algorithme de recherche de similarités se base sur une détection de clique pour détecter les sous-graphes similaires formés de paires de « *pseudocenters* » de propriétés équivalentes et situés à une distance acceptable. Une certaine tolérance de superposition est permise par l'algorithme afin de gérer au mieux la flexibilité conformationnelle des protéines. Une cavité requête (dite « *query-cavity* ») peut donc être comparée à l'ensemble de cavités stockées dans la base Cavbase. Les cavités retrouvées aux meilleurs rangs sont celles qui ont le plus de similarités locales et dont les modes de liaisons ressemblent le plus à ceux de la cavité requête. Cette approche étant basée entièrement sur les propriétés physico-chimiques

---

exposées à la surface et non des acides aminés, est donc indépendante d'homologie de structure et de séquence. Aucune information sur le temps de comparaison nécessaire à Cavbase n'est précisée. Il est simplement spécifié que l'algorithme de comparaison est assez lourd et long du fait de la recherche de clique. Cavbase est un module inclus dans le package de Relibase [139] qui regroupe un ensemble de base de données, ainsi qu'une interface graphique. Celle-ci se compose d'une version commerciale et d'une version gratuite pour les académiques.

**SiteEngine** [136-138] est une méthode qui permet de rechercher un site fonctionnel précis sur les surfaces d'un ensemble de protéines, de comparer un site de liaison à un ensemble de sites de liaison et de détecter un site de liaison sur la surface d'une protéine. Les applications impliquent deux types de comparaisons : la recherche sur une surface de protéine entière d'un site donnée, et la comparaison entre deux sites de liaison. Le même algorithme de comparaison est utilisé dans les deux cas. Comme pour Cavbase, la première étape consiste à détecter les *pseudocentres* sur les résidus de la surface considérée (pour un site ou la surface entière). Chacun de ces pseudocentres représente un centre d'interaction possible. Les propriétés chimiques sont les mêmes que pour Cavbase : donneur de liaison hydrogène, accepteur de liaison hydrogène, objet mixte donneur/accepteur, groupe hydrophobe et groupe aromatique. La première étape de la méthode consiste à détecter les *pseudocentres* sur les résidus. Lorsque la surface entière de la protéine est caractérisée, seuls les *pseudocentres* situés sur la surface telle qu'elle est définie par une approche de « Connolly » sont conservés (cf. Annexe 1). Pour les sites de liaisons, les *pseudocentres* situés dans les 4 Å autour du ligand sont conservés. Les *pseudocentres* obtenus sont assemblés en triplets. Les triplets dont la taille des arêtes est acceptable sont stockés dans une table de hachage. Un triplet est représenté par une clé qui contient des informations sur la taille des arêtes et un index physico-chimique. L'algorithme est basé sur la recherche de paires de triplets similaires entre les deux sites comparés. Les triplets compatibles qui sont adjacents sont regroupés. Chaque ensemble de paires de triplets représente des superpositions d'une partie de la surface des deux protéines comparées. Les résultats sont ensuite classés en fonction de leur score. Ce score dépend des *pseudocentres* compatibles entre les surfaces similaires détectées. Un serveur web est disponible. Il permet de comparer un site de liaison requête ou la surface entière d'une protéine à une base de cavités pré-compilées (<http://bioinfo3d.cs.tau.ac.il/SiteEngine/bin/getSiteEngineData.pl>)

---

**MED-SuMo** est le produit dérivé à vocation industrielle de SuMo [7,9,10] dont le développement a été réalisé par Martin Jambon lors de sa thèse sous la direction de Christophe Geourjon, au laboratoire CNRS, Institut de Biologie et Chimie des Protéines (IBCP) [8]. Cette méthode localise rapidement les surfaces d'interaction similaires entre macromolécules. Si son approche possède des similitudes avec SiteEngine et Cavbase, MED-SuMo est tout de même très original. Tout d'abord, il dispose d'un dictionnaire puissant et reconfigurable pour décrire les différents groupements chimiques à retenir pour cartographier et comparer les surfaces d'interaction. Ensuite, il intègre des coordonnées pseudo-internes pour mieux gérer les variabilités conformationnelles. Il propose de plus une fonction de score avancée (cf. partie II.D.3.). Enfin, il s'appuie sur une transformation de l'espace 3D de la surface d'interaction en un graphe de triplet de groupements chimiques afin permettre une accélération significative du temps de calcul. Comme MED-SuMo est l'outil au centre de mon travail de thèse, j'ai choisi d'y consacrer une partie entière (cf. partie II).

Toutes les méthodes de détection ou de comparaison de surfaces d'interactions n'ont pas été décrites. Notamment, la méthode « Evolutionary Trace » (ET) [111] qui se base sur des études phylogénétiques pour identifier les résidus fonctionnellement importants. Ces résidus sont ensuite localisés sur la surface des structures résolues de protéines et recherchés dans d'autres structures. CPASS [140] extrait les résidus autour des ligands co-cristallisés dans les structures des protéines ainsi que leurs positions et les stocke dans une base de données. Un algorithme d'alignement géométrique permet de comparer un site requête à ceux de la base. Enfin, ProFunc [133] est un serveur qui combine plusieurs méthodes pour analyser une protéine. La séquence et la structure d'une protéine sont analysées en même temps, détectant ainsi des séquences voisines (PSI-BLAST), des structures voisines (DALI), et la présence de motifs PROSITE ou PFAM particuliers. ProFunc permet aussi de détecter quels résidus sont potentiellement impliqués dans l'activité de la protéine étudiée. Aucune de ces méthodes prises séparément ne peut systématiquement assurer la caractérisation de toute protéine donnée. Le principe de ce serveur est donc d'en combiner plusieurs dans le but de rassembler plus d'information possible sur la structure étudiée. Toutefois, la profusion des résultats rend fastidieux les études portant sur plusieurs protéines. Aucun résumé simple efficace n'est en effet proposé.

La détection d'une surface d'interaction d'une protéine n'est en aucun cas triviale. Les méthodes décrites dans cette partie sont complexes et chacune est validée sur différents jeux

---

de protéines. Il est donc difficile de comparer leurs performances. Pour ce faire, il faudrait pouvoir accéder à chacune des méthodes, analyser des protéines particulièrement étudiées et connues, et accéder à tous les résultats. Certaines de ces méthodes nécessitent l'achat de licence commercial comme Cavbase. Il n'a donc pas été possible dans le cadre de ma thèse de comparer en détail les résultats et les performances obtenues entre ces logiciels. Dans son étude, Rotterman et ses collaborateurs [131] valident le modèle FOD de détection de site de liaison en comparant les données expérimentales de la base CSA (cf. I.i) avec leurs résultats et ceux des approches SuMo et ProFunc sur un jeu de 33 structures. Malgré cette comparaison, il est difficile se prononcer en faveur d'une des méthodes. Les résultats sont surtout tous très différents et ils mettent en avant la complexité du problème posé.



## PARTIE II : Présentation de MED-SuMo

---

## II. Présentation de MED-SuMo

### A. Heuristique de la méthode

MED-SuMo est la version dérivée à vocation industrielle du logiciel SuMo conçu pour localiser les régions semblables sur les surfaces des protéines associées à une fonction définie [9,10]. Si la majorité des détails de l'heuristique SuMo sont présentés dans la thèse de Martin Jambon [8], je rappellerai ici les principaux paramètres et présenterai certaines des parties développées à MEDIT et intégrées dans MED-SuMo.

L'heuristique est basée sur une représentation en 3D des macromolécules biologiques avec des groupements chimiques particuliers situés à la surface des protéines. Ces groupes sont nommés les « *Surface Chemical Feature* » (SCFs). Ils désignent des types présents dans Cavbase et SiteEngine : les donneurs et accepteurs de liaisons hydrogène non liés, les charges positives et négatives, les objets aromatiques et hydrophobes, mais aussi d'autres groupements spécifiques des acides aminés tel le groupement amide, hydroxyle ou le groupement guanidinium présent sur le résidu arginine. Les molécules d'eau identifiées pour stabiliser le ligand avec la protéine peuvent être optionnellement représentées par l'objet spécifique « *structural water* ». Chaque SCF est encodé par ses caractéristiques chimiques mais surtout par des règles géométriques précises d'ailleurs configurables au sein d'un dictionnaire séparé.

L'algorithme global de comparaison de surface d'interaction disponible dans MED-SuMo est illustré Figure 14. Il se divise en deux étapes principales : la construction d'un graphe requête et la comparaison de graphes.

#### Construction d'un graphe requête

Les SCFs sont détectés sur la surface de la protéine choisie par une analyse lexicographique du fichier PDB: un résidu est décrit par un ensemble de SCFs, par exemple la phénylalanine est représentée par deux accepteurs de liaison hydrogène, un donneur, un aromatique et trois objets hydrophobes. La surface analysée peut être un site de liaison, une surface entière de protéine ou encore une sélection manuelle d'un utilisateur. Une fois les SCFs détectés, leurs positions et orientations sont vérifiées afin d'éliminer ceux qui sont trop enfouis (paramètre « *densmax* ») et ceux déjà impliqués dans une interaction intra-protéique (cf. Figure 14a et b). Les SCFs restants sont assemblés en triplets ayant des caractéristiques géométriques strictes (tailles d'arête et angles minimum et maximum, périmètre maximum...)

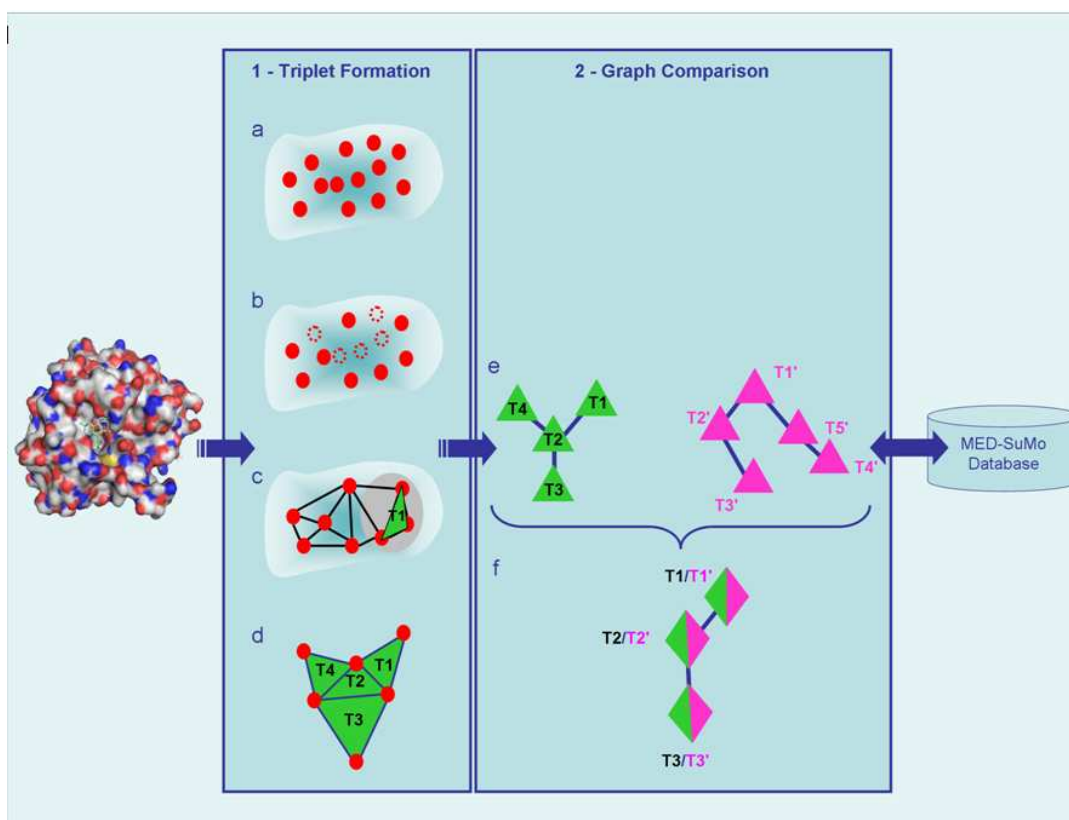
---

(cf. Figure 14c). Ces triplets vont constituer les sommets d'un graphe dont les arêtes ne relient que les triplets adjacents (cf. Figure 14d). Le graphe formé peut être utilisé comme graphe requête ou être stocké dans une base de données.

### Comparaison de graphes

Pour comparer deux graphes, l'algorithme utilisé est la recherche de sous-graphes communs. MED-SuMo recherche toutes les paires des triplets ayant à la fois les mêmes types de SCFs (cf Figure 14e) et partageant un ensemble de propriétés communes (longueur des arêtes, orientation du triplet, enfouissement, orientation des SCFs, forme locale). À partir de ces paires de triplets, qui sont des graines de comparaison, MED-SuMo étend les comparaisons aux sommets voisins, représentant les triplets adjacents, jusqu'à ce qu'aucune autre similitude ne soit trouvée. Ce procédé permet la formation de *patches* comportant donc au moins 3 SCFs en commun à partir des deux graphes comparés. Chaque paire de *patches* a donc une signature en SCFs commune, et sont classés en fonction du score MED-SuMo. Ce score est la somme de l'influence de chaque SCF sur le *patch*, chaque SCF étant défini par un poids par défaut qui est pondéré par l'environnement atomique des SCFs (cf. partie II.D.3). Les graphes similaires détectés par MED-SuMo sont appelés des *hits*.

Ces comparaisons sont habituellement effectuées entre une requête et une base de données de graphes pré-compilés. Une commande (cf. partie II.D.32.i) permet de convertir automatiquement chaque fichier au format PDB permet de construire ces graphes soit à partir de l'ensemble des atome de la structure de la protéine, soit à partir des atomes à proximité de chaque ligand co-cristallisé avec la structure. Ainsi, deux sortes de bases sont disponibles: la base de surfaces entières dont les graphes sont composés de SCFs sur toute la surface des structures, et la base de sites de liaisons dont les graphes sont composés de SCFs autour de ligands co-cristallisés. Les caractéristiques des bases sont définies par trois paramètres essentiels : (1) La taille de l'environnement considéré pour les bases de sites (paramètre *ligand\_radius*), les valeurs 4,5 Å et 6,0 Å sont les plus communément utilisées. (2) La distance maximale entre deux SCFs pour qu'ils soient inclus dans un triplet (paramètre *edge\_max*). (3) Le périmètre maximal d'un triplet (paramètre *max\_edge\_sum*). Le détail mathématique de cette heuristique est présenté de manière complète par Martin Jambon dans son manuscrit de thèse (page 45 [9]), le score est re-détaillé par la suite (cf. partie I.D.3).



**Figure 14 : Procédure de comparaison disponible dans MED-SuMo.**

Elle se divise en deux étapes : (1) Construction d'un graphe requête : (a) Les SCFs sont détectés sur la structure de la protéine. (b) Ils sont ensuite filtrés selon des critères d'enfouissement ou d'implication éventuelle dans des liaisons intra-protéines. (c) Les SCFs restant sont rassemblés en triplets. (d) Ces triplets constituent les sommets d'un graphe. Deux sommets sont connectés si les triplets qu'ils représentent sont adjacents. (2) Comparaison de graphes. (e) Le graphe requête (en vert) est comparé aux graphes d'une base de données (en rose). Les triplets compatibles sont détectés, i.e., ils sont composés de SCFs compatibles et présentent un ensemble de propriétés géométriques communes. (f) Les *hits* sont classés par score MED-SuMo et peuvent être analysés à l'aide de l'interface graphique MED-SuMo GUI.

La représentation des surfaces en graphe de triplets ayant des caractéristiques géométriques précises permet à l'algorithme de comparaison d'être rapide et puissant. En effet, la comparaison d'un site de liaison à l'ensemble des sites de liaisons des structures de la PDB prend environ 15 minutes sur une machine linux datant de 2006 (bi-Opteron AMD Dual Core 270, 6GB RAM). De plus, une superposition peut dépasser le millier de SCFs si nécessaire. MED-SuMo gère la flexibilité au niveau de la détection des similitudes de la détection des similitudes de surfaces d'interaction en tolérant une certaine déviation entre les SCFs, ainsi qu'un certain angle pour les directions des vecteurs des SCFs. (cf. les SCFs encerclés dans la Figure 39).

La compilation des bases de surfaces d'interaction MED-SuMo, ainsi que les recherches (ou *runs*) lancés utilisent des fonctions qui distribuent les calculs si le noyau (*kernel*) linux utilisé tourne sur plusieurs CPUs et cœur. Un des paramètres *sumo\_process\_num* sert à

---

spécifier le nombre de CPU utilisé. Ainsi, MED-SuMo est capable d'exploiter au mieux la puissance des machines sur lequel il est utilisé. Plus la machine du serveur dispose de processeur et cœur multiples, plus les calculs seront rapides.

## **B. Utilisation du logiciel**

MED-SuMo fonctionne avec un système client-serveur. Le serveur est écrit en OCaml [141] et fonctionne sous Linux. La version originale (SuMo) est disponible sur un serveur web à l'adresse: <http://sumo-pbil.ibcp.fr/cgi-bin/sumo-welcome> [10]. Afin d'augmenter la robustesse d'accès aux données, MED-SuMo stocke ses bases de surface d'interaction à travers le système de gestion de base de données relationnelle SQLite. La librairie mlsqLite [142] permet l'interfaçage entre ce moteur de base de données et le serveur MED-SuMo. Le principal avantage de SQLite est que les bases MED-SuMo sont stockées sous forme de fichiers binaires directement gérés par le programme. L'installation d'un système de gestion de bases de données indépendant n'est pas nécessaire. L'utilisation de MED-SuMo est possible à travers trois interfaçages : les deux premiers concernent exclusivement le serveur et le troisième est la base des communications client-serveur.

- **MED-sumo-clui** (« *command line user interface* ») est une interface en lignes de commande Shell qui gère l'essentiel des fonctionnalités de MED-SuMo telles que la construction des différentes bases (*pdb\_index.db*, *sites*, *full*) décrites dans la partie II.D.2. La liste des commandes disponibles, accompagnées d'une description, se trouve dans l'Annexe 2.
- **MED-sumo-lua** est une interface avec le langage de script Lua [143]. Elle permet l'utilisation de certaines fonctions de MED-SuMo serveur à travers des scripts Lua. Il est par exemple possible de lancer un *run*, et d'exporter les résultats aux formats texte, CSV ou XML. L'interfaçage avec le Lua permet le lancement de calculs en ligne de commande. Cette fonctionnalité permet l'utilisation de MED-SuMo pour des calculs à grande échelle.
- **MED-sumo-client** est l'interface qui permet la communication client-serveur de MED-SuMo. Cette communication se fait via deux programmes différents *sumo-client.exe* et *sumo-client-batch.exe*. Le premier est utilisé par

l'interface graphique, MED-SuMo Graphical User Interface (MED-SuMo GUI), décrite dans la partie I.C qui fonctionne sous MS-Windows et qui constitue aujourd'hui l'un des atouts majeurs du logiciel par rapport aux autres approches (Cavbase, SiteEngine). Le second permet le lancement des fonctions utilisées par la GUI mais en mode batch. Toutes les fonctionnalités de MED-SuMo sont ainsi disponibles en ligne de commande. Le programme *sumo-client-batch.exe* a permis notamment le développement d'un composant pour le logiciel Scitegic Pipeline Pilot<sup>TM</sup> [144] permettant l'intégration de MED-SuMo dans des enchaînements de programmes (cf. Figure 51).

La Figure 15 représente deux manières de lancer un même *run* MED-SuMo, une utilisant une requête MED-sumo-client et l'autre un script MED-sumo-lua.

**Requête MED-sumo-client**

```
sumo-client -address 192.168.0.100 -port 9393 -u anonymous -p "" --command
"{function = "sumo-run" ;
  query = {{
    scan = {
      pdb_id = "1ATP";
      selection = "Ligand_binding_site 1 ";
      database = "ligands";
    }
  }} ;
}"
```

**Script MED-sumo-lua**

```
molec = SuMo.read_pdb_id("1ATP");
graph = SuMo.graph(molec, {select="{Ligand_binding_site 1}"});
result = SuMo.compare_scan(graph, "sites");
SuMo.output_comparison_text(result, "sumo-results-1ATP_1.txt");
SuMo.output_comparison_csv(result, "sumo-results-1ATP_1.csv");
```

**Figure 15 : Lancement du même *run* MED-SuMo de deux manières différentes.**

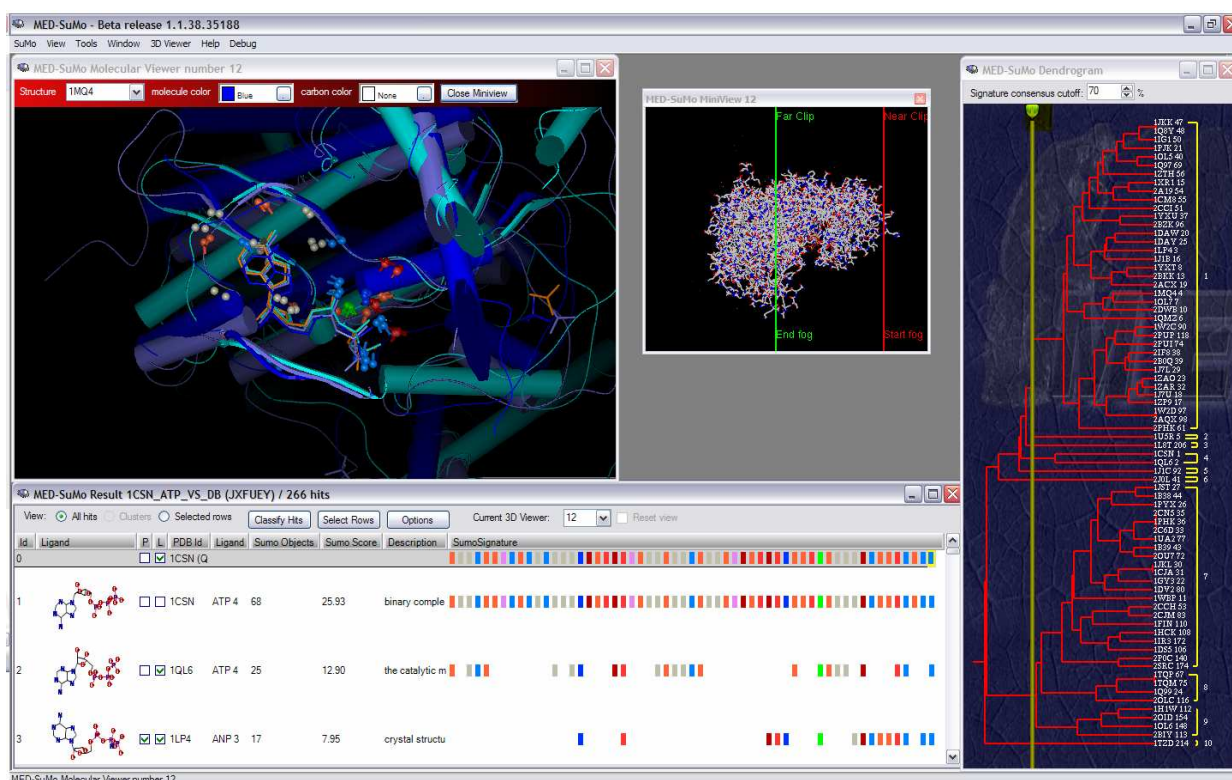
La partie supérieure contient une requête MED-sumo-client en ligne de commande. La partie inférieure représente le contenu d'un script Lua appelé *sumo-run.lua* qu'il faut lancer avec l'exécutable MED-sumo-lua. Le script se lance avec la commande : "# sumo-lua sumo-run.lua".

---

### C. Interface Graphique : MED-SuMo GUI

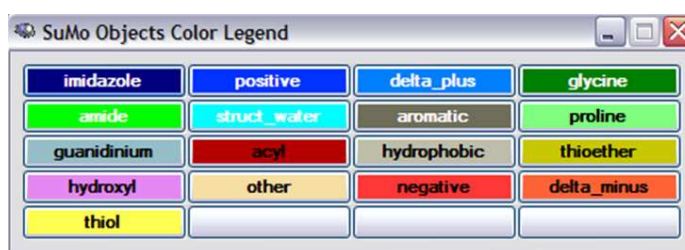
L'interface graphique développée à MEDIT, permet de simplifier l'utilisation de MED-SuMo au travers d'une technique simple et puissante. Elle permet la visualisation de la structure 3D d'une protéine, la sélection des atomes à considérer dans la requête MED-SuMo et le lancement de comparaison sur des bases de sites ou de surface entières pré-compilées. La Figure 16 est une capture d'écran de MED-SuMo GUI.

Pour commencer une recherche avec MED-SuMo, l'utilisateur charge sa protéine requête dans le visualisateur 3D ou bien spécifie un code PDB s'il est déjà disponible dans la base de surface d'interaction. Les sites de liaisons sont détectés automatiquement par la présence de ligands ou de peptides co-cristallisés. L'utilisateur peut alors sélectionner soit un ligand, et donc la surface d'interaction entourant ce ligand, soit la surface entière de la protéine, ou encore construire une sélection manuelle autour de résidus, atomes, ou SCFs précis. L'utilisateur peut ensuite lancer la comparaison sur une base de sites de liaison ou sur une base de surface entière. Les résultats obtenus par MED-SuMo sont affichés dans un tableau de résultats contenant sur chaque ligne la structure 3D projetée du ligand, le score MED-SuMo, la signature SCFs et d'autres annotations sont édités (cf. Figure 16). Les similitudes entre la requête et les résultats (*hits*) peuvent être examinées dans le visualisateur 3D où les superpositions sont disponibles. Ces superpositions permettent de comparer en 3D les poches et sous-poches de SCFs communs détectés. Dans le cas des bases MED-SuMo de sites de liaison, il est possible de visualiser les ligands co-cristallisés de la requête et des hits afin de mieux comprendre les interactions protéine-ligand mises en jeu. La colonne la plus à droite dans le tableau de résultat représente la signature globale des SCFs communs entre la requête et les *hits*. Une classification hiérarchique ascendante basée sur le calcul d'une distance entre les signatures SCFs de chaque hit mais aussi la requête, permet de regrouper les *hits* en fonction de leur signature. Les *hits* s'affichent dans un dendrogramme représenté sur la Figure 16. J'ai participé au développement de l'outil, dans le cadre du savoir-faire MED-SuMo mis en œuvre par MEDIT, dans l'implémentation de nouvelles fonctions MED-sumo-client, mais aussi, comme la plupart du personnel de MEDIT, au débogage et aux tests de l'outil.



**Figure 16: L'interface graphique de MED-SuMo: MED-SuMo GUI.**

Quatre fenêtres sont représentées ici: (1) Le visualisateur 3D est en haut à gauche: trois éléments y sont superposés: les protéines, les ligands co-cristallisés en représentation bâton et enfin les SCFs qui ont permis cette superposition. Le visualisateur est un Active X qui permet à l'utilisateur de réorienter ou de déplacer les structures (rotation, translation). (2) Le tableau de résultat situé en bas à gauche: La première ligne correspond à la requête ; son nom et sa signature en SCFs. Chaque autre ligne correspond à un *hit* détecté par MED-SuMo. Les *hits* sont classés par ordre de score décroissant (8<sup>ème</sup> colonne). Pour chaque ligne, différents éléments sont disponibles : projection 2D du ligand, le score MED-SuMo, le nombre d'SCFs commun avec la requête, le nom du ligand, le HEADER du fichier PDB du *hit*. La colonne la plus important contient la liste de SCFs commun entre la requête et le hit: la signature SCF. Chacun des SCFs représente une similitude structurale et fonctionnelle. Ils sont représentés par des rectangles colorés dont la légende est représentée Figure 17. (3) Le mini visualisateur permet à l'utilisateur de régler la profondeur graphique. (4) Un algorithme de classification a été intégré au MED-SuMo GUI afin de classer les hits en fonction de leur signature SCFs. La quatrième fenêtre contient un dendrogramme où les hits sont classifiés. Le nombre de groupes (*clusters*) peut être choisi par l'utilisateur en déplaçant le curseur jaune.



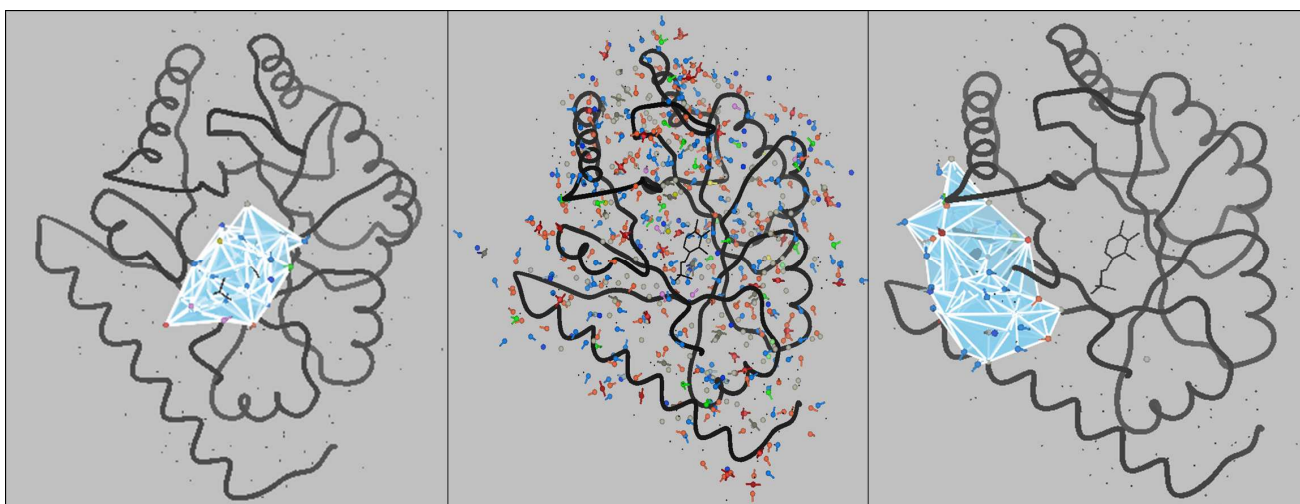
**Figure 17 : Liste et couleurs des SCFs utilisés dans MED-SuMo GUI**



## D. Détails sur les composants de MED-SuMo

### 1. Différents types de requêtes

MED-SuMo détecte les régions similaires sur la surface des structures protéiques. Un langage de script de sélection est directement intégré au serveur de MED-SuMo afin de rendre le type de requête modulable. En effet, l'utilisateur peut choisir de soumettre une surface entière de protéine, un ou plusieurs sites de liaisons détectés par la présence de ligands co-cristallisés. Il est également possible de choisir une des chaînes de la protéine, un ensemble de résidus ou plus spécifiquement un ensemble de SCFs. La Figure 18 montre trois différents types de requêtes possibles.



**Figure 18: Différents types de requêtes possibles pour MED-SuMo sur la structure hypothétique 1B54.**

De gauche à droite: (1) Requête la plus classique impliquant les SCFs situés à proximité d'un des sites de liaison détectés. Les triplets formés sont affichés en bleu. (2) Ici, la requête soumise est la surface entière de la protéine.

Des SCFs sont détectés tout autour de la protéine. Les triplets n'ont pas été affichés par souci de clarté. (3) La requête soumise ici est manuelle. Quelques résidus du fichier PDB ont des facteurs B élevés (résidus 227 à 234, d'une valeur supérieure à 15), ces résidus étant instable peuvent être stabilisés par une interaction avec une autre molécule d'après Barlett et ses collaborateurs [40]. Ils sont sélectionnés ici pour construire une requête MED-SuMo.

### 2. Différents types de bases de données

La requête est ensuite comparée à un ensemble de graphes MED-SuMo pré-compilés. Deux types de bases sont utilisés : la base de sites et la base de surfaces entières. Chaque base est définie par différents paramètres. Elles se calquent toutes les deux sur les mises à jour de la PDB qui est elle-même importée dans une table séparée. Ces trois bases de données sont au format SQLite, et sont décrites ci-dessous.

---

### i. *L'index de la PDB*

Les fichiers de la PDB sont indispensables pour le fonctionnement de MED-SuMo. La base *pdb\_index.db* permet le référencement de chacun des fichiers de la PDB, et fait le lien avec les bases de graphes. Cet index ne contient pas d'informations sur les structures mais le chemin relatif d'accès à chaque fichier PDB. Il rend aussi possible l'intégration de n'importe quelle structure de protéines (structures privées ou modèles structuraux) à condition que le format des fichiers soit le format PDB. Son but principal est la synchronisation des différentes bases MED-SuMo. La commande suivante permet la construction et la mise à jour de l'index PDB :

```
# sumo pdb update $path
```

*\$path* étant le chemin du répertoire contenant les fichiers PDB.

Des annotations sur les structures de chaque fichier PDB peuvent aussi être importés dans cet index. Au départ, seules des annotations extraites du fichier PDB étaient stockables, il est maintenant possible d'importer d'autres annotations stockées alors dans des colonnes supplémentaires de cette table (cf. III.ii). Ces annotations sont lues ensuite par le client graphique et peuvent être affichées dans le tableau de résultats. Il est aussi possible de créer un *pdb\_index.db* ne référençant qu'un petit nombre de structure et non la PDB entière.

En termes de procédure de mise à jour, il est d'abord nécessaire d'utiliser le script de mise à jour délivré par la PDB afin de télécharger les fichiers dans un répertoire local du serveur Linux. Il faut ensuite lancer la mise à jour de l'index de la PDB, et ensuite mettre à jour les bases de graphes décrites ci-dessous.

### ii. *La base de sites*

Un site de liaison est défini ici par la présence d'un ligand co-cristallisé avec la structure protéique. Plusieurs ligands peuvent être présents pour une même protéine, alors que certaines sont résolues sans ligands. Les bases de sites sont caractérisées par trois paramètres essentiels: la taille de l'environnement considéré protéine autour de chaque atome du ligand (paramètre *ligand\_radius*), 4,5 Å et 6,0 Å sont les rayons les plus communément utilisés (4.5 Å permet de définir clairement les modes de liaisons des ligands. 6.0 Å permet une analyse plus large de la surface du site actif). Les paramètres de triplets {*edge\_max* et *max\_edge\_sum*} fixent respectivement des limites sur la distance maximale entre deux SCFs pour qu'ils forment une arête d'un triplet et le périmètre maximal d'un triplet. En fonction des valeurs fixées, en

---

général {13 Å -39 Å} et {20 Å -60 Å}, le nombre de triplets créés est plus ou moins élevé, les graphes formés sont donc plus ou moins denses. Dans les bases sites, aucune limite sur le taux d'enfouissement des SCFs n'est appliquée, tous les SCFs autour du ligand considéré sont utilisés pour construire le graphe.

### iii. *La base de surface entière : la base Full*

MED-SuMo définit une surface d'interaction avec les SCFs présents à la surface de la protéine. Le calcul de la densité atomique de chaque atome permet d'éliminer les SCFs trop enfouis (cf. [8]). Les SCFs restant sont les groupements chimiques accessibles aux partenaires de la protéine étudiée. Comme pour la base de sites, la taille des triplets est fixée par les deux paramètres : *edge\_max* et *max\_edge\_sum*. Pour la base full, les valeurs {13 Å -39 Å} sont les seuls utilisés, la base SQLite associée faisant déjà environ 400G octets. Les commandes pour construire et mettre à jour les bases en fonction de l'index PDB sont :

```
# sumo graphdb update sites
# sumo graphdb update full
```

Les bases *sites* et *full* de MED-SuMo sont régulièrement mise à jour, elles contiennent, respectivement, tous les sites de liaison de toutes les structures de la PDB, soit au 27/01/09, 193 502 graphes *sites* et 53 413 graphes *full* pour 55 410 structures.

### iv. *Plusieurs modes d'utilisation*

L'accès à différents types de requêtes et à différentes bases permet de diversifier les applications du logiciel (cf. Figure 19). La comparaison de sites de liaison permet d'identifier dans certains cas des corrélations potentielles entre la contribution de groupements chimiques du site actif de la requête et les affinités et sélectivités thermodynamiques observées de ses ligands. Parce que MED-SuMo superpose non seulement les protéines de la PDB mais aussi les ligands co-cristallisés de chacun de ces hits en appliquant la même matrice de rotation/translation, la comparaison de sites de liaison permet aussi l'identification de nouvelles molécules candidates à partir de ces ligands co-cristallisés provenant des hits MED-SuMo. Le « *Drug repurposing* », se base sur le principe qu'une molécule active peut interagir avec plusieurs protéines cibles. Cette application est très appréciée par l'industrie pharmaceutique car elle permet le recyclage de molécules déjà classées comme « *drug-like* »

mais mises hors parcours à cause d'effets secondaires ou d'autres problèmes qu'elles peuvent générer sur les protéines cibles.

Certaines protéines cibles sont co-cristallisées avec les molécules actives qui vont inhiber ou activer leurs mécanismes. Ces ligands caractérisent pour MED-SuMo des surfaces d'interactions intéressantes à exploiter.

Pour ces protéines, l'approche de MED-SuMo est très intéressante car si des sites de liaisons très semblables sont détectés dans la PDB, il est probable que la molécule interagisse aussi avec eux. La comparaison d'une surface entière de protéine à une base de sites de liaisons permet ainsi de localiser un site d'interaction et éventuellement d'apporter une annotation fonctionnelle sur la protéine requête. Cette utilisation de MED-SuMo est au centre d'une partie de mes travaux de thèse exposés dans la partie I.A.2.





Base de données Requête	Site 	Full 
Site 	<ul style="list-style-type: none"> <li>• Caractérisation de sites de liaison</li> <li>• Prédiction d'interactions ligands-protéine</li> <li>• Caractérisation de nouvelles cibles pour une molécule active</li> </ul>	<ul style="list-style-type: none"> <li>• Meilleure caractérisation des sites</li> </ul>
Full 	<ul style="list-style-type: none"> <li>• Détection de sites</li> <li>• Application pour de l'annotation fonctionnelle 3D</li> </ul>	<ul style="list-style-type: none"> <li>• Caractérisation des interfaces protéine-protéine</li> <li>• Comparaison de templates issus de la modélisation par homologie aux structures de la PDB.</li> </ul>

Figure 19 : Les différents modes d'utilisation du logiciel MED-SuMo et leurs diverses applications.

### 3. Le score MED-SuMo

Le score MED-SuMo est la somme de l'influence de chaque SCFs des deux patches similaires détectés par MED-SuMo. Un patch est un ensemble de SCFs communs entre deux surfaces d'interaction. L'influence d'un SCF dépend de la « densité locale atomique » (cf. [8], page 82), son environnement. Plus un SCFs est éloigné d'autres SCFs, plus son score est élevé. En revanche, si un SCF est au milieu de plusieurs autres SCFs, son score sera diminué. Cette propriété se quantifie par les deux propriétés suivantes :

- $w$  est le poids de chaque SCF comme il est défini dans le fichier *sumo\_groups* (dictionnaire contenant la définition de tous les types de SCFs et qui permet leur positionnement sur les résidus d'une protéine). Par

---

exemple, un donneur/accepteur de liaison hydrogène à un poids de 0,6, une charge a un poids de 0,75. Ces valeurs ont été calibrées au moment de la validation de l'approche dans l'article [9].

- $m$  est le score du SCF, c'est-à-dire le poids original  $w$  modéré par une fonction d'influence. La fonction d'influence pondère le poids d'un SCF en mesurant sa proximité avec les SCFs de son environnement. Cette fonction vaut 1 lorsque deux SCFs sont très proches (si leur distance tend vers une distance nulle), l'influence du second SCF est donc maximale. Elle tend vers 0 si les SCFs assez éloignés, aucune influence n'est alors exercée par le second SCF [8].

Le score d'un SCF se calcule avec la fonction suivante :

$$m_i = w_i \times f(i, env_i) \quad (3)$$

Le score d'un patch se calcule avec la fonction suivant :

$$score_p = \sum_{i=1}^n m_i \quad (4)$$

Le score pour un hit dans les résultats de MED-SuMo est la moyenne des scores des deux patches similaires ( $p_1$  et  $p_2$ ) détectés, soit :

$$MEDSuMo\_score = \frac{score_{p1} + score_{p2}}{2} \quad (5)$$

Cette partie décrit le logiciel MED-SuMo tel qu'il était au début de ma thèse (décembre 2005). Au cours de ces trois dernières années, j'ai essentiellement travaillé sur la partie serveur de MED-SuMo sur lequel j'ai pu apporter certaines optimisations notamment sur la base *sites* de MED-SuMo ou sur les communications client-serveur dans le cadre du savoir-faire mis en œuvre par MEDIT. Ces optimisations sont décrites dans la partie I.A. La partie la plus importante de mon travail est la nouvelle approche MED-SuMo\_Multi de comparaison multiple et de classification. Ce travail avait été initié pour SuMo mais n'était pas fonctionnel. MEDIT avait décidé de revoir cette partie. Nous avons donc entièrement développé et implémenté une nouvelle méthode qui repose sur la même comparaison multiple mais dont

---

l'interprétation des résultats est différente. L'approche MED-SuMo Multi (MED-SMA) permet maintenant de classer tous jeux de sites de liaison, nous finissons d'ailleurs tout juste à MEDIT d'adapter le protocole aux grands jeux de données dans le but de classer tous les sites de liaisons de la PDB (grand challenge que nous faisons dans le cadre du projet POPS financé par le pôle de compétitivité SYSTEM@TIC Paris-Région [145,146]). Ce protocole, ainsi que des applications sur deux jeux de données plus réduites sont décrits dans la partie III.B. La dernière partie de ce manuscrit concernera une toute nouvelle approche de MED-SuMo permettant la conception *de novo* de molécules se fixant à une protéine cible donnée. Ma participation a concerné l'ajout de fonctionnalités sur le serveur MED-SuMo pour qu'il communique avec le programme de fragmentation implémenté à MEDIT et aussi la création d'un nouveau type de base pour MED-SuMo la base *fragments*. L'ensemble de ces travaux ont été réalisé avec le savoir faire mis en œuvre par MEDIT.

PARTIE III : MED-SuMo : optimisations et  
nouveaux développements

---

### III. MED-SuMo : optimisations et nouveaux développements

#### A. MED-SuMo serveur

##### 1. Nouvelles fonctionnalités

Les développements du serveur de MED-SuMo auxquels j'ai contribué, dans le savoir-faire mis en œuvre par MEDIT, ont permis des améliorations notables et l'ajout de certaines fonctionnalités décrites dans cette partie.

###### i. La base de sites

###### Définition des ligands

La définition des ligands de SuMo concernait majoritairement les molécules de taille inférieure à 100 atomes et dont tous les atomes étaient caractérisés par le champ HETATM dans le fichier PDB. Toute molécule contenant un acide aminé était donc exclue, tels les hétéropeptides et les homopeptides. Les ligands covalents n'étaient par ailleurs pas inclus non plus.

Dans le but d'enrichir les bases de sites de MED-SuMo, les caractéristiques des ligands ont été modifiées et les deux dernières des trois règles suivantes ont été ajoutées. (1) Un ligand est une molécule de moins de 100 atomes (décrits dans le fichier PDB). (2) Si une protéine contient une chaîne d'au moins 100 résidus, toute autre chaîne de moins de 10 résidus est considérée comme un ligand. (3) Si un atome est caractérisé par le mot HETATM dans le fichier PDB, le résidu dont il fait partie est considéré comme ligand même s'il est composé de moins de 100 atomes et s'il est lié de manière covalente au reste de la protéine. Ainsi avec une version récente de la PDB qui contient 55 410 structures, 121 110 sites étaient auparavant détectés, alors que 157 320 sites le sont avec les nouvelles règles.

###### Analyse des redondances de la PDB

La PDB est une base de données de dépôt des structures protéiques résolues. Certaines protéines y sont représentées plusieurs fois, avec des résolutions différentes ou co-cristallisées avec des ligands différents. À titre d'information, l'alpha thrombine humaine était présente dans 232 enregistrements de la PDB en janvier 2009. Par exemple, les structures 1DWC, 1DWD et 1DWE [147] contiennent chacune l'alpha thrombine humaine résolues par le même auteur et déposées le même jour. Elles sont toutefois co-cristallisées avec des inhibiteurs



---

différents. La PDB est donc une base redondante. Pour beaucoup d'approches, seul un représentant de cette protéine serait utile. Cependant pour MED-SuMo, cette redondance enrichit la base de surfaces d'interaction de deux manières. La première concerne les variations conformationnelles observées pour une même protéine. La seconde touche la diversité de définition d'un même site actif caractérisée par les ligands co-cristallisés. Lorsqu'un hit est détecté par MED-SuMo, son ligand peut être superposé à la structure de la protéine contenant le site requête permettant ainsi l'analyse d'interaction possible entre la requête et un autre ligand

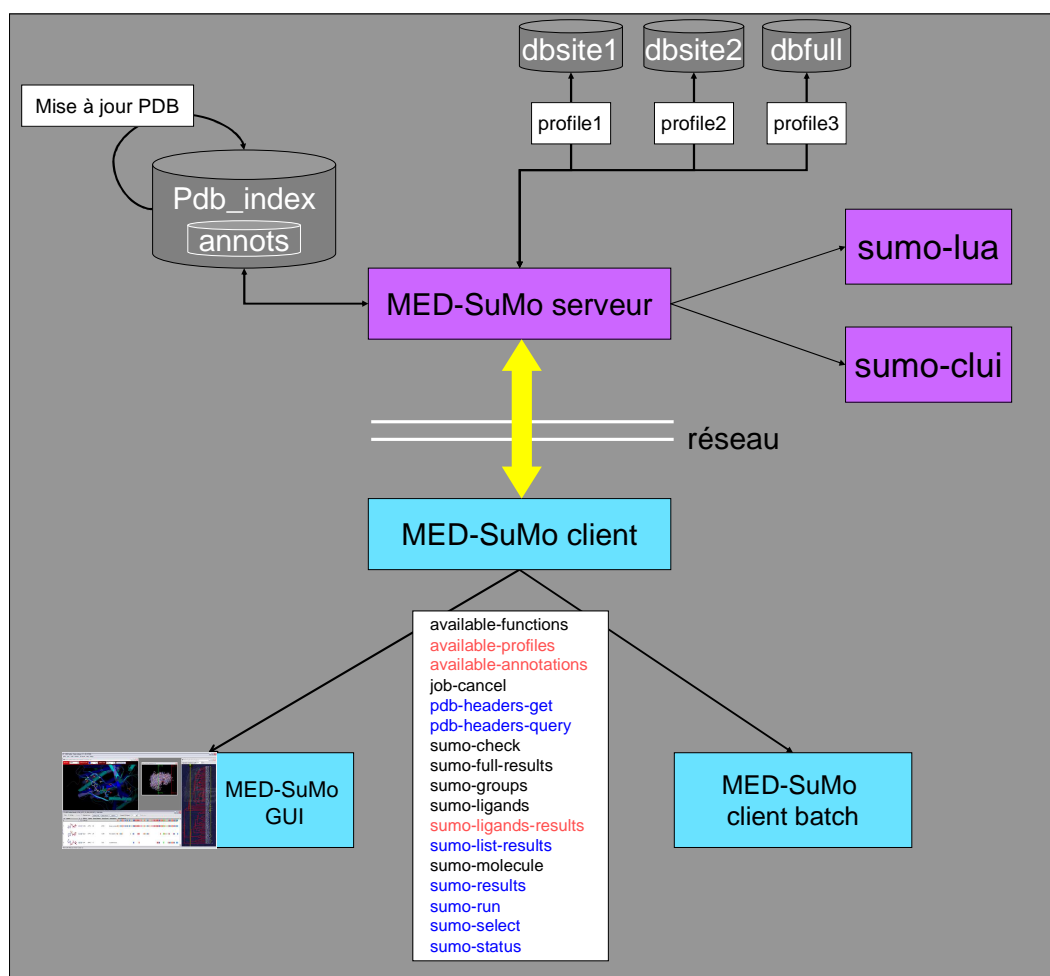
Un autre type de redondance se retrouve dans la PDB. Certaines structures contiennent plusieurs fois la même chaîne. Initialement, les chaînes redondantes d'une même structure étaient détectées en analysant leurs séquences. Si des chaînes redondantes étaient présentes, seuls les graphes des ligands d'une seule des chaînes redondantes étaient stockés dans la base de sites. Cette règle permettait de limiter la taille de la base de sites. Cependant, même si une chaîne apparaît plusieurs fois dans une structure, il se peut que les ligands qui y sont fixés, soient différents. Par exemple, la stromélysine, une enzyme primordiale de destruction de la matrice extracellulaire, est un homodimère (code PDB: 1HY7 [148]). Le ligand MBS est fixé sur la chaîne B de la structure résolue. La règle précédente ne considère naturellement pas la chaîne B et comme aucun ligand n'est fixé sur la chaîne A, aucun site de liaison est stocké pour cette structure. Une règle supplémentaire a donc été ajoutée. Lorsque des chaînes redondantes sont détectées dans une même structure, les sites de liaisons de la chaîne redondante sont exclus de la base de sites seulement si les ligands qui y sont fixés sont identiques. Avec la même version de la PDB (55 410 structures), la prise en compte de cette information permet de passer d'une base contenant 157 320 sites, à 193 502 sites. Cette base est celle utilisée actuellement.

Ces deux modifications permettent à la base de sites de MED-SuMo d'être plus exhaustive et donc plus représentative des richesses structurales de la PDB.

## ii. *Communication client-serveur*

Le serveur gère l'essentiel des fonctionnalités de MED-SuMo et définit, à l'aide de requêtes précises, les communications client-serveur (au travers du programme *sumo.client.exe*, cf. III.A.1.ii). Le développement de la nouvelle interface graphique a impliqué la création de nouvelles requêtes. La Figure 20 illustre l'architecture globale de

MED-SuMo. Le bas de la figure décrit la partie client de MED-SuMo qui communique au serveur via le réseau. La liste de fonctions en bas au centre constitue la liste des requêtes MED-sumo-client disponibles. Toutes sont implémentées pour le MED-SuMo GUI, alors que les bleues sont aussi utilisées au client *batch*. Les rouges sont les plus récentes et elles sont précisément décrites dans le paragraphe suivant. Une description complète de toutes les requêtes MED-sumo-client est disponible en Annexe 3.



**Figure 20 : Architecture globale de MED-SuMo**

La partie haute représente le côté serveur, les interactions avec l'index PDB, les profils utilisateurs ou les interfaces locales. La partie basse représente le côté client composé du programme en ligne de commande sumo-client-batch.exe et le client graphique. Les requêtes disponibles sont colorées: en rouges pour les nouvelles, en bleues pour celles communes aux deux interfaces client et en noir pour celle exclusivement utilisées par le GUI.

### Description des trois nouvelles requêtes

#### **Gestion des profils utilisateurs (available-profiles)**

Une amélioration notable du serveur de MED-SuMo a été l'intégration de profils pour les utilisateurs. Cette nouvelle fonctionnalité permet à un serveur MED-SuMo-serveur unique,

---

lancé pour écouter un port unique, de gérer plusieurs types de bases MED-SuMo en même temps. Par exemple, le profil « *db60\_1339* » accède à la base dont les paramètres sont: [*ligand\_radius* = 6,0 Å, *edge\_max* = 13 Å, *max\_edge\_sum* = 39 Å] et le profil « *db45\_2060* » accède à une autre base dont les paramètres sont: [*ligand\_radius* = 4,5 Å, *edge\_max* = 20 Å, *max\_edge\_sum* = 60 Å]. Comme expliqué dans la section.II.D.2.iv, ces bases sont utilisées pour des applications différentes de MED-SuMo. Certains profils peuvent aussi n'être accessibles qu'à un nombre réduit d'utilisateurs, permettant par exemple de limiter l'accès à la base *full* qui génère des calculs plus lourds pour le serveur. La requête *available-profiles* est utilisée par l'interface graphique pour récupérer la liste des profils disponibles au moment de la connexion au serveur MED-SuMo. L'utilisateur doit alors s'authentifier et choisir le profil de la base avec lequel il veut travailler.

### **Prise en compte des annotations (available-annotations)**

Les fichiers de la PDB sont indispensables au fonctionnement de MED-SuMo. Ils doivent en effet constamment être accessibles au serveur. Ces fichiers sont référencés dans la base *pdb\_index.db* représentée en haut à gauche dans la Figure 20. Au moment de sa création, des annotations issues de divers champs du fichier PDB, tels que le *HEADER*, le type de méthode de résolution, le titre qui contient une brève description de la structure, sont extraites et stockées dans la base *pdb\_index.db*. Cette table d'annotations a été améliorée pour inclure des informations supplémentaires. Des annotations autres que celles extraites des fichiers PDB peuvent être ajoutées à la base. Pour ce faire, il faut tout d'abord créer un fichier texte contenant les annotations: pour chaque structure, une ligne ayant le format: ID\_PDB, ID\_ANNOTATION, ANNOTATION. Ensuite, les annotations sont ajoutées à la base *pdb\_index.db* en lançant la commande :

```
# sumo pdb annot ID_FICHER
```

Il est, par exemple, possible, de stocker des annotations UNIPROT [2] donnant accès à des informations sur le gène dont la structure de la protéine est issue. Ces annotations sont particulièrement utiles pour les utilisateurs de l'interface graphique car elles sont disponibles dans les colonnes du tableau de résultats (cf. Figure 16). La requête *available-annotations* est utilisée par l'interface graphique au moment de l'affichage des résultats d'un calcul MED-SuMo afin de visualiser dans des colonnes séparées, toutes les annotations présentes dans la base *pdb\_index.db*.

---

## Renvoi des informations sur les ligands des hits (sumo-ligands-results)

Lorsque la base de sites est sollicitée, une projection en deux dimensions des ligands des *hits* est calculée par une routine de conversion de structures 3D en structures 2D afin d'être affichée dans le tableau de résultats de l'interface graphique. Cette représentation permet d'avoir une idée rapide sur le type de ligand qui se fixe dans les résultats (*hits*). La routine nécessite la liste des atomes des ligands de tous les *hits* et leurs coordonnées. Cette liste est envoyée à l'interface graphique par la requête *sumo-ligand-results*.

La partie supérieure de la Figure 20 décrit les interactions du serveur MED-SuMo. Le serveur peut donc gérer différentes bases grâce aux profils utilisateurs. Il interagit avec la base *pdb\_index.db* afin de récupérer diverses annotations bien que son utilité principale soit de permettre la synchronisation des bases MED-SuMo en fonction des actualisations de la *Protein Data Bank*. Toutes les semaines, des structures sont éliminées, d'autres sont corrigées alors que de nombreuses nouvelles sont insérées. L'index PDB doit dans un premier temps être actualisé (cf. partie II.D.2.i). Une commande permet ensuite la mise à jour des bases (cf. partie III.D.2.ii). L'interfaçage avec le langage de script Lua [149] est aussi représentée sur la Figure 20.

Les optimisations et nouveautés du logiciel MED-SuMo permettent son utilisation dans un large choix d'applications de modélisation moléculaire. Une revue décrivant les applications possible a récemment été acceptée [14] (cf. article 4). L'utilisation de scripts shell (MED-sumo-client) et de scripts Lua (MED-sumo-lua) a permis d'adapter le logiciel pour des applications sur des grands jeux de données. En effet, en 2005, plus de 2 000 surfaces entières de protéines hypothétiques de la PDB avaient été comparées à la base de sites. Nous avons pris un exemple représentatif pour montrer l'intérêt de MED-SuMo pour fonctionnellement annoter des structures protéiques [13] (cf. article 1).

## 2. Annotation fonctionnelle avec MED-SuMo

Une structure sans fonction connue est une protéine qui a été résolue sans que le rôle du gène dont elle est issue ne soit déterminé. Ces structures sont de plus en plus nombreuses dans la PDB. Les consortiums de génomique structurale ont pour objectif de résoudre le plus grand nombre de structures possible, sachant que pour certaine familles, la connaissance de la fonction de leur gène n'est pas toujours connue. Ces structures résolues sont stockées dans la PDB et identifiées par les annotations « *Structural Genomics Unknown function* » ou

---

« *Hypothetical Protein* ». Une requête avec ces deux expressions clés sur le site de la PDB [5] permet de recueillir plus de 3000 structures. Les chercheurs de ces consortiums espèrent entre autre pouvoir dériver la fonction de ces protéines par comparaison de leur repliement à ceux déjà connus. La disponibilité de leurs structures particulières est une première étape et permet la recherche d'indices concernant les mécanismes dans lesquels ces protéines sont impliquées. Les méthodes de modélisation moléculaire 3D les plus puissantes pour l'annotation fonctionnelle de structures sont celles qui se basent sur la comparaison des surfaces des protéines. En effet, la fonction d'une protéine ne s'exprime qu'à partir du moment où elle interagit avec un ou plusieurs partenaires. L'avantage majeur de MED-SuMo est sa capacité à exploiter la richesse de la PDB. Aujourd'hui, de nombreuses protéines dont le rôle est connu sont co-cristallisées plusieurs fois avec des ligands différents. Ces structures fournissent des informations capitales et permettent une analyse poussée de leurs différents modes de liaisons. Le protocole MED-SuMo où la surface totale de la protéine hypothétique est comparée à la base de surfaces d'interaction de site actif, est particulièrement adapté pour annoter fonctionnellement les structures de protéines, et cela pour plusieurs raisons :

- (1) MED-SuMo caractérise chaque association protéine-ligand de la PDB par un graphe dans la base *sites*.
- (2) MED-SuMo permet la soumission de toute sorte de requête, ce qui inclut la possibilité de comparer toute la surface d'une protéine à toutes les associations protéine-ligand de la PDB (cf. Figure 18, milieu).
- (3) L'approche de MED-SuMo est originale et détecte les similitudes structurales et fonctionnelles à la surface des protéines. La détection de *hits* signifie qu'une signature commune en termes de SCFs avec celle de la requête a été trouvée. Des groupements chimiques à la surface des deux protéines sont donc disposés et orientés de la même manière. Ces similitudes peuvent laisser supposer que les protéines en question vont interagir de la même manière et donc avoir des fonctions biochimiques similaires.

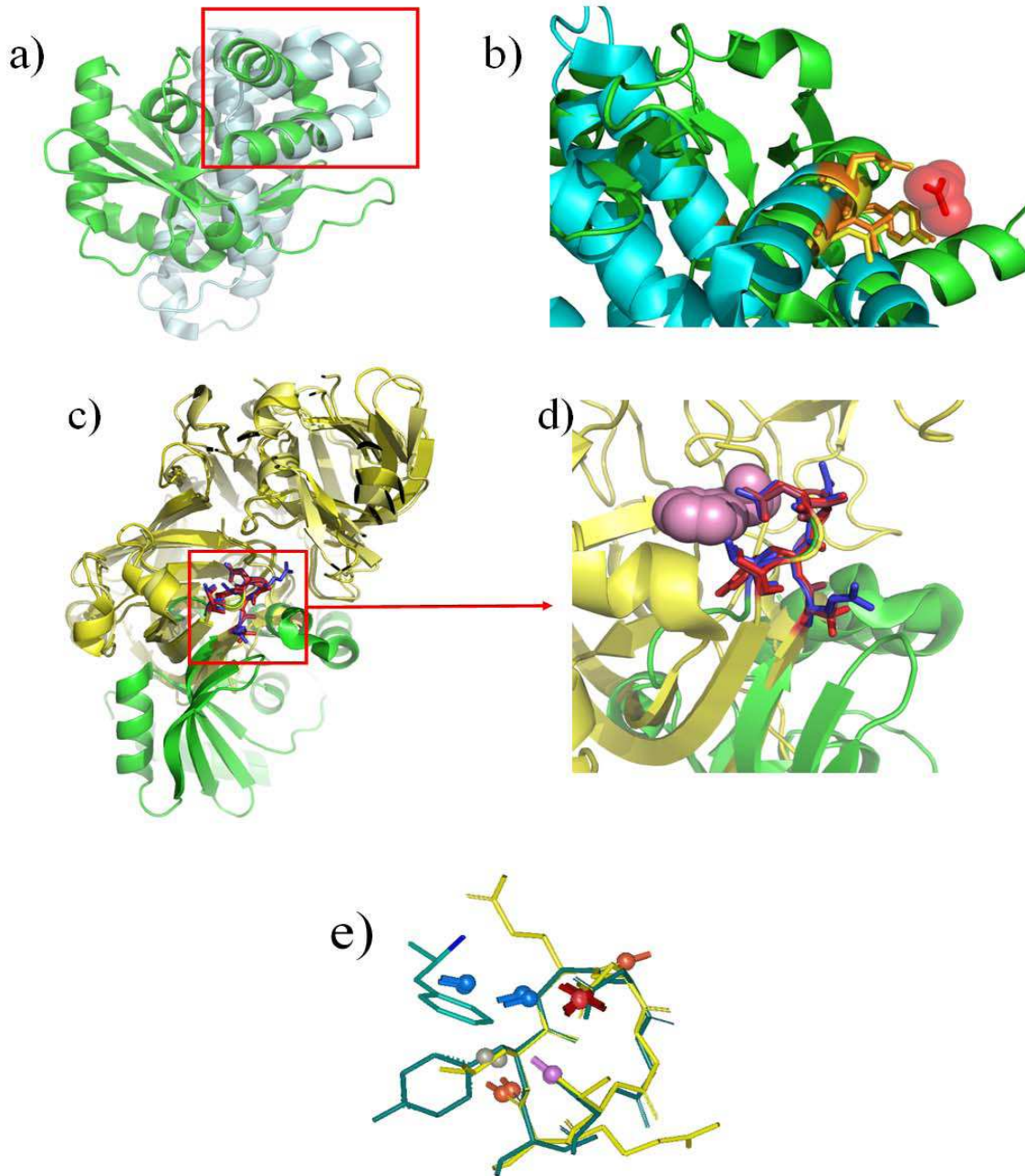
Deux applications sont présentées ici : les protéines TM1012 (code PDB 2EWR [www.jcsg.org](http://www.jcsg.org)) et YBL036C (code PDB 1B54 [www.nysgxrc.org](http://www.nysgxrc.org)) font parties des 3 000 considérées sans fonction connue. Il convient de noter que cette quantité est probablement légèrement surestimée, sachant qu'un certain nombre de ces protéines dispose de séquences homologues *a priori* aisément identifiables.

---

## Protéine TM1012

La structure de la protéine TM1012 a été résolue par le consortium « *Joint Center for Structural Genomics* » (JCSG, <http://www.jcsg.org/>) et déposée en novembre 2005. Ce consortium a résolu plus de 750 structures et, bien que plus de la moitié n'aient pas de fonction connue, la plupart ont des similitudes de séquences ou de structures avec des protéines déjà connues. La protéine TM1012 est une protéine hypothétique de *Thermotoga maritime* (code PDB: 2EWR). Elle ne peut être associée à aucune protéine connue. Les approches classiques comme PSI-BLAST [21] lancées sur la base de données de séquence non redondante (via le web service NCBI), et le serveur ProFunc [133] ne permettent pas d'identifier des séquences d'un gène connu semblables, ni de localiser un ensemble de résidus impliqués dans un profil particulier.

Nous avons donc utilisé la surface de la protéine TM1012 comme requête que nous avons soumise à la base des sites de MED-SuMo. Les paramètres de la base utilisée sont  $\{ligand\_radius = 6 \text{ \AA}; edge\_max = 13 \text{ \AA}; max\_edge\_sum = 39 \text{ \AA}\}$ . Les trois meilleurs *hits* correspondent à la même protéine, toujours résolue par le JCSG, mais dans des conditions différentes (par exemple, code PDB 2FCL). Ensuite, les *hits* 2CJ5, 5APR et 1OD1 ne sont pas directement liés à la requête. Les structures ne se superposent pas et ont de très faibles taux d'identité de séquence avec la requête. Les protéines 5APR et 1OD1 ont un taux d'identité de séquences de 38%, et leurs structures se superposent globalement avec un RMSD de 1,8 Å. La protéine 2CJ5 est distincte des deux autres avec des taux d'identité de séquence de 22% et les structures qui ne se superposent que sur 20% de leurs structures.



**Figure 21: Exemples de résultats obtenus par MED-SuMo sur la protéine hypothétique TM1012 de l'espèce *Thermotoga maritime* (code PDB 2EWR en vert).**

a) Superposition globale des protéines TM1012 et l'inhibiteur invertase du mur cellulaire de la plante *Nicotiana tobacum* (code PDB 2CJ5 en bleu clair). b) Vue rapprochée des résidus superposés par MED-SuMo. c) Superposition de TM1012 et la protéine rhizopuspepsin (code PDB 5APR en jaune). d) Vue rapprochée avec le ligand statine. e) Superposition de TM1012 et du site de liaison de la protéine endothiapepsin (Code PDB 1OD1). Les SCFs y sont aussi représentés. Les molécules sont représentées à l'aide du logiciel PyMol [43] et de MED-SuMo GUI.

La Figure 21 montre deux régions d'intérêt sur la protéine TM1012 identifiées par MED-SuMo. La première implique les résidus 134, 135 et 138 qui se superposent aux résidus

---

17, 18 et 31 de la protéine 2CJ5 (« inhibiteur invertase du mur cellulaire » de la plante *Nicotiana tobacum*). Si la Figure 21a souligne des repliements différents, la Figure 21b affiche une vue rapprochée de la superposition locale des résidus, ainsi que du ligand, un ion acétate ( $\text{CH}_3\text{COO}^-$ ). Le RMSD local est de 0,5 Å et les résidus des deux régions sont identiques (YQ—L). L'ion acétate représentant plus un artefact de cristallographie, cette superposition présente un intérêt limité pour détecter une fonction biologique

La seconde région présentant une surface d'interaction similaire implique un nombre d'acides aminés plus important. Les Figure 21c et d montre une superposition de TM1012 et la protéine rhizopuspepsin, (code PDB 5APR). Les résidus  $\text{LS}^{76}\text{Y}^{77}\text{G}^{78}\text{D}^{79}\text{-S}^{81}$  de 5APR et  $\text{R}^{99,100}\text{E}^{101}\text{D}^{102}\text{-T}^{104}$  de TM1012 sont superposés. Les mêmes résidus de la protéine 1OD1 se superposent (cf. Figure 21e). Le RMSD local sur ces résidus est faible malgré le fait que les résidus impliqués soient différents. De manière intéressante, seul l'aspartate est commun, alors que neuf SCFs des sites de liaison de 1OD1 et 5APR se superposent avec ceux de la requête 2EWR. En outre, l'analyse phylogénétique de la séquence du motif du site de liaison de la rhizopuspepsin montre que 3 résidus sur 5 sont conservés, alors que 8 SCFs sur 9 le sont, ce qui montre l'intérêt d'une approche comme MED-SuMo.

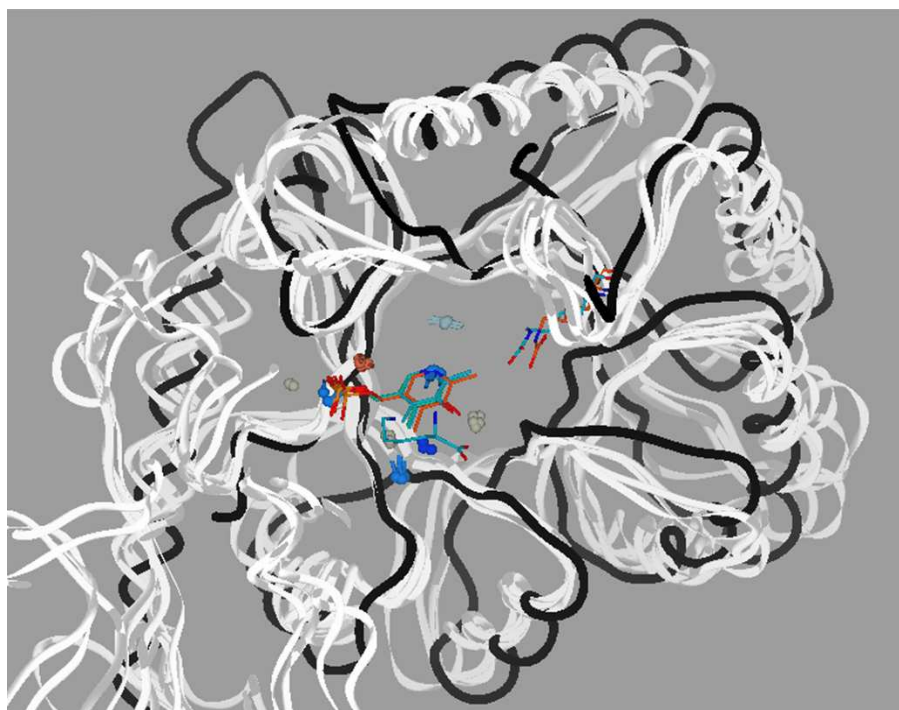
Des analyses complémentaires sont naturellement requises pour confirmer une fonction réelle à cette protéine. Toutefois, il est intéressant de noter que lorsque les approches d'étude classique des protéines ne donnent aucun résultat exploitable, MED-SuMo permet d'initier un processus de détermination de fonction en localisant un jeu de résidus contenant des groupements chimiques positionnés similairement à une surface d'interaction connue et pouvant interagir ainsi avec un ligand. La protéine *TM1012* pourrait donc interagir par un mécanisme plus ou moins similaire à celui de la protéine rhizopuspepsin.

### Protéine YBL036C

La protéine YBL036C (code PDB 1B54) a été résolue par le consortium NYSGXRC (<http://www.nysgxrc.org/>) et déposée en janvier 1999. Son repliement est proche d'un *fold* connu pour sa diversité fonctionnelle: le célèbre *fold TIM-barrel* [85]. Contrairement à la protéine TM1012, quelques éléments sur la fonction de YBL036C sont disponibles grâce à la méthode PSI-BLAST et au serveur ProFunc. Les 18 premiers *hits* de PSI-BLAST sont des séquences de protéines non caractérisées. Les 19, 23, 31 et 39<sup>ème</sup> rang sont des séquences issues de la famille des alanines racémases. Leurs taux d'identité de séquence sont inférieurs à 44%. D'autres séquences intéressantes de la famille des prolines synthétase sont aux 35, 42, 44, 50 et 52<sup>ème</sup> et ont des taux d'identité de séquence inférieurs à 46%. Ces taux d'identité de



séquence sont suffisants pour déduire que les séquences ont une origine commune mais pas pour attribuer une fonction à la protéine d'une manière définitive [150]. Le serveur ProFunc permet, entre autre, de comparer la séquence d'une protéine aux différentes bases PROSITE [116] et PFAM [117]. Un *hit* de la base PFAM est détecté impliquant le domaine PF01168 qui représente la famille des alanines racémases. La fonction de cette protéine est donc probablement liée à cette famille. La structure d'une protéine étant plus conservée que sa séquence, MED-SuMo permet un autre type d'analyse sur cette protéine. Une requête considérant la surface entière de la protéine est construite; 664 SCFs sont détectés pour former les 12 949 triplets d'un graphe qui est comparé à la base de sites de MED-SuMo. Ici aussi les paramètres de la base utilisée sont  $\{ligand\_radius = 6 \text{ \AA}; edge\_max = 13 \text{ \AA}; max\_edge\_sum = 39 \text{ \AA}\}$ . Les quatre meilleurs résultats sont des protéines hypothétiques. Le sixième est la structure d'une alanine racémase. La possibilité de classer les résultats en fonction de la similarité des signatures en SCFs observées permet de retrouver aisément les autres alanines racémases détectées par MED-SuMo (codes PDB : 1XFC, 1VFH, 1SFT). La Figure 22 représente la superposition de la protéine hypothétique YBL036C avec les alanines racémases détectées par MED-SuMo et montre clairement qu'il pourrait s'agir d'une protéine de cette famille.



**Figure 22 : Superposition de la protéine YBL036C avec les alanines racémases détectées par MED-SuMo**  
(code PDB 1XFC, 1VFH, 1SFT en blanc et 1B54 en noir).

---

Les résultats de PSI-BLAST, de PFAM et de MED-SuMo vont donc dans le même sens. La fonction de la protéine YBL036C est probablement liée aux alanines racémases qui sont des protéines qui catalysent l'interconversion des énantiomères alanines. Cette hypothèse nécessite naturellement d'être validée expérimentalement.

Ces deux applications démontrent l'intérêt de MED-SuMo dans l'annotation fonctionnelle de protéines. En effet, pour la protéine TM1012 alors que les approches d'études classiques des protéines ne donnent aucun résultat exploitable, MED-SuMo permet d'initier un processus de détermination de fonction en détectant des surfaces d'interactions similaires à d'autres sites de liaison dont la fonction est définie. Pour la protéine YBL036C, la situation est différente. La protéine est annotée dans la PDB comme « hypothétique », cependant, PSI-BLAST détecte des alanines racémases ayant un faible taux d'identité de séquence avec la protéine alors que des alanines racémases sont aussi détectées par le serveur Profunc (notamment par la méthode de comparaison de structure DALI). Ici MED-SuMo confirme ce que les autres méthodes détectent en localisant quatre alanines racémases dont les sites de liaison se superposent parfaitement avec celui de la requête (cf. Figure 22). L'intérêt de MED-SuMo s'illustre donc ici. Une application à grande échelle sur toutes les protéines hypothétiques de la PDB avait été tentée au début de ma thèse, mais l'analyse avait dû être arrêtée pour pouvoir libérer du temps pour de nouveaux sujets. Une des limitations était la difficulté de gérer les résultats, la nouvelle interface graphique permettrait éventuellement de la faciliter aujourd'hui, et ensuite de valider expérimentalement ces hypothèses (résultats publiés dans [14], cf. article 4).

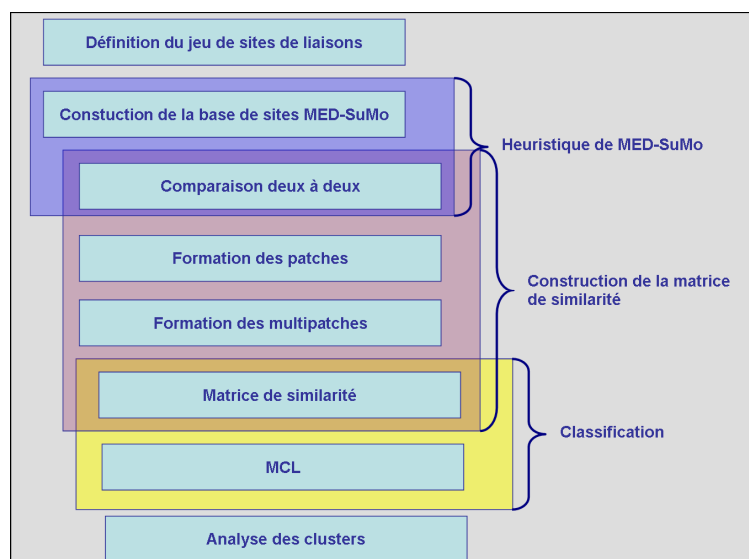
---

## **B. Une nouvelle méthode performante de classification des surfaces protéiques d'interaction: MED-SMA**

Lorsqu'une méthode de comparaison structurale est efficace, établir une classification se basant sur celle-ci est un prolongement logique de son approche. Ainsi, la classification FSSP [46,51] (cf. I.iii) résulte des alignements de DALI [49]. Une méthode basée sur Cavbase a aussi été utilisée pour procéder à la classification d'ensembles de sites de liaison [151,152]. L'approche de MED-SuMo est originale et rapide. Son intérêt à détecter des similitudes structurales et fonctionnelles entre surfaces d'interaction a été démontré à travers plusieurs applications [8-10,13]. De ce fait, nous avons décidé d'étendre ses capacités à la comparaison multiple d'un ensemble de sites de liaison dans le but de classer des jeux de sites. La méthode permettant cette classification est décrite dans les deux parties suivantes: deux applications qui ont abouties à l'écriture de deux publications sont présentées (article 2 [15] et article 5 [16] soumis).

### **1. Description globale**

L'approche MED-SuMo Multi (MED-SMA : *MED-SuMo Multi Approach*) est une nouvelle méthode de classification qui se base sur l'algorithme de MED-SuMo pour rassembler des sites fonctionnellement liés. Cette méthode se base sur la comparaison deux à deux de tous les sites d'une base MED-SuMo afin d'y extraire des régions similaires détectées non plus entre paires de sites mais pour un ensemble de sites. Un score calculé entre ces régions permet la construction d'une matrice de similarité qui est classifié par l'algorithme Markov CLustering (MCL) [153]. Cette nouvelle approche regroupe des surfaces d'interaction fonctionnellement liées. La Figure 23 décrit la procédure globale de la méthode, et la Figure 24 décrit en détail, les six étapes du protocole de classification.



**Figure 23 : Procédure globale de MED-SMA.**

Trois étapes majeures: (i) La comparaison multiple (rectangle bleu foncé), (ii) la construction du graphe de similarité (rectangle violet) et (iii) la classification du graphe par MCL [153] (rectangle jaune). La création du jeu de sites en amont est aussi nécessaire, ainsi que l'analyse de clusters créés.

Pour toute classification, la première étape est la constitution du jeu de données. Pour sélectionner les codes PDB à comparer, il est possible, par exemple, de classer une famille de protéines fixant un type particulier de ligands ou des protéines d'une famille extraite d'un autre type de classification tel SCOP ou PFAM. Une fois la liste de structures construite, la base de graphes peut avoir deux propriétés: (i) La base peut contenir tous les sites de liaison détectés sur les protéines choisies. D'un point de vue technique, il suffit de créer un index référençant les fichiers PDB des protéines choisies et de générer la base avec une commande MED-sumo-clui (cf. Annexe 2):

```
# sumo graphdb update sites
```

Dans ce cas, certaines règles sur les caractéristiques des ligands détectés peuvent être fixées : nombre minimum d'atomes, nombre maximum d'atomes... (ii) La base peut aussi contenir exclusivement certains sites, par exemple des sites co-cristallisés avec des ligands puriques. Pour ce faire, l'interface MED-sumo-clui sert tout d'abord à lister les ligands des sites et leurs indices respectifs. Un script MED-sumo-lua permet ensuite la construction la base en spécifiant précisément quels sont les indices des sites de liaison à inclure pour chaque structure. Une fois la base des sites créée, le protocole peut être appliqué.

MED-SuMo compare tous les sites de la base deux à deux. Un score minimal toléré entre paires de sites, définit lesquelles sont sélectionnées pour l'étape suivante, le paramètre

---

est nommé *minimal\_sumo\_score*. Ces comparaisons font ressortir les SCFs communs entre paires de sites (cf. Figure 24b) qui forment les *singlepatches* (cf. Figure 24c). Chaque site de liaison peut contenir plusieurs *singlepatches* associés à plusieurs sites. Les *singlepatches* de chaque site sont ensuite analysés un par un. Si deux *singlepatches* d'un même site partagent une quantité suffisante de SCFs, ils sont fusionnés en *multipatch* (cf. Figure 24d). La quantité de SCFs suffisante pour la fusion de deux patches est définie par le paramètre *covering\_factor*, le plus souvent fixé 0,6 (60% des SCFs de deux *singlepatches* doivent être communs pour qu'ils soient fusionnés). Le *multipatch* représente la partie globale commune unifiée entre plusieurs sites de liaison d'un jeu de données. Il assure qu'assez de SCFs sont communs, donc que les sites de liaisons représentés par chacun de ces multipatches ont des similitudes structurales et fonctionnelles suffisantes. Le score MED-SuMo calculé entre chaque paire de *multipatches* permet la construction d'une matrice de similarité (cf. Figure 24e). Finalement, MCL est utilisé pour classer la matrice et créé des groupes de sites de liaisons. La visualisation de la classification en deux dimensions est possible avec certains outils dédiés tel Biolayout [154] (cf. Figure 24f).

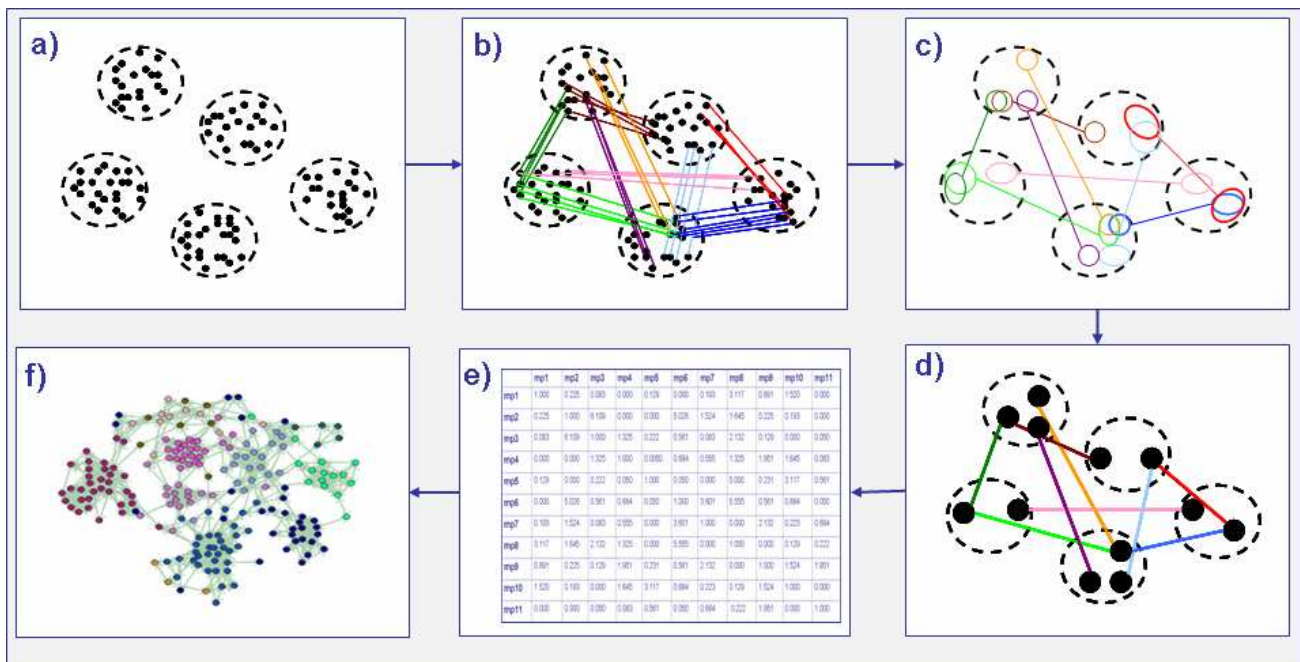


Figure 24 : Les six étapes de la méthode MED-SMA.

a) Construction du jeu de sites de liaison, cinq sites sont représentés ici. Les points noirs représentent les SCFs. b) Les SCFs en commun sont détectés grâce aux comparaisons deux à deux. Ils sont chacun reliés par une ligne colorée. Chaque couleur correspond à des sous sites semblables entre paires de sites. c) Les SCFs en commun entre paires de sites forment des *singlepatches* représentés par des cercles de couleur différente. Certains *singlepatches* se chevauchent. d) Les sites sont analysés un par un. Si un site contient des *singlepatches* suffisamment chevauchants, ils sont fusionnés en *multipatch* (représenté par des cercles noirs). e) Les scores MED-SuMo sont calculés entre chaque paire de *multipatches* pour la construction de la matrice de similarité. f) La matrice est finalement classifiée avec MCL. Le logiciel Biolayout [154] est utilisé pour visualiser les clusters créés par MED-SMA.

## 2. Détails d'implémentation

### i. Comparaison deux à deux

La première étape de MED-SMA est une comparaison multiple des sites de liaison composant la base MED-SuMo. Ces comparaisons sont effectuées par une fonction MED-sumo-lua nommée **SuMo.compare\_graph\_cross** qui retourne un résultat sous la forme d'une liste de similarités dans un fichier binaire. À l'issue de cette étape, nous savons par exemple que le site A et le site B ont un certain score MED-SuMo, ainsi que la liste des SCFs de A équivalents à ceux de B, leurs types, leurs positions... (cf. Figure 24b). Si le site A et le site C ont un score inférieur au paramètre *minimal\_sumo\_score*, aucune information concernant les paires A et C n'apparaît dans ce fichier de résultats.

---

ii. *Score MED-SuMo entre multipatches*

Le score entre deux *singlepatches* se calcule de la même manière que le score MED-SuMo pour un *hit* détecté (cf. équation 5). Pour le calcul du score entre *multipatches*, différents éléments sont à considérer. D'une part, un *multipatch* se forme par la fusion entre plusieurs *singlepatches* (cf. Figure 24c et 24d). D'autre part, la fusion de deux *singlepatches* signifie qu'ils ont plusieurs SCFs commun, il ne faut donc pas compter plusieurs fois le score des SCFs redondant dans le calcul du score du *multipatch*. Dans un *multipatch*, si un SCF est présent dans plusieurs *singlepatches*, il sera compté qu'une seule fois mais son score sera exceptionnellement égal à son poids par défaut (cf. équation 7 et la Figure 25).

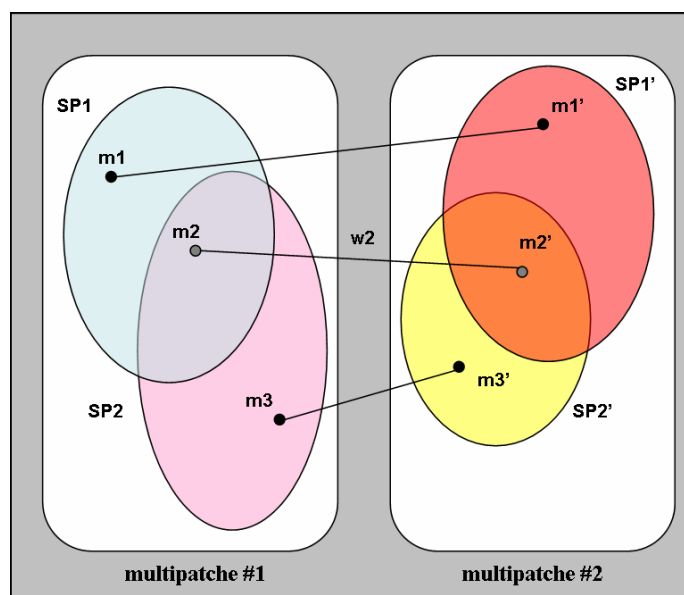
$\forall$   $SCF_i \in mp_1$  et  $SCF_j \in mp_2$  ayant chacun un score respectif de  $m_i$  et  $m_j$ , le score entre deux *multipatches* se calcule de la manière suivant :

$$score(mp_1, mp_2) = \frac{\sum_{i=1}^k m_i + \sum_{j=1}^l m_j}{2} \quad (6)$$

Le score entre les *multipatches* de la Figure 25 se calcule de la manière suivante :

$$score(mp_1, mp_2) = \frac{(m_1 + w_2 + m_3) + (m'_1 + w_2 + m'_3)}{2} \quad (7)$$

Où  $w_2$  est le poids par défaut du  $SCF_2$ .



**Figure 25 : Exemple simplifié illustrant le calcul du score entre deux *multipatches*.**

Chacun des *singlepatches* contient deux SCFs, dont un est en commun. Le calcul du score entre ces deux

$$\text{multipatches se calcule avec l'équation 7 } (score(mp_1, mp_2) = \frac{(m_1 + w_2 + m_3) + (m'_1 + w_2 + m'_3)}{2})$$

### iii. *Matrice de Similarité*

Contrairement à l'approche MED-SuMo où la requête est toujours la même, MED-SMA compare tous les graphes du jeu de données entre eux. Il faut donc prendre un certain recul quant aux scores MED-SuMo qui ne sont pas bornés comme un indice de similarité. En effet, si deux grands sites sont très similaires, ils vont avoir un score élevé, alors que si deux petits sites sont similaires, leur score sera faible, même si la similarité entre les petits sites est de meilleure qualité. Pour que les grands sites ne soient pas excessivement favorisés par rapport aux petits, nous avons décidé d'appliquer une transformation pour borner les valeurs de la matrice. Pour ce faire, chaque score de la ligne est divisé par le score obtenu entre le *multipatch* et lui-même sur la même ligne, soit le score de la diagonale de la matrice, nommé le *self\_score*. Les Figure 26a et b montre les modifications induites par la méthode de normalisation des lignes, appliquée sur une petite matrice de similarité.

Toutefois, comme présenté sur la Figure 26, une fois transformée, la matrice n'est plus symétrique, chaque élément et son symétrique représentant les valeurs min et max normalisées par l'un des deux sites sur eux-mêmes. L'algorithme MCL est moins efficace sur un graphe orienté que sur un graphe non-orienté [143]. La matrice est donc rendue symétrique. Plusieurs stratégies sont possibles et ont été testées. Pour les applications décrites dans les parties III.i et III.ii, les scores des cases  $m_{ij}$  et  $m_{ji}$ , sont remplacés par leurs valeurs maximales (cf. Figure 26).



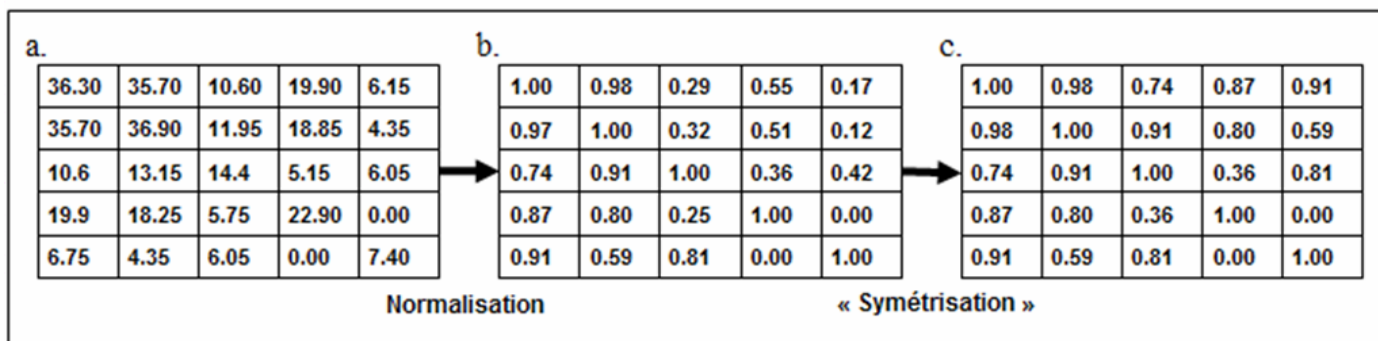


Figure 26 : Exemple des transformations appliquées à la matrice de similarité entre cinq *multipatches*.

Un des avantages de l'algorithme de classification MCL [153] est qu'il n'utilise que l'information topologique entre les éléments qu'il classe et non une information de distance à laquelle nous n'avons pas directement accès avec MED-SuMo. Il se base sur la présence ou l'absence d'un lien de similarité entre deux *multipatches*. De ce fait, la structure graphe est utilisée pour représenter la matrice de similarité, cette structure suffit pour MCL. Elle est aussi moins lourde à stocker qu'une matrice complète comportant beaucoup de valeurs non significatives (matrice contenant des zéro dès que des protéines différentes sont comparées).

#### iv. Lancement du programme

Bien que MED-SMA se déroule en six étapes consécutives, seulement deux fonctions sont nécessaires pour lancer une classification:

1. **SuMo.compare\_graph\_cross** est la fonction de comparaison multiple qui lance les comparaisons deux à deux de tous les sites d'une base MED-SuMo. Les résultats sont retournés sous forme d'un fichier binaire uniquement lisible par MED-SuMo (format propriétaire).
2. **SuMo.multi** gère le reste de MED-SMA. Cette fonction interprète les résultats obtenus par la fonction de comparaison multiple pour détecter les *singlepatches*, former les *multipatches*, construire la matrice de similarité et enfin classer avec MCL. Son argument est le fichier binaire et son résultat est la classification.

---

## v. Analyse des résultats

Les résultats générés par MED-SMA sont sous la forme d'une liste de *multipatches*, chacun associé à un numéro de groupe. L'identifiant du *multipatch* est composé du code PDB de la structure, de l'indice du site, ainsi qu'un identifiant numérique du *multipatch* lui-même, par exemple 1B38\_1\_1099. Pour visualiser les classifications en deux dimensions (cf. Figure 37), un fichier texte lisible par le logiciel Biolayout [154] est généré.

### Ajout d'informations supplémentaires pour l'analyse des résultats

Pour analyser les résultats, l'utilisation d'informations annexes est indispensable. Les bases de données PROSITE et PFAM permettent de faire facilement le lien entre les fichiers de la PDB et leurs annotations fonctionnelles. Ainsi, le site internet de PROSITE (<http://www.expasy.org/prosite/>) met à disposition un fichier texte contenant les correspondances entre les structures de la PDB et ces différents motifs. Par exemple, le motif PDOC00100 décrit dans la partie I.i est associé à l'annotation "*eukaryotic protein kinase*". Il est donc possible d'obtenir la liste de toutes les structures de la PDB contenant ces signatures. Le 27 janvier 2009, le motif PDOC00100 est associé à 535 structures de la PDB.

Ces banques de données sont donc très utiles pour annoter les structures, il est toutefois important de souligner deux points. Le premier concerne l'élément de la structure qui contient réellement le motif de séquence, en effet il ne concerne généralement qu'une partie de la structure comme une des chaînes de la protéine. Par exemple, pour l'analyse d'une classification de sites, la chaîne liée au motif n'est pas forcément celle qui contient le site de liaison. Le second point concerne la couverture du jeu de données par les annotations. En effet, si seulement 50% des structures possèdent une annotation, 50% des résultats ne peuvent être analysés. Une correspondance entre les annotations Prosite et la PDB extraite en septembre 2008 illustre d'ailleurs cet exemple ; 27843 fichiers PDB contenaient au moins une annotation Prosite à cette date où la PDB contenait plus de 53000 structures protéiques, soit en taux de recouvrement d'environ 52%.

L'étude décrite dans la partie à venir III.ii est une application de MED-SMA sur un jeu de protéines fixant des ligands puriques. 2322 sites de liaisons sont classifiés en 247 clusters. Pour l'analyse des clusters, nous avons initialement décidé d'utiliser les annotations de PROSITE, cependant seulement la moitié des clusters sont composés de structures possédant une annotation PROSITE. Nous avons néanmoins inclus l'analyse de leur distribution aux résultats, mais d'autres annotations ont aussi été utilisées. Même si de plus en plus de structures sans fonction sont résolues, la majeure partie des nouvelles structures publiées

concerne des protéines de fonction connue au moment de son dépôt dans la PDB. De ce fait, les champs *MOLECULE*, *HEADER* ou *TITLE* des fichiers PDB sont souvent renseignés et contiennent des informations utiles. Par exemple, ces champs pour la protéine cycline dépendante kinase 2 (CDK2) 1B38, sont respectivement « CELL DIVISION PROTEIN KINASE 2 », « TRANSFERASE » et « HUMAN CYCLIN-DEPENDENT KINASE 2 ». Alors que le champ *HEADER* n'est pas assez précis, et que le champ *TITLE* peut l'être trop (il peut contenir des informations sur la résolution de la structure), le champ *MOLECULE* contient une annotation fonctionnelle précise et facilement évaluable. Sur les 2322 sites de notre jeu, seulement six protéines n'ont pas de champ *MOLECULE* exploitable. Après vérification systématique (et manuelle) de toutes les annotations, ces six structures ont été annotées comme protéines hypothétiques.

La qualité d'un algorithme de classification de site actif est définie par sa capacité à rassembler des éléments similaires pour former des clusters à partir d'une métrique capable de réellement détecter les similarités d'interaction. MED-SMA est une méthode de classification fonctionnelle des sites de liaison; pour l'évaluer, il faut analyser l'homogénéité des fonctions biochimiques des sites présents dans chacun des clusters formés. Pour ce faire, différentes solutions existent, nous avons choisi l'utilisation des deux solutions suivantes.

#### Nombres d'états équivalent : le $N_{eq}$

Une mesure dérivée de l'entropie de Shannon [155] a été définie par le Pr Serge Hazout pour évaluer l'homogénéité d'un cluster: « le nombre d'états équivalents ou  $N_{eq}$  » [156]. Il représente la répartition du nombre d'états présents par rapport au nombre maximal d'états observés. Ici, il est équivalent au nombre de fonctions observées par groupe. Son calcul est simple et repose sur l'entropie du groupe  $c$ ,  $H(c)$  (cf. équation 8). Le  $N_{eq}$  est simplement l'exponentielle de l'entropie de Shannon (cf. équation 9).

$$H(c) = -\sum_{i=1}^{i=F} p(i_c) \cdot \ln F_s p(i_c) \quad (8)$$

Où  $p(i_c)$  est la probabilité que la fonction  $i$  dans le groupe  $c$  et  $F$  est le nombre de fonctions observées dans le groupe  $c$ .

$$Neq(c) = \exp[H(c)] \quad (9)$$

Le  $N_{eq}$  varie donc entre 1 (valeur du  $N_{eq}$  lorsqu'il n'y a qu'une unique fonction dans le groupe) et  $F$  (valeur si chaque élément possède une fonction différente). Dans l'application III.ii, le  $N_{eq}$  est calculé sur les annotations extraites du champ MOLECULE de tous les fichiers PDB. Ces annotations ont toutes été vérifiées manuellement et des corrections ont été effectuées.

### Spécificité et sensibilité

La spécificité et la sensibilité des annotations sont des mesures statistiques qui permettent d'évaluer la pureté des groupes formés par une méthode de classification. Contrairement aux mesures statistiques définies habituellement [157], ces valeurs ne fournissent pas les taux de faux-positifs ou -négatifs des groupes formés. En effet, la méthode de classification que nous présentons est difficilement comparable à des résultats issus d'autres méthodes. Sans classification référence, il n'est pas possible de calculer ces taux. Ici, la spécificité et la sensibilité permettent d'évaluer les groupes formés en fonction d'un type d'annotation de référence:

La sensibilité mesure la diversité des annotations présentes dans un groupe, elle permet de spécifier la pureté d'un groupe. Elle se calcule par groupe, de la manière suivante:

$Nbre\_sites_{X \rightarrow Y}$  = Nombre de sites dans le cluster X liés à l'annotation Y

$Nbre\_sites\_X$  = Nombre de sites dans le cluster X

$$Sensibilité = \frac{Nbre\_sites_{X \rightarrow Y}}{Nbre\_sites\_X} \quad (10)$$

Une sensibilité de 1 signifie que tous les sites de liaison du groupe sont annotés identiquement.

La spécificité d'une annotation fournit sa distribution dans les groupes par rapport à sa distribution dans tout le jeu de données. Elle se calcule, aussi pour chaque groupe, de la manière suivante :

$Nbre\_sites_{X \rightarrow Y}$  = Nombre de sites dans le cluster X liés à l'annotation Y

$Nbre\_sites \rightarrow Y$  = Nombre de sites dans tout le jeu de données liés à l'annotation Y

---

$$\text{Spécificité} = \frac{\text{Nbre\_sitesX} \rightarrow Y}{\text{Nbre\_sites} \rightarrow Y} \quad (11)$$

Une spécificité de 1 signifie que toutes les annotations Y sont regroupées ensemble.

Les mesures  $N_{eq}$ , spécificité et sensibilité peuvent être utilisées pour analyser les groupes créés par MED-SMA. Ils fournissent une manière rapide d'évaluer la pureté des groupes. Seul le  $N_{eq}$  a été utilisé pour une des applications de MED-SMA décrites dans III.ii. Il est aussi important de préciser que l'objectif n'est pas «la création de clusters exclusivement homogènes», mais «de mettre en évidence des similitudes entre sites non détectées par des méthodes de comparaison structurales traditionnelles». En effet, ce système de classification permet de rassembler des sites de liaison structuralement et fonctionnellement semblables. Cependant, si les annotations sont différentes et si les sites sont associés au même groupe, leurs mécanismes ou leurs modes de liaison aux ligands devraient être proches. Deux applications de MED-SMA sont présentées dans la partie suivante: l'une sur les protéines du repliement GHKL et l'autre sur les protéines fixant les ligands puriques. Ces deux applications ont fait l'objet de deux articles scientifiques dont le premier est accepté alors que le second est soumis (article 2, article 5).

### 3. Deux applications de MED-SMA

#### i. *GHKL fold*

Les HSP90s (*Heat Shock Protein*) sont abondantes dans le cytoplasme des cellules. Cette protéine existe sous différentes isoformes et exerce principalement des rôles de protéines chaperonnes tels que le contrôle du repliement des protéines, la survie cellulaire [158] (apoptose), ou la répression de certaines tumeurs [159]. Elle interagit avec ATP (cf. Figure 28a). Elle est aussi la cible de molécules actives innovantes telles le geldanamycin, dont l'action permet une réduction de 50% de la croissance de certaines tumeurs [160], ou le celasterol qui perturbe les interactions entre les HSP90 et la protéine Cdc37 dans les cellules cancéreuses du pancréas [161]. Actuellement, le radicicol (RDC) est une molécule très étudiée (cf. Figure 27). Son affinité est très grande pour le site de liaison des HSP90 (de l'ordre de 20nM) [162] (cf. Figure 28b).

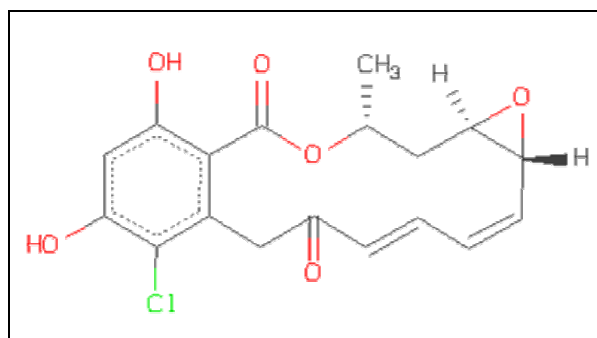


Figure 27: Vue 2D de la molécule radical RDC.

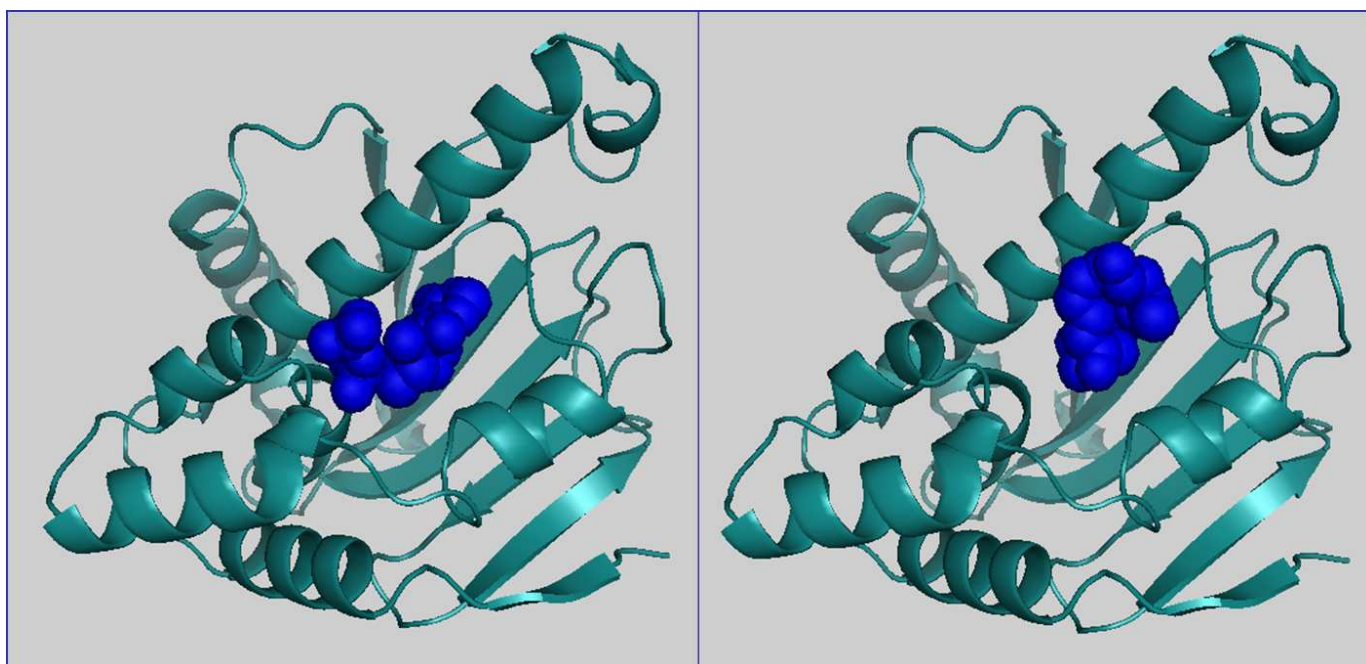
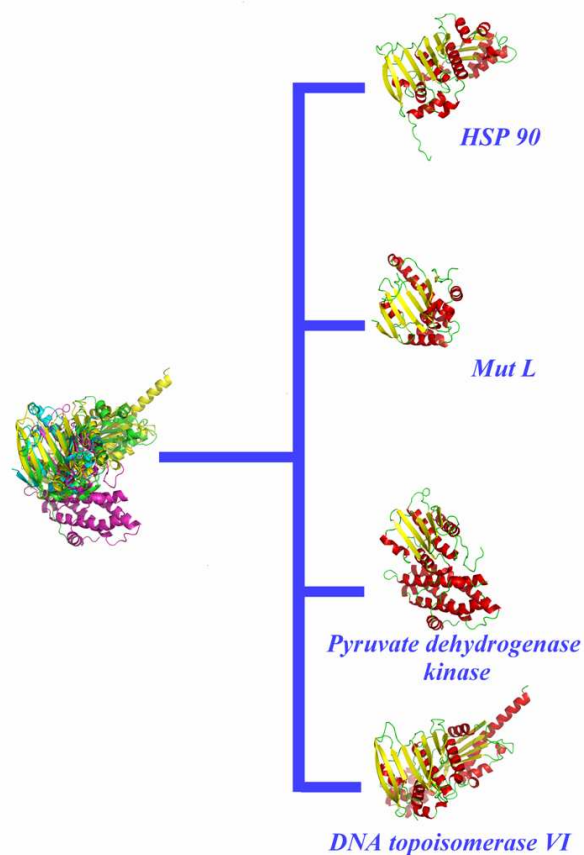


Figure 28 : Représentation de deux HSP90.

La structure de gauche 1AMW [163] est co-cristallisée avec le ligand ADP alors que la structure de droite 1BGQ [162] est co-cristallisée avec le radical (RCD).

Toutefois, le radical n'est pas spécifique des HSP90. Il a été démontré expérimentalement qu'il se fixe aussi aux Sensor Kinases PhoQ [164] et aux topoisomerase VI [165]. De manière intéressante, les HSP90, ADN topoisomerase / MutL et histidines kinases sont rassemblées dans une même superfamille SCOP (cf. Figure 29) (<http://scop.berkeley.edu/data/scop.b.e.ccg.A.html>) car elles partagent un repliement commun : le *Bergerat fold* [78]. Par ailleurs, leur site d'interaction avec l'ATP semble avoir de fortes similitudes structurales.



**Figure 29 : Protéines membres de la superfamille SCOP GHKL.**

Les HSP90, les protéines MutL, les pyruvate déhydrogénase kinases et les DNA topoisomérases VI ont ce *fold* en commun.

Les 116 structures composant cette superfamille (ID SCOP: 55874) ont été sélectionnées afin d'analyser et de classer, de manière précise, leurs sites de liaisons par l'approche MED-SMA.

### Construction de la base

Seuls les sites de liaisons impliquant des ligands de plus de 10 atomes ont été inclus dans la base MED-SuMo, soit 101 structures PDB sur les 116 initiales. Au total 146 sites de liaisons composent la base de sites. Différents types de ligands sont présents : des purines telles l'ATP ou le N-Ethyl-5'-Carboxamido Adénosine mais aussi des molécules actives comme le radicicol ou la novobiocine. Les ligands de cristallisation tels le tétraéthylène glycol (PG4) et le pentaéthylène glycol (1PE) ont été manuellement retirés. Sur les 146 sites de liaison, 78 sont issus d'HSP90, 38 de Topoisomérases/mutL, 26 d'histidine kinases et 4 d' $\alpha$ -ketoacid déhydrogénase kinases C (BCK). Les paramètres de la base MED-SuMo qui ont été utilisés dans cette étude sont [*ligand\_radius* = 6,0 Å, *edge\_max* = 13 Å, *max\_edge\_sum* = 39

Å]. Les paramètres de classification sont [*covering\_factor* = 0,6, *score\_min* = 4,0] La classification dure 2 minutes sur un machine de type *2Xeon Quad Core 5335, 16 GB RAM*.

### Analyse des groupes

MED-SMA forme cinq groupes (ou MED-clusters). La distribution des sites dans ces groupes est montrée par le Tableau 3, la composition de chaque groupe est présentée dans l'Annexe 4. Deux types de groupes sont à distinguer. D'une part les trois groupes homogènes, car ils ne contiennent que des protéines issues d'une seule famille SCOP (MED-clusters 1, 3 et 5) ; les MED-clusters 1 et 3 sont spécifiques à la famille Topoisomérase/MutL alors que le cluster 5 ne contient que des histidines kinases. D'autres part, les deux autres groupes hétérogènes: le MED-cluster 2 contient des sites de liaisons de deux familles: BCK et les histidine kinases et le MED-cluster 4 contient des sites de liaison issus de trois familles: HSP90, Topoisomerase/MutL et BCK.

SCOP fam. MED-Clusters	HSP90	DNA gyrase MutL	Histidine Kinase	$\alpha$ -ketoacid dehydrogénase kinase C
1	0	22	0	0
2	0	0	15	3
3	0	6	0	0
4	78	10	0	1
5	0	0	11	0

**Tableau 3 : Description des MED-clusters obtenus à l'issue de la classification par MED-SMA par rapport aux familles SCOP.** Les MED-clusters sont représentés verticalement alors que les familles SCOP sont disposées horizontalement. Ce tableau met en évidence l'homogénéité des MED-clusters 1, 3 et 5 et l'hétérogénéité des MED-clusters 2 et 4.

### MED-cluster 1 et 3

Les MED-clusters 1 et 3 contiennent respectivement 22 et 6 sites de liaison des 38 protéines de la famille topoisomerase / MutL / DNA gyrase. Les deux formes de topoisomerase VI d'*Escherichia coli* (code PDB 1S14 et 1S16) ont un taux d'identité de séquence de 99.5%. Leur seule différence provient d'une insertion de 23 résidus dans la séquence de la protéine de 1S16. Elles sont cependant séparées par MED-SMA. En observant précisément leur site de liaison à l'ATP, les deux sites ont en effet des ressemblances, mais possèdent aussi de fortes différences. La Figure 30 représente une superposition en 3D de ces deux protéines. Les régions notées (1) montrent une très grande ressemblance entre plusieurs feuillets  $\beta$  et deux hélices  $\alpha$  ainsi qu'au niveau de la partie gauche du site de liaison: partie



---

mise en évidence par 5 SCFs et notée (2) sur la Figure 30. Inversement, l'autre partie du site de liaison notée (3) sur la Figure 30 est très différente.

Les ligands de ces deux topoisomérases sont la novobiocine pour 1S14 et le Phosphoaminophosphonic Acid-Adenylate Ester (ANP) pour 1S16. Leurs structures sont mal superposées dans les sites de liaison. Leur chevauchement est assez faible, à peine 10 atomes, alors que ces molécules contiennent 44 atomes pour la novobiocine et 31 atomes pour l'ANP. De plus, la novobiocine ne peut pas se positionner dans le site de liaison de 1S16 car un clash stérique apparaît entre le ligand et une des hélices  $\alpha$  de 1S16, partie notée (4) sur la Figure 30. Il est donc évident que les sites de liaison des groupes 1 et 3 sont trop différents pour fixer le même type de molécule avec le même mode de liaison ou pour être rassemblés dans le même MED-cluster. Cette conclusion est surprenante au premier abord, les deux formes de topoisomérases étant très proches, mais l'insertion des 23 résidus provoque des différences majeures quant à leurs affinités envers les molécules liantes. Notre étude renforce celle de Bellon et ses collaborateurs [166]. Elle met de plus en évidence « avec élégance » le fait que deux conformations locales distinctes peuvent être détectées même dans des protéines particulièrement proches.

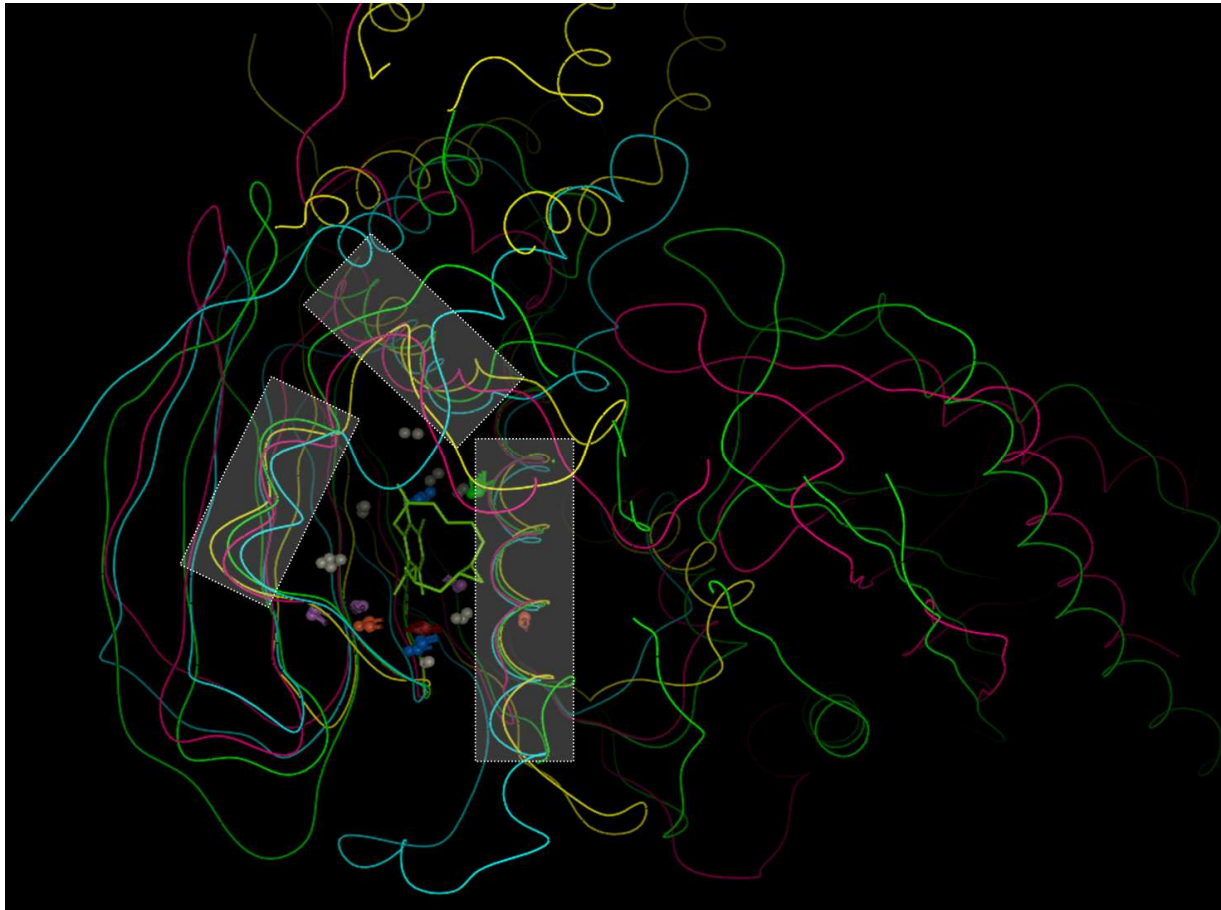


**Figure 30 : Superposition de deux topoisomérases VI séparées par MED-SMA.**

Les structures des topoisomérases PDB 1S16 (rouge) et 1S14 (vert) sont superposées mais leurs sites de liaisons ne se sont pas assez semblables pour être rassemblés dans le même cluster. Cette figure est divisée en différentes régions numérotées : (1) Les parties similaires des deux structures: deux hélices  $\alpha$  et plusieurs feuillets  $\beta$  sont communs aux deux structures. (2) La faible ressemblance au niveau de leur sites de liaisons est démontrée par la présence de seulement 5 SCFs. (3) Les parties différentes des deux structures. (4) clash stérique potentiel entre les ligands des deux sites de liaison.

#### MED cluster 4

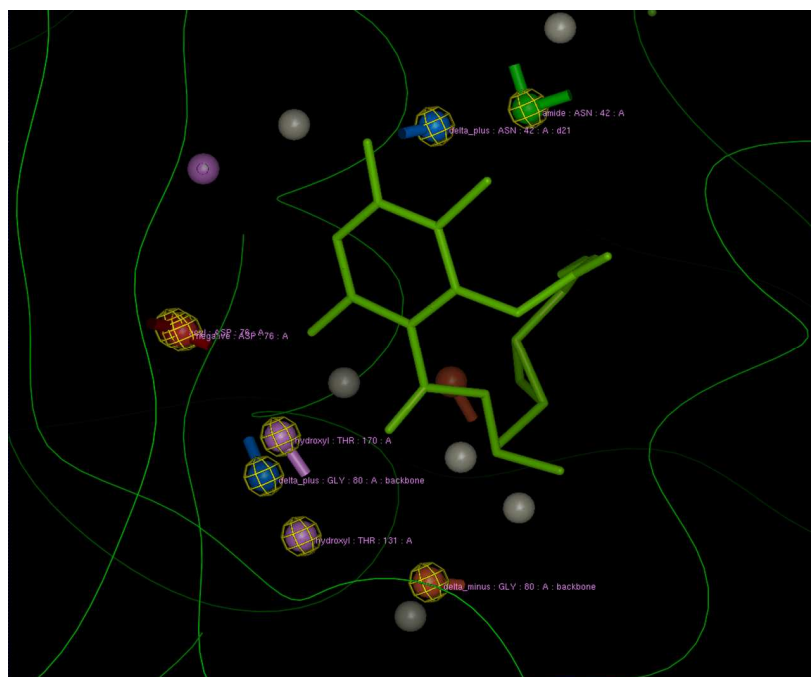
Le groupe 4 contient des protéines de trois différentes familles SCOP. Il s'agit aussi du groupe le plus grand; il contient 89 sites de liaison. Toutes les HSP90 du jeu de données sont présentes (78 sites), 10 de la famille mutL/DNA topoisomerase (1 topoisomerase, 5 mutL et 4 PMS2) et 1 de la famille BCK. Les ligands sont aussi très diversifiés, 48 types de ligands différents sont présents.



**Figure 31 : Superposition de quatre protéines issues de trois familles SCOP rassemblés par MED-SMA.** (Codes PDB 2HKJ (vert), 2CCT (cyan), 1B63 (rose), 1JM6 (jaune)). Les rectangles blancs montrent les similitudes autour des ligands ainsi que les hélices du repliement de type Bergerat. Le reste des structures est assez différent.

Les sites de liaison de ce groupe ont un ensemble de SCFs en commun. La Figure 31 représente une superposition globale de quatre structures issues des familles. Les rectangles blancs soulignent les régions similaires détectées par MED-SMA alors que le reste est très différent. La Figure 32 est une vue rapprochée autour du ligand radicicol. Les huit SCFs portant un label descriptif sont présents dans toutes les signatures SCF des structures superposées dans la Figure 31. Ils entourent complètement le ligand, ce qui signifie que les similitudes concernent tout le site de liaison.

Le fait que MED-SMA regroupe les sites de liaisons de trois familles SCOP suggère que leurs modes de liaison sont excessivement proches. Comme le radicicol est une molécule non spécifique qui se fixe à la fois aux HSP90 et aux topoisomérases VI, elle pourrait aussi se lier aux autres protéines incluses dans le MED-cluster 4, la famille PMS2.



**Figure 32 : Vue rapprochée autour du ligand radical.**

Les huit SCFs entourés de jaune sont communs à toutes les structures superposées dans la Figure 31. Ils entourent complètement le ligand ce qui suggère que les similitudes concernent tout le site de liaison.

### MED-clusters 2 et 5

Les MED-clusters 2 et 5 sont principalement composés d'histidine kinases. Le MED-cluster 2 est hétérogène alors que le 5 est homogène. La taille des sites de liaisons composant le groupe 5 est aussi homogène car tous les ligands sont puriques (et de taille comparable). Les sites de liaisons rassemblés par MED-SMA ont des modes de liaison aux ligands similaires, alors que les protéines sont distinctes. Ce groupe pourrait être utilisé pour la recherche de molécules spécifiques aux histidine kinases CheA afin d'inhiber leur action.

De manière surprenante, le groupe 2 contient aussi deux histidine kinases CheA (Codes PDB 2CH4 et 1I5D). L'association de ces protéines à ce groupe et non au précédent est logiquement due à des différences au niveau de leur site de liaison. Il faut noter que le score MED-SuMo entre les sites est inférieur au seuil fixé lors de l'étape de comparaison multiple. Ainsi, une molécule conçue pour inhiber les sites de liaison du groupe 5 ne devrait se fixer qu'avec une affinité trop faible aux histidine kinases du groupe 2, pour inhiber leur action.

Pour finir, le radical inhibe les HSP90s mais aussi l'action des protéines des familles BCK et anti-sigma factor spoIIab. Si MED-SuMo détecte les modes de liaison similaires, MED-SMA en regroupant les sites d'interaction proches sépare les HSP90s des deux autres familles. Cette classification met en évidence le fait que les modes de liaison au radical des

---

sites du groupe 2 et des sites du groupe 4 n'impliquent pas les mêmes types d'interaction, ils sont différents.

### **Conclusion/Discussion**

Pour cette étude, l'approche MED-SMA est utilisée pour analyser les protéines de la superfamille SCOP « ATPase domain of HSP90 chaperone / DNA topoisomerase II / histidine kinase ». Les protéines de cette superfamille sont globalement différentes mais leurs surfaces d'interaction à l'ATP présentent certaines ressemblances. MED-SMA forme cinq clusters parmi lesquels trois sont homogènes. Ces trois MED-clusters reflètent la spécificité des sites de liaison. On peut ainsi établir l'hypothèse qu'aucune molécule se liant aux sites du groupe 1, ne se fixerait aussi aux protéines du groupe 2 en impliquant les mêmes interactions. Le fait que les ligands soient des molécules puriques dans les groupes 1 et 2 amplifie l'observation suivante: les ligands sont identiques mais leurs modes de liaison aux surfaces des protéines sont différents. En revanche, le MED-cluster 4 rassemble des sites issus de trois familles différentes. La superposition en 3D effectuée avec MED-SuMo GUI, illustre des différences entre les structures globales, alors que le repliement Bergerat est visible et conservé (rectangle blanc sur la Figure 31). Des SCFs communs entourent le ligand radicicol, indiquant que les similitudes détectées dans les sites de liaisons sont présentes dans les trois familles. De plus, ces résultats sont en total accord avec des données expérimentales qui montrent que les protéines des trois familles SCOP se lient au radicicol [166].

Cet exemple illustre l'efficacité de la méthode de classification à détecter les similitudes et les différences entre les surfaces d'interaction des protéines. Certaines protéines proches ont des différences locales au niveau des interactions dans lesquelles elles sont impliquées alors que des protéines éloignées peuvent partager des modes de liaison similaires. Par exemple, les protéines connues pour se lier au radicicol sont regroupées dans le même MED-cluster alors que leurs structures globales sont différentes.

#### *ii. Sites de liaison aux purines*

La deuxième application concerne un jeu de protéines largement plus grand rassemblant 2229 structures des protéines se fixant aux ligands puriques, soit le purinôme. Ces protéines sont particulièrement étudiées [167]. En effet, la recherche de molécules actives permettant leur activation ou leur inhibition est primordiale pour lutter contre différentes pathologies. Ainsi, la conception de molécules actives sur les protéines HMG CoA réductases et

---

dihydrofolate réductases ont permis la mise au point de traitement pour l'artériosclérose, l'arthrite rhumatoïde et le cancer. Par exemple, les statines inhibent l'action des HMG CoA réductases en mimant l'action du HMG CoA. Ces molécules permettent la réduction du taux de cholestérol sous forme de LDL (*Low density lipoprotein level*) [167].

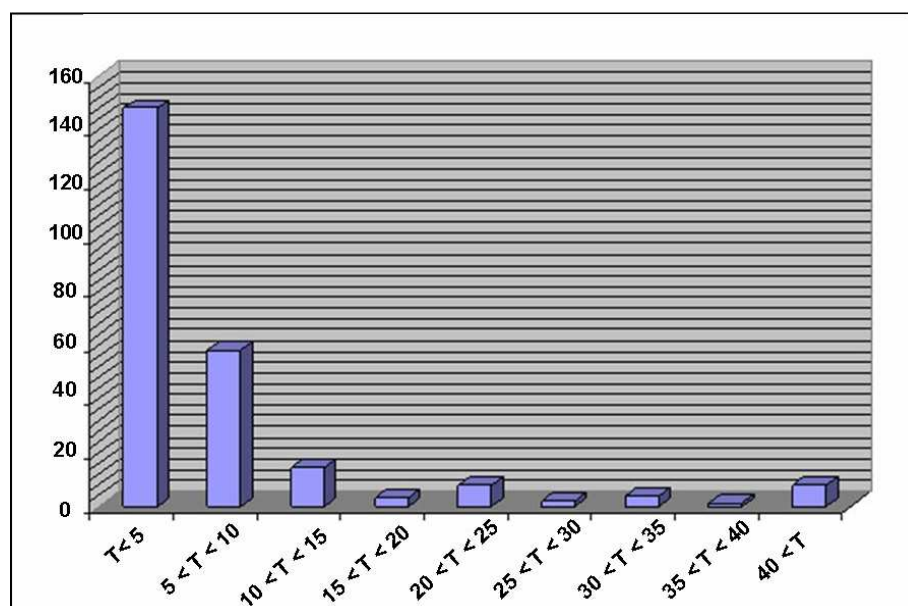
Dans cette étude, au lieu de réunir des sites de liaison fonctionnellement liés, MED-SMA rassemble des sites dont les modes de liaison sont similaires, donc qui peuvent potentiellement être activés ou inhibés par les mêmes molécules. Le protocole est identique à celui précédemment présenté.

### **Construction de la base et classification**

La base MED-SuMo utilisée dans cette étude ne contient que les sites de liaison aux 9 ligands ATP, ADP, AMP, ANP, GTP, GDP, GMP, GNP et NAD. Cette sélection permet de rassembler 2322 sites indépendants. Les paramètres de la base sont [*ligand\_radius* = 6,0 Å, *edge\_max* = 13 Å, *max\_edge\_sum* = 39 Å]. Les paramètres de la classification sont [*covering\_factor* = 0,6, *score\_min* = 5,5]. La valeur du score minimum pour les comparaisons deux à deux a été fixée en analysant les résultats de calculs MED-SuMo obtenu en comparant une sélection de sites puriques au reste des 2322 sites. La valeur seuil 5,5 pour le paramètre *score\_min* a été choisie car il correspond au score limite laissant apparaître les premiers « faux positifs ». Un « faux positif » est un *hit* qui se superpose d'une manière non satisfaisante à la protéine requête, la diversité de la PDB permettant dans certains cas des superpositions avec 3, 4, 5 ou plus de SCFs n'ayant aucun sens biochimique. Ces faux positifs ont en général un RMSD entre les SCFs trop élevé. La classification par MED-SMA prend 4 heures sur une machine à 8 cœurs de type '*bi-Xeon Quad Core 5335, 16 GB RAM*', dont 3h30 pour les comparaisons deux à deux et 30 minutes pour la construction de la matrice de similarité et la classification MCL.

### **Analyse des groupes**

L'analyse des données par MED-SMA montre 247 groupes contenant 2115 sites de liaison. 207 n'ont pu être associés à aucun autre site de la base et sont éliminés au moment de la fusion des *singlepatches* (cf. Figure 24b). Ces sites sont donc classifiés comme « singletons ». La Figure 33 représente la distribution de la taille des groupes: 60 % ont une taille inférieure à 5, 25 % entre 5 et 10 et, 14 ont plus de 30 sites de liaisons. Malgré la prédominance des groupes de petite taille, de nombreux groupes de taille moyenne et quelques gros sont présents.



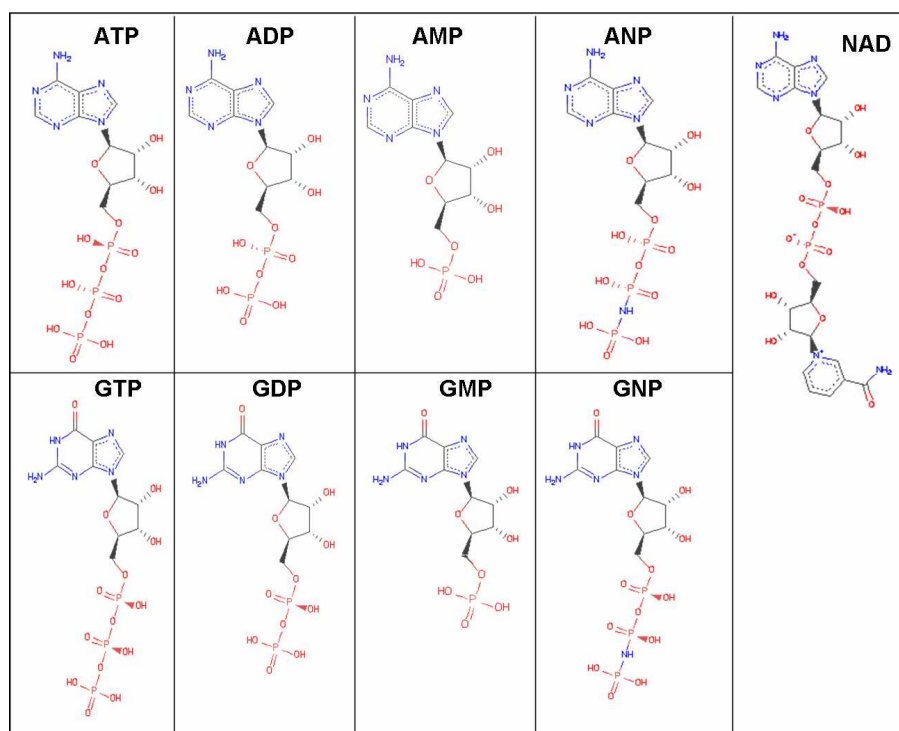
**Figure 33 : Distribution de la taille des groupes.**

163 sur 247 groupes ont une taille inférieure à 5 alors que 14 contiennent plus de 30 sites.  
(La lettre T est mise pour le mot taille.)

Une analyse précise de la classification et des propriétés fonctionnelles des groupes est réalisée dans les prochains paragraphes.

### Distribution des ligands

Bien que MED-SMA classe la surface d'interaction exposée aux ligands, chaque site est associé à un ligand purique co-cristallisé. Nous nommons AXP l'ensemble: ATP, ADP, AMP, et ANP et GXP l'ensemble: GTP, GDP, GMP, et GNP. Ces ligands contiennent soit une adénine soit une guanine qui ne diffère que par deux groupements chimiques (cf. Figure 34). NAD et AXP diffèrent dans la région des groupements phosphoriques.



**Figure 34 : Structures 2D des ligands puriques sélectionnés pour constituer le jeu de données**

Bien que les ligands puriques se ressemblent fortement (cf. Figure 34), la plupart des groupes sont spécifiques d'un certain type de ligand. Le

Tableau 4a répertorie le nombre de fois où un ligand est présent au moins une fois dans un des groupes. Par exemple, 80 groupes ne contiennent que le ligand ATP alors que 42 ont de l'ATP et de l'ADP, 8 ont de l'ATP et du NAD. Ces valeurs sont moins élevées pour les GXPs pour lesquels les groupes sont plus homogènes. Ces ligands apparaissent dans 70 groupes parmi lesquels 54 ne contiennent que des ligands GXP. Il est aussi remarquable que seuls trois groupes contiennent des ligands ATP et GTP alors que leurs structures sont très semblables. D'autres mélanges sont aussi détectables, cependant le nucléotide est souvent le même (Adénine ou Guanine). Le

Tableau 4b est une synthèse du

Tableau 4a. 83% ( $303 \div (303+31+22)$ ) des groupes contenant des AXP ne contiennent que des AXP, 71% ( $76 \div (76+31+0)$ ) pour les GXP et 70% ( $53 \div (53+22+0)$ ) des groupes de NAD. Le mélange du NAD et du GXP n'est jamais observé alors que 30% des groupes contenant du NAD ont aussi des AXP et que 17% des groupes contenant des AXP contiennent aussi des ligands GXP et NAD. MED-SMA est indépendant des types de ligands et permet de différencier les modes de liaisons des ligands des sites classifiés.



a)

	ATP	ADP	AMP	ANP	GTP	GDP	GMP	GNP	NAD
ATP	80	42	21	28	3	3	0	2	8
ADP	42	104	14	47	1	6	1	3	8
AMP	21	14	56	8	1	1	5	0	6
ANP	28	47	8	61	0	2	0	3	0
GTP	3	1	1	0	14	4	0	0	0
GDP	3	6	1	2	4	21	3	1	0
GMP	0	1	5	0	0	3	13	0	0
GNP	2	3	0	3	0	1	0	6	0
NAD	8	8	6	0	0	0	0	0	53

b)

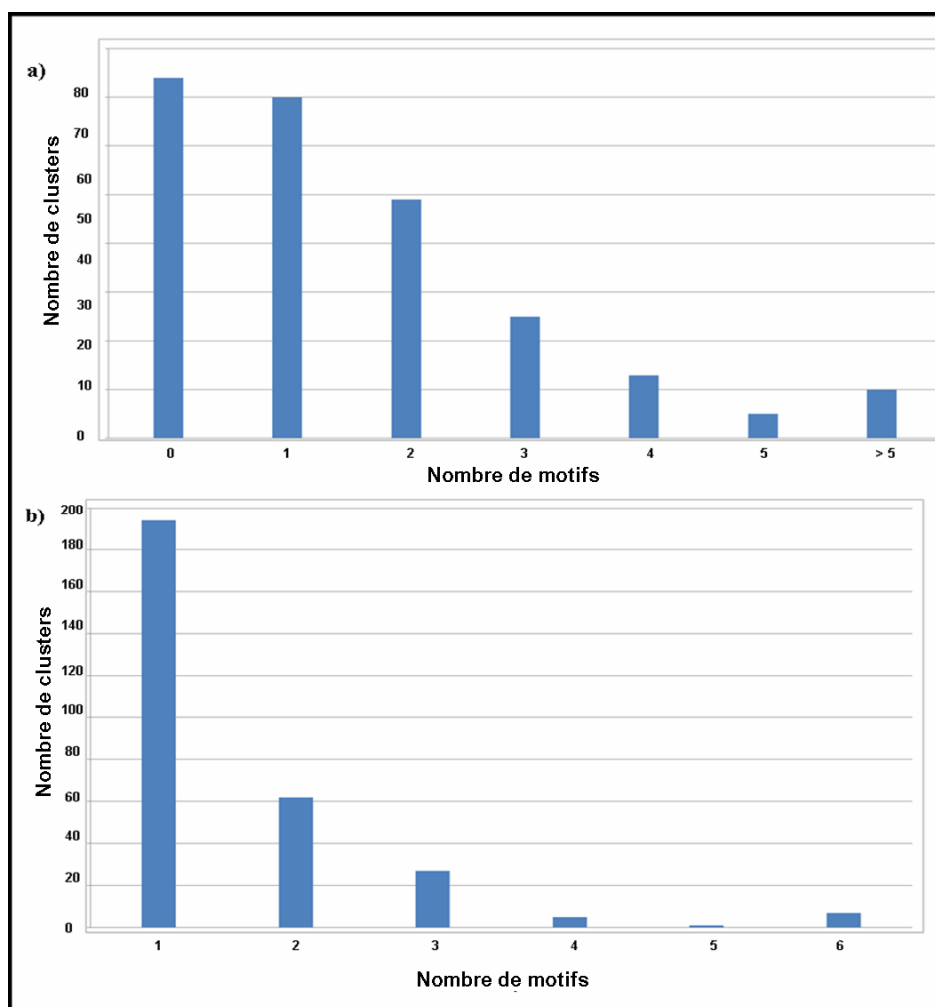
	AXP	GXP	NAD
AXP	303	31	22
GXP	31	76	0
NAD	22	0	53

**Tableau 4 : Matrice distribution des ligands dans les MED-clusters.**

Le tableau a représenté la matrice de confusion pour les 9 ligands. Le tableau est une synthèse représentant la matrice de confusion entre AXP, GXP et NAD. Ces tables montrent que les ligands AXP et GXP se mélangent comme les ligands AXP et NAD. En revanche, les ligands GXP et NAD ne sont jamais présents dans le même groupe.

#### Distribution des motifs / profiles PROSITE

Les protéines des groupes créés par MED-SMA sont associées à 296 motifs PROSITE. L'association entre une structure et un motif se fait par son identifiant PDB. Il est important de noter que d'une part une structure protéique contient souvent plusieurs chaînes et qu'un motif PROSITE n'est souvent associé qu'à une d'elle. L'association effectuée ici présente un biais car le motif n'est pas toujours associé à la chaîne contenant le site d'interaction classifié. L'analyse de la distribution des motifs PROSITE est cependant intéressante. Elle est exposée dans la cette partie.



**Figure 35 : Distribution des motifs PROSITE au sein des MED-clusters.**

- a) Distribution du nombre de motifs dans les groupes, par exemple, 84 groupes ne sont associés à aucun pattern.  
 b) Distribution du nombre de groupes contenant 1, 2...6 motifs PROSITE, par exemple, plus de cinq groupes sont associés à plus de 5 motifs.

La Figure 35 montre deux distributions des motifs PROSITE dans les 247 MED-clusters. Le graphe du haut (cf. Figure 35a) représente la distribution du nombre de motifs dans les groupes: 30% (84/247) ne sont associés à aucun motif, alors que 28% (80/247) ne sont associés qu'à un seul motif. 3% (10/247) sont associés à plus de cinq motifs. Cette catégorie correspond à deux types de groupes. Les premiers sont de grandes tailles comme les groupes 40 et 157. Ils contiennent respectivement 402 et 60 sites de liaisons et sont fonctionnellement très hétérogènes. Les seconds contiennent des structures associées à plusieurs motifs PROSITE tel que les groupes 105 qui rassemblent 70 sites de liaison de structures d'actine. Ce groupe est fonctionnellement très homogène mais 93% des structures sont associées à trois signatures PROSITE (PS00406, PS00432 et PS01132) regroupées dans le motif PDOC00340, alors que quelques autres sont associés à d'autres motifs. Par exemple, le complexe Actin-DNase (Code PDB 1ATN [168]) est aussi lié aux deux autres signatures du

---

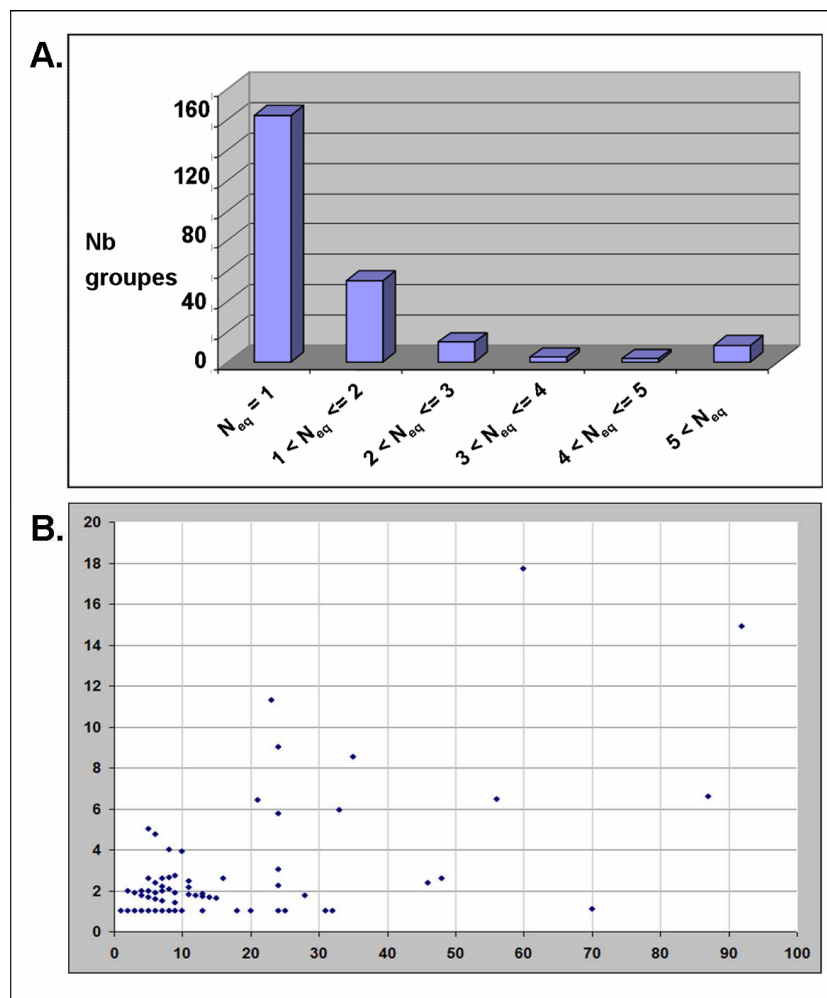
motif PDOC00711 (PS00919 et PS00918). Un autre exemple est relevé dans le groupe 104 où une structure est associée à sept motifs correspondant aux sept domaines des neuf chaînes des trois complexes DNA-directed RNA polymerase II (Code PDB 1TWA [169]) Pour la plupart des groupes associés à des motifs PROSITE, un seul y est prépondérant alors que la présence des autres est souvent dû au fait que d'autres chaînes sont aussi cristallisées dans les structures des protéines qui présentent alors plusieurs sites.

La Figure 35b représente la distribution de chaque motif PROSITE dans les MED-clusters. Plus de 190 motifs sont spécifiques d'un seul groupe, 61 sont présents dans deux groupes et curieusement sept motifs sont présents dans plus de cinq groupes. Les trois premiers sont issus du motif des protéines kinases PDOC00100 : PS00107 « *a protein kinase ATP-binding region signature* », PS00108 « *a Serine/Threonine protein kinase active-site signature* » et PS50011 « *Protein kinase domain profile* ». 61 protéines kinases sont associées à ces trois signatures et elles font parties de cinq groupes différents. Une analyse plus détaillée des groupes contenant des protéines kinase sera réalisée plus loin. Une première observation est que les motifs associés aux protéines kinases sont dans les mêmes groupes ce qui permet de penser que ces protéines peuvent être rassemblées en fonction de leurs modes de liaison.

#### Annotation fonctionnelle

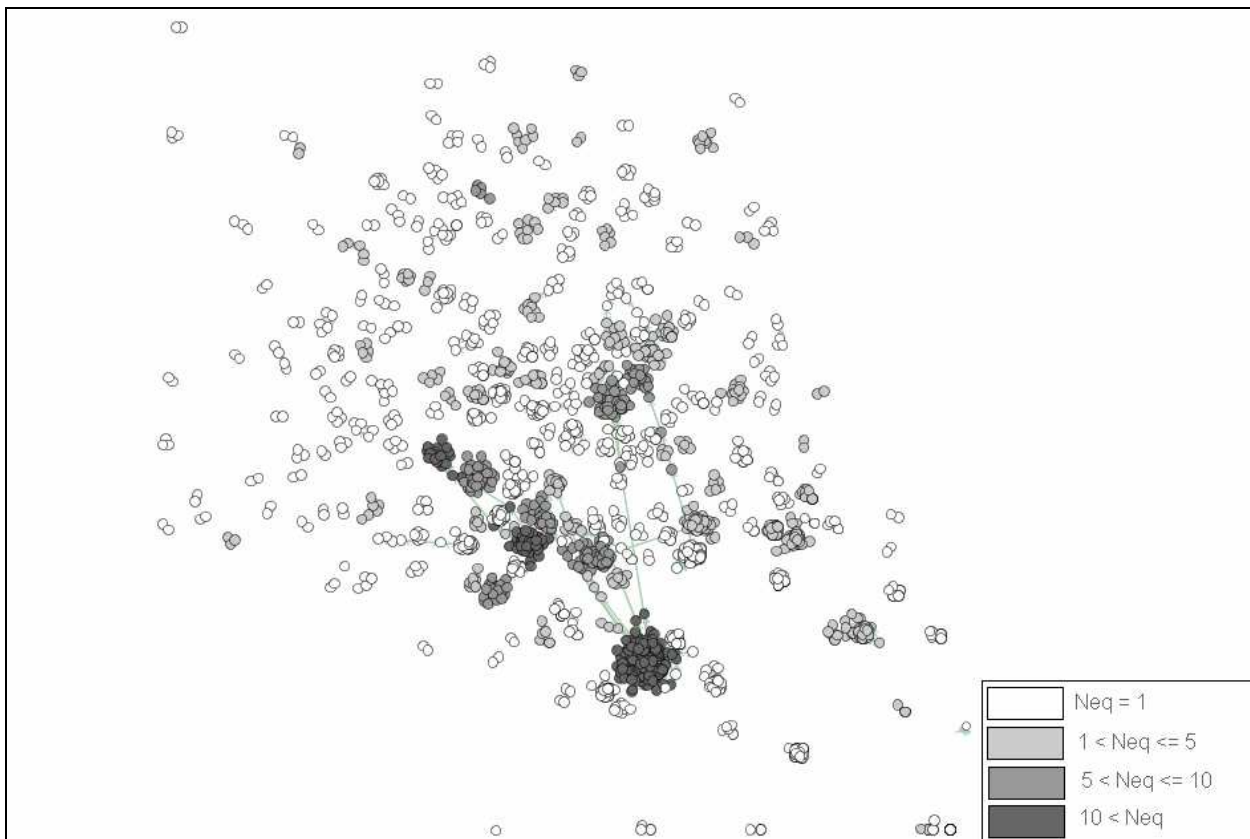
Le champ MOLECULE des fichiers PDB a été utilisé pour extraire des annotations fonctionnelles pour chaque structure du jeu (cf. partie III.B.2.v.). Les fonctions assignées ont toutes été vérifiées manuellement avant l'analyse. 442 fonctions distinctes sont ainsi annotées. Pour analyser l'homogénéité des annotations dans les MED-clusters, le  $N_{eq}$  (nombre équivalent d'états), présenté dans [75] et démontré dans [156], a été calculé pour chaque groupe. Basé sur la probabilité de présence d'une fonction spécifique, le  $N_{eq}$  représente ici le nombre équivalent de fonctions dans un groupe. Il évalue le nombre d'états (ici de fonctions) et leur distribution au sein d'un groupe. Par exemple, si un groupe ne contient qu'une fonction, son  $N_{eq}$  vaut 1, plus le MED-cluster est hétérogène plus sa valeur augmente : si deux fonctions sont chacune présente à 50%, le  $N_{eq}$  vaut 2 alors que si une fonction est présente à 90% et que la seconde à 10%, la valeur du  $N_{eq}$  est de 1.13. La Figure 36 (gauche) représente la distribution des  $N_{eq}$  dans les MED-clusters. 163 groupes ont un  $N_{eq}$  de 1. Un tiers des groupes sont donc au moins associés à deux fonctions. La Figure 36 (droite) représente les valeurs de  $N_{eq}$  calculées en fonction de la taille des groupes. Logiquement, le  $N_{eq}$  augmente en fonction de la taille des groupes bien que quelques groupes de grande taille aient un  $N_{eq}$  bas, et que des petits groupes peuvent avoir un  $N_{eq}$  élevé. La Figure 37 montre une représentation

2D des groupes avec le logiciel Biolayout [154] colorés en fonction de leur  $N_{eq}$ . Huit groupes ont un  $N_{eq}$  supérieur à 5, trois sont analysés en détail dans le paragraphe suivant.



**Figure 36 : Distribution des valeurs de  $N_{eq}$**

A. représente la distribution des valeurs des  $N_{eq}$  dans les différents groupes. Par exemple, 163 groupes ont un  $N_{eq}$  égale à 1. B. représente la distribution du  $N_{eq}$  en fonction de la taille des différents groupes.



**Figure 37 : Distribution des valeurs de  $N_{eq}$  dans une représentation 2D des MED-clusters.**

Les groupes sont colorés en fonction de la valeur de leur  $N_{eq}$ . Les lignes qui apparaissent représentent les liens inter-groupes décrits dans un des paragraphes de l'application.

Le **MED-cluster 4** a un  $N_{eq}$  de 14,92, ce groupe est très hétérogène. Il est composé des sites de liaison de protéines de 27 fonctions différentes. Ces protéines sont principalement des épimérasés, des déhydratases et des déhydrogénésés, par exemple hydroxysteroid dehydrogenases, butanediol dehydrogenases ou la protéine de biosynthèse sulfolipide SQD1. Toutes les protéines de ce groupe fixent le NAD sauf la protéine Arna (code PDB 1Z7E [170]) co-cristallisée avec l'ATP. La Figure 38 montre la superposition de la protéine dTDP-D-glucose 4,6-déhydratase (code PDB : 1KEP [171]) avec trois autres sites de liaison du même groupe. Pour les trois images, la position de 1KEP est fixe, elle sert de référence. L'image de gauche représente une superposition avec le site de liaison de la protéine UDP-Glucose-4-épipérasé (code PDB: 2P5Y). Des SCFs situés tout autour du ligand sont bien superposés, les deux sites de liaison sont très similaires. L'image du milieu représente une superposition avec le site de la protéine GDP-Mannose-4,6-Déhydratase (code PDB 1RPN [172]). De manière surprenante, les ligands sont assez bien superposés, malgré un léger décalage. Toutefois, alors que 14 SCFs en commun sont détectés sur le côté droit sur l'image

de gauche, seulement 2 SCFs sont présents ici. Des différences sont aussi notables du côté de la protéine où certaines hélices sont éloignées. L'image de droite représente une superposition de sites où les ligands sont différents. Les similitudes au niveau des surfaces d'interaction détectées par MED-SuMo permettent de superposer les ligands au niveau de leurs sous-structures communes, l'adénosine. La partie gauche des sites de liaison est semblable pour les quatre protéines (1KEP, 2P5Y, 1RPN et 1Z7E).



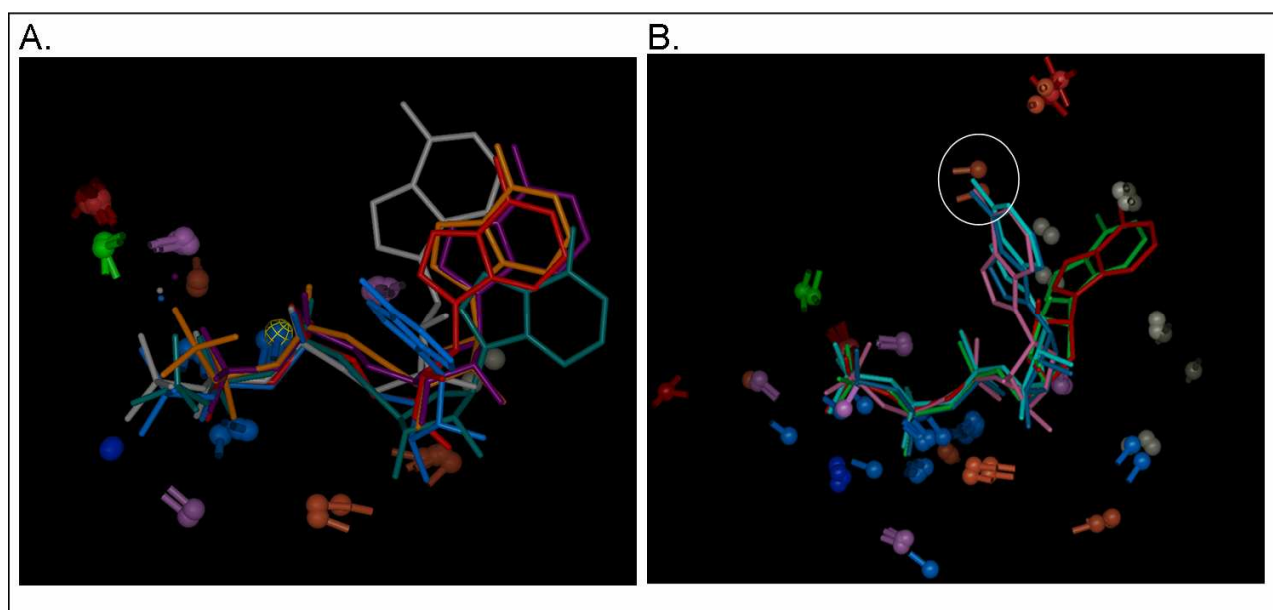
**Figure 38 : Exemples de superposition de sites de liaison en 3D par MED-SuMo.**

Le site de liaison de la protéine dTDP-D-glucose 4,6-déhydratase 1KEP (cyan) est superposé avec trois sites de liaison différents du groupe 4. Chacun possède une fonction différente et implique des parties distinctes du site de liaison. (1) superposition globale des sites de 1KEP et de la protéine UDP-Glucose-4-épimerase 2P5Y (rose). (2) superposition de la partie gauche du site de liaison au NAD de la protéine GDP-Mannose-4,6-Déhydratase 1RPN (bleu). (3) superposition de la partie gauche et haute (celle du nucléotide) du site de liaison de 1KEP avec le site de liaison à l'ATP de la protéine Arna 1Z7E (vert)

MED-SMA détecte les sous-poches communes entre plusieurs sites, il est fort probable que la partie gauche des sites de liaison ait permis le rassemblement de ces sites aux multiples fonctions.

Le **MED-cluster 33** a un  $N_{eq}$  de 8,53. Il est aussi fonctionnellement hétérogène. Toutefois, les fonctions sont toutes liées au transport à travers la membrane, par exemple *cystic fibrosis transmembrane conductance* ou transporteur ABC. Les ligands sont homogènes, seuls des AXP sont présents, et principalement de l'ATP et de l'ADP. Les similitudes détectées sur les sites se situent autour de leur région phosphate. La Figure 39a souligne que les similitudes des sites actifs ne peuvent pas être les parties fixant le nucléotide car leurs positions sont très distinctes et cela implique des différences au niveau de la surface d'interaction. Quelques SCFs sont tout de même présents (9 SCFs), mais la plupart se trouvent du côté des groupements phosphate (25 SCFs). La Figure 39b représente la superposition de six ligands issus de six protéines associées à différentes fonctions: histidine

perméase (code PDB 1B0U [173]), maltose / maltodextrin transport ATP-binding (code PDB 1Q12 [174]), *cystic fibrosis transmembrane transductance* (code PDB 1R0X [175]), *multidrug resistance-associated protein* (code PDB 2CBZ [176]),  $\alpha$ -hemolysin translocation ATP-binding protein HLYB (code PDB 2FF7 [177]) et le peptide transporteur TAP1 (code PDB 1JJ7 [178]). Quatre sont co-cristallisées avec de l'ATP et deux avec de l'ADP. Ce groupe donne un autre exemple des similarités locales détectées par MED-SuMo. Ici encore, des similitudes locales permettent le regroupement de sites de liaison. Ce groupe illustre aussi la flexibilité des ligands puriques qui explique les diversités de modes de liaison existant pour ces types de molécule. Pour finir, il convient de remarquer les SCFs encerclés en blanc qui soulignent aussi la prise en compte, implicite de la flexibilité des protéines par MED-SuMo (cf. Figure 39B). En effet, les règles de superposition sont assez flexibles pour que ces deux SCFs de même type, séparés de 0.5 Å mais orientés dans la même direction, soient considérés comme équivalents.



**Figure 39: Superposition de ligands du MED-cluster 33**

A. Superposition de six ligands du groupe 33. Chaque ligand est extrait de structures de protéine ayant des fonctions différentes: quatre sont ATP (provenant de 2CBZ en gris, de 1B0U en orange, de 1R0X en bleu, 1Q12 en vert), et deux sont ADP (provenant de 2FF7 en rouge et de 1JJ7 en violet). B. Superposition d'autres ligands du groupe 33. La région fixant le groupement phosphate est dense en SCFs, donc hautement semblable dans la plupart des sites de liaison du groupe alors que la région fixant le nucléotide est pauvre en SCFs. Les conformations du ligand sont d'ailleurs assez différentes.

Le fait que plusieurs fonctions soient présentes dans ce groupe souligne que fixer un groupement phosphate, n'est pas spécifique d'une seule fonction. Cette propriété est aussi très commune aux protéines de transport transmembranaire. La fixation du ligand de ces protéines

---

fonctionnellement différentes se fait par le mode de liaison caractérisé dans le MED-cluster 33.

Le **MED-cluster 40** a le  $N_{eq}$  le plus élevé, 53,36. Il est aussi le groupe le plus peuplé et contient 402 sites de liaison issus de 386 protéines. Il regroupe toutes les petites protéines G du jeu de site, 279 (72% du groupe), tous les « facteurs d'élongation 2 » du jeu de sites, 40 soit 10%. La valeur élevée de  $N_{eq}$  provient des 18% restant, ces protéines englobent une quantité importante de fonctions variées.

Il faut aussi souligner quelques groupes de grande taille ayant un  $N_{eq}$  bas et donc bien homogène. En l'occurrence, le groupe 162, dont la valeur de  $N_{eq}$  est 1,74, contient 28 sites de liaison de protéines de choc thermique, les HSP70. D'ailleurs, son  $N_{eq}$  aurait dû être égal 1 car le descriptif utilisé pour désigner HSP70 est différent dans certains cas et rend l'annotation texte non homogène. Le groupe 106 a un  $N_{eq}$  de 1,07, il rassemble les 70 sites de liaisons d'actine du jeu de données. Ces deux groupes confirment le fait que MED-SMA rassemble des protéines de fonction ou de mode d'activation similaires.

MED-SMA forme différents types de groupes, certains sont fonctionnellement très hétérogènes, alors que d'autres sont homogènes. Les sites de liaison de protéines ayant des fonctions différentes mais rassemblées dans le même groupe (par exemple les petites protéines G) peuvent partager leurs mécanismes d'activation ou d'inhibition et donc interagir avec le même type de molécule. De même pour les actines du groupe 106: elles sont séparées de tous les autres sites de liaison du jeu de données, elles doivent donc interagir avec un mode de liaison très spécifique.

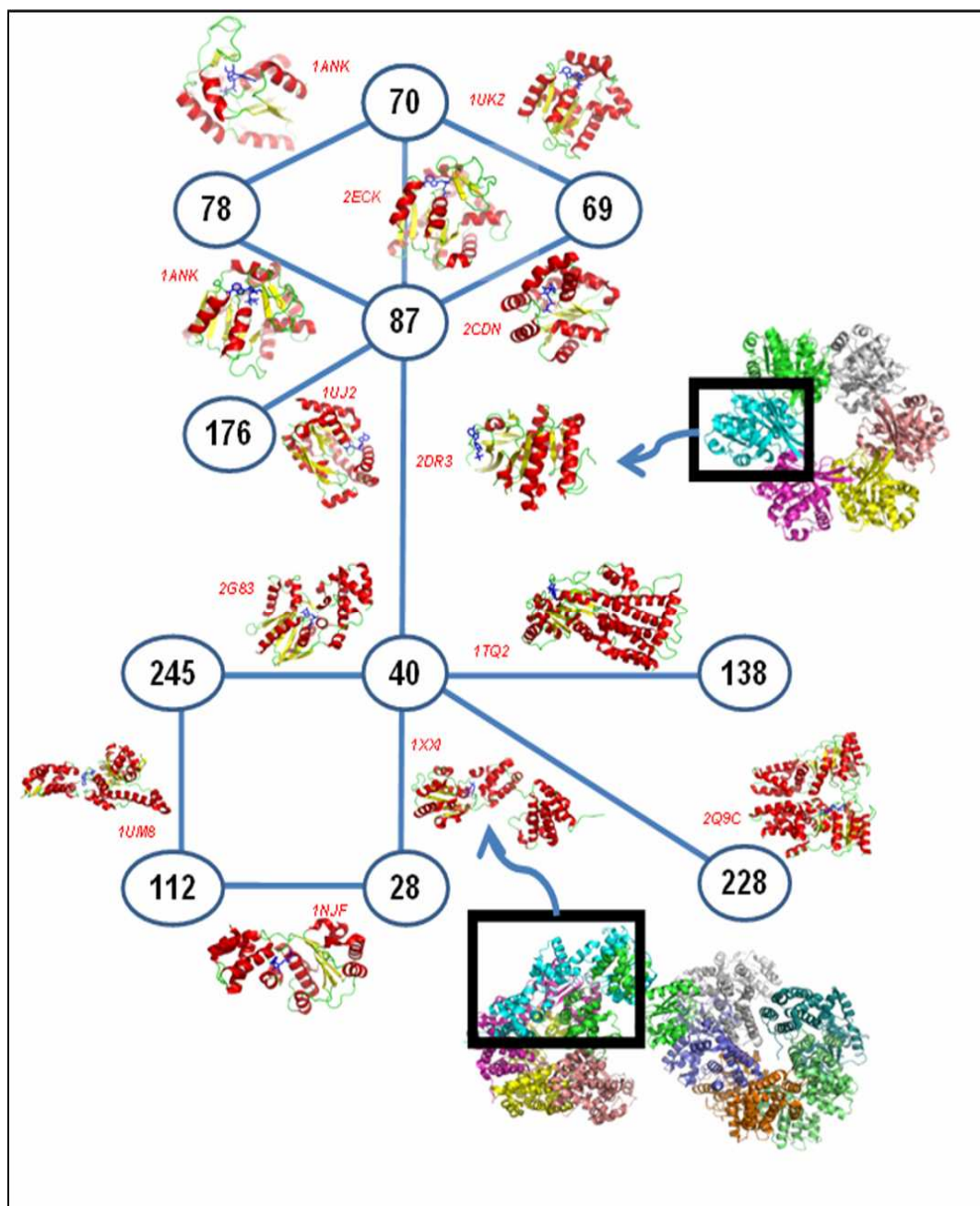
### Liens entre groupes

Dans cette étude, nous considérons que deux groupes sont reliés si chacun contient 1 *multipatch* issu du même site de liaison. Une partie d'un site peut être proche d'une famille de protéines alors qu'une autre partie est proche d'une famille différente. Deux *singlepatches* ne contenant pas assez de SCFs (paramètre *covering\_factor*) pour être fusionnés sont considérés alors comme deux *multipatches* distincts. Cette division va permettre de refléter une certaine flexibilité des sites. La Figure 40 montre un exemple intéressant de lien entre groupes concernant, entre autres le groupe 40 qui est le plus grand des groupes formés. Dans la Figure 37, ce groupe est au centre de la représentation 2D, il est le plus foncé. Le cluster 40 est au centre d'un réseau impliquant 11 groupes qui se divise en deux parties.



---

La partie haute inclut cinq groupes: MED-cluster 70 ( $N_{eq}=1,6$ , taille=5), MED-cluster 69 ( $N_{eq}=1,0$ , taille=2) et MED-cluster 78 ( $N_{eq}=1,0$ , taille=2) sont des groupes d'adénylate kinases. Les adénylate kinases sont des enzymes phosphotransférases qui catalysent la réaction d'interconversion des nucléotides adénine. Ils jouent un rôle important dans l'homéostasie de l'énergie cellulaire [179]. Le MED-cluster 87 ( $N_{eq}=6,5$ , taille=56) contient aussi des adénylate kinases. Son  $N_{eq}$  traduit une hétérogénéité fonctionnelle plus importante, il contient d'autres types de nucléotides kinases comme le thymidylate kinase ou l'uridylate kinase. Toutefois, toutes les protéines du MED-cluster 87 sont des enzymes qui catalysent le transfert d'un phosphate d'un ATP vers l'atome 5' des nucléotides. Le groupe 176 ( $N_{eq}=2,3$ , taille=16) contient d'autres nucléotides kinases comme déoxycytidine kinase (68%). Cependant le déoxycytidine ne fait pas partie des nucléotides naturels. La déoxycytidine kinase humaine est responsable de la phosphorylation d'un certain nombre de nucléosides analogues cliniquement importants [180]. Son mode de liaison doit être différent des autres nucléotides kinases détectées dans les groupes cités antérieurement.



**Figure 40 : Représentation d'un réseau de groupes dans la classification.**

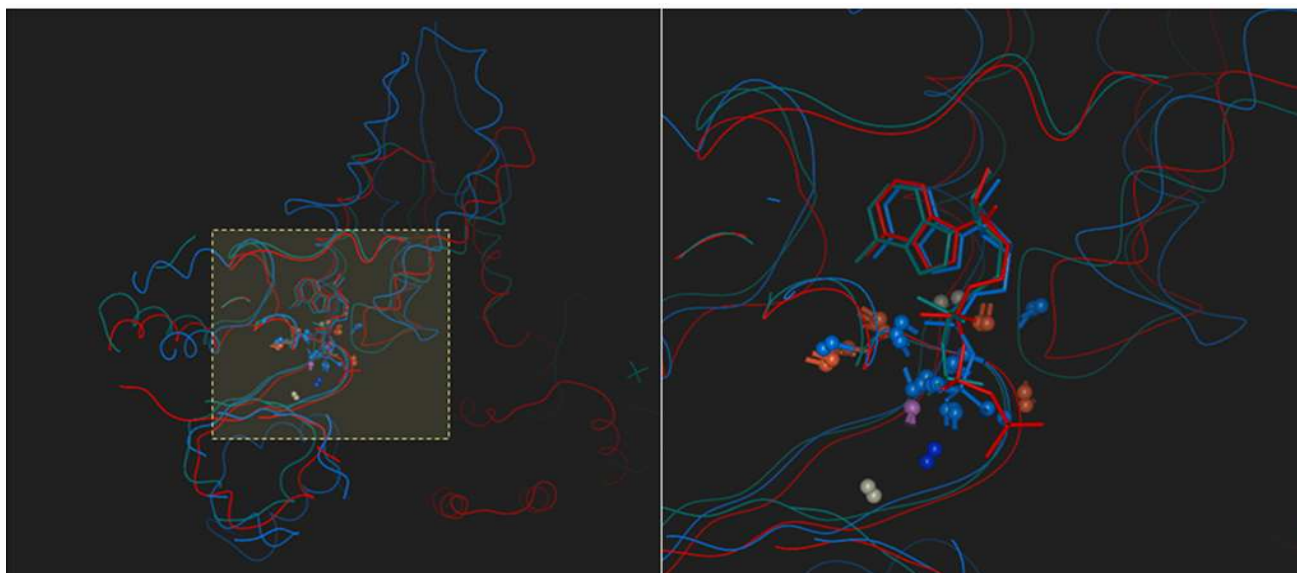
Chaque nombre encerclé est un identifiant. Si deux cercles sont connectés, c'est qu'ils ont chacun un *multipatch* issu d'un même site de liaison. Les structures des protéines affichées sur les lignes sont celles dont est extrait le site de liaison partagé. Par exemple, le site de liaison à l'ADP de la protéine 1XXI est présent dans les groupes 40 et 28. Treize liens impliquant onze groupes sont représentés. Ces liens soulignent la mise en évidence des sous-poches communes par MED-SMA.

La partie basse du réseau inclut six groupes. Le groupe 138 ( $N_{eq} = 1,0$ , taille = 2) contient des GTPases. MED-cluster 228 ( $N_{eq} = 1,0$ , taille = 1) contient un site de liaison conservé dans le domaine GTPase commun à la fois aux protéines SRP (*signal recognition particle*) et aux récepteurs conjugués [181]. Le MED-cluster 28 ( $N_{eq} = 1,9$ , taille = 3) est un petit groupe d'holoenzymes ADN polymérase III. Cet holoenzyme est le premier complexe impliqué dans

---

la réplication de l'ADN chez les procaryotes [182]. MED-Cluster 112 ( $N_{eq} = 1,0$ , taille =15) rassemble des protéines Heat Shock Locus (HSL), une ADN polymérase III et des protéases CLP. HSL et CLP ont des activités de protéines chaperons ; ils aident la formation des complexes protéiques. MED-Cluster 245 ( $N_{eq} = 5,9$ , taille = 33) est un groupe hétérogène qui rassemble principalement des ATPase F1, des protéines RecA ainsi qu'une myosine.

Treize liens sont représentés sur la Figure 40. Le MED-cluster 40 est connecté à cinq groupes, le MED-cluster 70 à trois groupes, cinq autres ont deux liens et les trois restants n'ont qu'un seul lien. Chacune des connexions peut être illustrée par des superpositions 3D. La Figure 41 présente deux liens entre trois groupes : les MED-cluster 245, 40 et 28. Nous avons donc sélectionné trois protéines (codes PDB : 1UM8, 1SXJ et 1XXI) dont les séquences ont un taux d'identité très faible (4,4%, le taux d'identité aléatoire étant de 12%, elles sont vraiment distinctes). Leurs structures ne se ressemblent pas non plus (cf. Figure 41 gauche). Malgré cela, l'observation de la région du site de liaison des ligands co-cristallisés (cf. Figure 41 droite) montre d'importantes similarités locales. Les ligands, un ATP et deux ADP sont bien superposés. La présence de plusieurs SCFs souligne une similarité importante dans la partie basse des trois sites du côté des groupements phosphates des ligands. Les structures des protéines sont d'ailleurs aussi bien superposées de ce côté du site. Toutefois, la séparation de ces sites par MED-SMA s'explique par le fait que le reste des sites est différent. Ces liens soulignent encore une fois des similarités de la sous-poche qui fixe des ligands contenant une partie mono, di ou triphosphate, seule partie de ces protéines qui rassemble leurs sites.



**Figure 41 : Illustration de deux liens inter-groupes**

Sur la gauche est représentée la superposition en 3D de trois structures des groupes 40, 112 et 245 ; code PDB 1UM8 (bleu), 1SXJ (rouge) et 1XXI (vert). Ces protéines ont un taux d'identité de séquence très faible (4,4%) et adoptent des repliements relativement différents. La figure de droite est une vue rapprochée du site actif. Les ligands ATP et deux ADP sont bien superposés et les similitudes locales sont mises en évidence par plusieurs SCFs. Cet exemple est représentatif des détections de sous-poches similaires par MED-SMA.

### Analyse de groupes contenant des protéines kinases

Les protéines kinases jouent un rôle central dans les voies de régulation cellulaire des eucaryotes [183]. Comme elles représentent la seconde plus grande famille de protéines cibles des entreprises pharmaceutiques, le concept d'approche *chemogenomics* systématique a d'ailleurs été fortement exploré sur cette famille dans les projets *kinomics* [184]. Bien qu'elles catalysent essentiellement la même réaction de transfert de groupement phosphate, les voies cellulaires dans lesquelles elles sont impliquées sont particulièrement variées tout comme le nombre de substrats qu'elles catalysent. De nombreuses analyses, ainsi que des classifications basées sur leur séquence ont déjà été réalisées par différentes approches. Certaines études ont pu aussi combiner des informations de séquences et de structures pour classer les protéines kinases. Ayant une forte affinité pour les ligands puriques, elles constituent une partie importante de notre jeu de données (122 sites). L'étude de leur distribution dans notre classification est réalisée dans le paragraphe suivant, elle constitue une partie attrayante de nos résultats.

En 2002, Manning et ses collaborateurs ont établi une classification des protéines kinases humaines, le Kinome [185]. Cette classification basée sur la séquence a permis l'identification de sept familles principales. TK: *tyrosine kinase*, TKL: *tyrosine kinase-like*, STE: *Homologs of yeast Sterile*, CK1; *Caseine Kinase 1*, AGC; *Protein kinase A, C, G*,

---

CAMK: *Calcium/calmodulin-dependent protein kinase*, CMGC: contenant les familles CDK, MAPK, GSK3 $\beta$ , CLK. Une famille de protéines atypiques a aussi été ajoutée, elle contient toutes les protéines kinases non encore caractérisées. Sept de nos MED-clusters incluent des protéines kinases : MED-Clusters 46 ( $N_{eq} = 2,14$ , taille = 11), 100 ( $N_{eq} = 2$ , taille = 2), 121 ( $N_{eq} = 2$ , taille = 2), 155 ( $N_{eq}=1,96$ , taille = 5), 167 ( $N_{eq}=17,71$ , taille = 60), 183 ( $N_{eq}=6,41$ , taille = 21), 211 ( $N_{eq}=11,29$ , taille = 23). Les groupes de taille supérieure à cinq ont été analysés.

Le MED-cluster 46 est le plus homogène, neuf protéines sur dix font partie de la famille AGC et le dixième est de la famille PTK. Le MED-cluster 211 est aussi assez homogène malgré son  $N_{eq}$  élevé; les sites de liaison proviennent principalement de deux branches de l'arbre du kinome, PTK et CMGC, ainsi qu'une protéine d'une branche intermédiaire de ces deux familles (code PDB 2A19). Ce site fait d'ailleurs un lien avec le MED-Cluster 157 dont le  $N_{eq}$  est élevé, mais 59 sur 60 des sites de liaison sont de protéines kinases qui proviennent des trois familles CMGC, PTK et CAMK et de la famille de protéines kinases atypiques. Ces différences entre la classification de Manning et la notre sont intéressantes. En effet, nous classifions les sites de liaison des kinases et non leurs séquences. Afin d'expliquer la disparité des familles de protéines kinases, nous avons lancé un calcul avec MED-SuMo en utilisant le site de liaison à l'ATP de la CDK2 (*cell division kinase 2*) (code PDB 1B38) comme requête que nous avons comparé au reste des sites du groupe 157 avec MED-SuMo. La Figure 42 représente les 31 premiers résultats détectés par MED-SuMo. Ils sont ordonnés par score décroissant. L'analyse fait apparaître en premier lieu la présence de toutes les CDK2 du jeu de données dans les premières lignes du tableau, elles sont surlignées en bleu foncé. La seconde observation concerne la présence de protéines de la même famille (CMGC), surlignées en bleu clair. Avec un taux d'identité de séquence de 23,5%, les CDK2 et les glycogène synthase kinases 3 $\beta$  (GSK3 $\beta$ ) sont de la même famille du kinome (CMGC). En 2004, une étude des relations structures-activités (SAR) des protéines kinases résolues dans la PDB a montré que ces deux types de protéines kinases avaient des activités comparables et qu'elles étaient inhibées par les mêmes types de molécules [184]. **MED-SMA les rassemble dans le même groupe. Ce résultat montre l'intérêt de cette nouvelle méthode de classification.**

Les sites d'autres familles sont aussi présents dans le tableau de résultats: autres couleurs (cf. Figure 42) telles que PTK (vert), CAMK (rose), TKL (gris) ou des protéines kinases atypiques (rouge clair). Le site de liaison à l'ANP de la tyrosine kinase Aurora-A (code PDB: 2DWB en vert) a un score plus élevé que d'autres CDK2, il est donc plus proche que certaines CDK2. MED-SuMo détecte donc des ressemblances structurales et

fonctionnelles entre les CDK2 et cette tyrosine kinase Aurora-A. Ces résultats sont en adéquation avec une étude récente qui a montré que les activités de ces deux types de protéines sont inhibées par les mêmes classes de molécule: les 1,4,5,6-tetrahydropyrrolo[3,4-c]pyrazoles [186]. Ces kinases se lient donc aux mêmes types de molécules, impliquant des similitudes fonctionnelles et structurales: leurs activités peuvent être inhibées ou activées par les mêmes types de principes actifs. MED-SMA détecte donc ce genre de similitude au sein d'un jeu de sites de liaison, cette méthode peut donc être utile pour la conception de médicament avec pour rôle de détecter d'autres protéines cibles qui pourraient être influencées par un certain type de molécule.

PDB Id	Ligand Id	SCF Count	Sumo Score	SumoSignature
1B38 (Query)	1B38 (Ligand 3 ATP)	38	0.000	
1B38	ATP 3	66	30.615	
1B39	ATP 3	50	24.424	
1HCK	ATP 2	46	22.784	
2CCH	ATP 2	30	15.621	
1FIN	ATP 1	27	12.759	
2DWB	ANP 2	22	11.864	
2CJM	ATP 3	25	11.698	
1GY3	ATP 7	20	11.243	
2CCI	ATP 3	21	10.845	
1QMZ	ATP 5	22	10.697	
1JST	ATP 3	20	9.756	
2PHK	ATP 4	17	9.670	
2SRC	ANP 1	16	9.303	
1MQ4	ADP 4	17	9.187	
1J1B	ANP 1	19	9.030	
1J1C	ADP 3	16	8.566	
1PYX	ANP 5	14	7.803	
1QSY	ADP 10	14	7.758	
1PHK	ATP 3	14	7.733	
2CN5	ADP 5	13	7.661	
1Q99	ANP 5	13	7.591	
2OID	ANP 1	13	7.059	
2A19	ANP 6	12	6.764	
1QL6	ATP 4	12	6.673	
2C6D	ANP 2	12	6.619	
1U5R	ATP 7	11	6.600	
1UA2	ATP 1	12	6.293	
1ZTH	ADP 5	13	6.040	
1OL6	ATP 1	11	5.794	
1OL5	ADP 7	10	5.786	
1ZP9	ATP 5	11	5.730	
2P0C	ANP 4	11	5.503	
1DS5	AMP 5	11	5.447	
2IVT	AMP 1	9	5.400	
1DAW	ANP 3	10	5.247	

**Figure 42 : Résultats d'une comparaison du site de liaison d'une cycline dépendante kinase 2 (code PDB : 1B38) avec tous les sites de liaison du jeu de données avec MED-SuMo.**

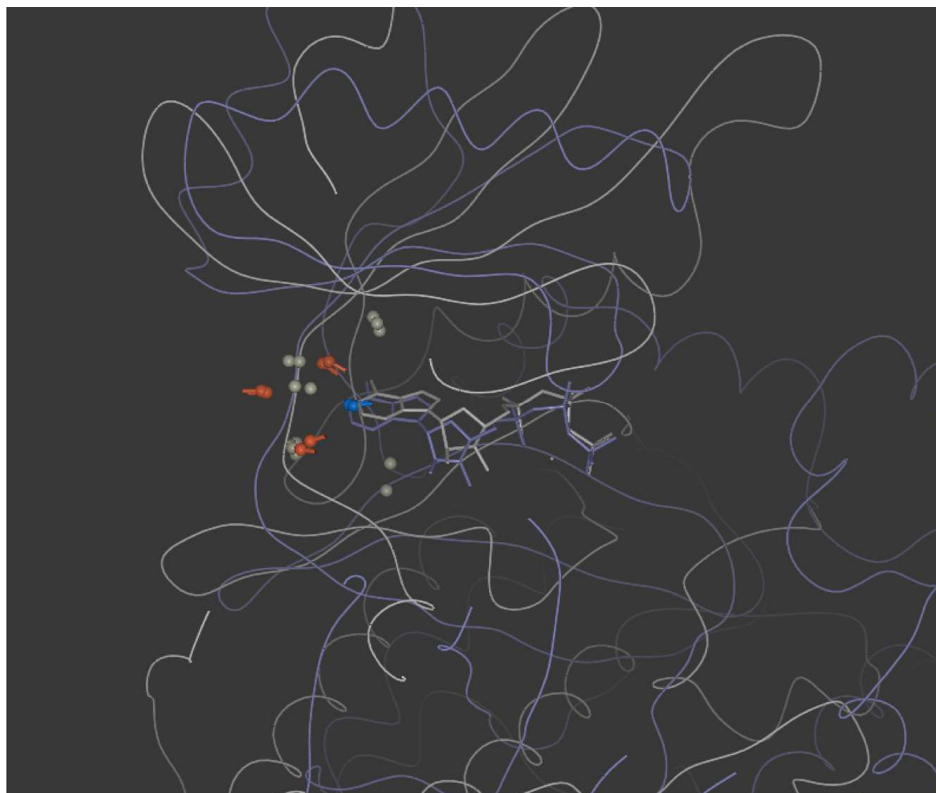
La ligne de la requête est jaune. Chacune des autres couleurs représente un famille de kinase : CMGC (bleu foncé pour CDK2, bleu clair pour les autres types de CMGC), PTK (vert), CAMK (rose), TKL (gris) et les protéines kinases atypiques (rouge clair). Les lignes blanches sont des protéines kinases d'autres espèces, elles ne sont pas associées aux protéines du kinome.

La dernière remarque concerne le 60<sup>ème</sup> site du groupe 157, site de liaison de l'enzyme biotine carboxylase (Code PDB 1DV2). Le motif PROSITE qui lui est associé confirme



---

l'appartenance à la superfamille des ATP-grasp. Aucune étude à ce jour n'évoque une quelconque relation entre les protéines kinases et la biotine carboxylase. Cependant, la superposition de la CDK2, 1QMZ et 1DV2 montre clairement des similitudes dans la région charnière du site de 1QMZ (cf. Figure 43).



**Figure 43 : Superposition 3D d'une biotine carboxylase et d'une CDK2.**

La protéine CDK2 (code PDB 1QMZ) est en gris et la biotine carboxylase (code PDB 1DV2) en violet. Les similitudes détectées par MED-SuMo sont situées dans la région charnière des sites.

### **Enrichissement des MED-clusters**

Initialement, PROSITE avait été utilisé pour constituer un second jeu de données de protéines se fixant aux ligands puriques. En sélectionnant des motifs liés aux ligands choisis, 3515 structures protéiques ont été rassemblées. 880 sont communes au jeu de données construit avec la PDB, 1492 ne sont pas co-cristallisées avec des ligands puriques et 1143 ne sont co-cristallisées sans aucun ligand. Afin d'essayer de caractériser ces structures sans ligand, nous avons décidé d'essayer de les comparer aux 247 groupes déjà identifiés par MED-SMA. Pour ce faire, nous avons utilisé MED-SuMo en mode surface entière des protéines contre les 2322 sites à purines (*full* versus *sites*). Deux critères ont été établis pour qu'un site soit associé à un groupe : (i) le score MED-SuMo doit être supérieur à 5,5, et (ii) au moins 60% des SCFs doivent être communs avec le site du groupe auquel il ressemble,

---

impliquant des similitudes structurales et fonctionnelles (paramètre *covering\_factor*). Le MED-sumo-client batch (cf. Figure 20) est utilisé pour lancer les simulations en ligne de commande.

En ne considérant que le premier filtre, 567 structures sans ligand (soit plus de la moitié) sont associées aux MED-Clusters qui seraient enrichis de 1038 sites. Certaines structures sont associées à plusieurs groupes et plusieurs sites de liaison peuvent être détectés dans une même structure. À l'issue du second filtre, seulement 196 structures sur les 1143 sont assignés aux groupes précédents. Ils correspondent à 203 sites de liaison. Sept structures sont associées à plus d'un groupe, ce qui est possible pour deux raisons : (i) tout d'abord une protéine peut avoir plusieurs sites puriques, (ii) une protéine peut être associée à deux groupes déjà connectés. Par exemple, 1FMK est associée à deux MED-clusters de protéines kinases ; les groupes 157 et 211. Ils contiennent déjà des *multipatches* issus d'un même site de liaison commun à l'ANP du facteur d'initiation  $\alpha 2$ , (code PDB 2A19).

L'enrichissement concerne 56 groupes. Par exemple, le groupe 40 augmente de 19 nouveaux sites. Les structures des protéines kinases qui sont le plus souvent résolues sans ligand ou co-cristallisés avec des inhibiteurs sont bien détectées par MED-SuMo. En effet, les MED-clusters de protéines kinases 157 et 211 sont les plus enrichis. Ils gagnent respectivement 26 et 9 sites. Ces nombres ne sont pas très élevés mais ceci est principalement dû à la rigueur des paramètres fixés. En effet, nous avons utilisé les mêmes paramètres pour MED-SMA (*score\_min* = 5,5, *covering\_factor* = 0,6) afin de (1) rassembler des sites similaires, (2) d'inclure quelques connexions entre groupes et (3) d'éviter le plus d'erreurs (faux positifs) possibles. Le reste des structures sans ligand n'est associé à aucun groupe, toutefois, au moins 371 ont des similarités avec des sites puriques. Des surfaces d'interaction aux purines ont donc été détectées par MED-SuMo, cependant il faudrait les analyser plus précisément pour pouvoir en tirer des conclusions plus précises.

Ainsi, seulement 1/5<sup>ème</sup> des structures sans ligand sont donc associées aux groupes. Cependant, les paramètres ont délibérément été fixés rigoureusement afin de ne pas induire de fausses associations aux groupes déjà fonctionnellement spécifiques.

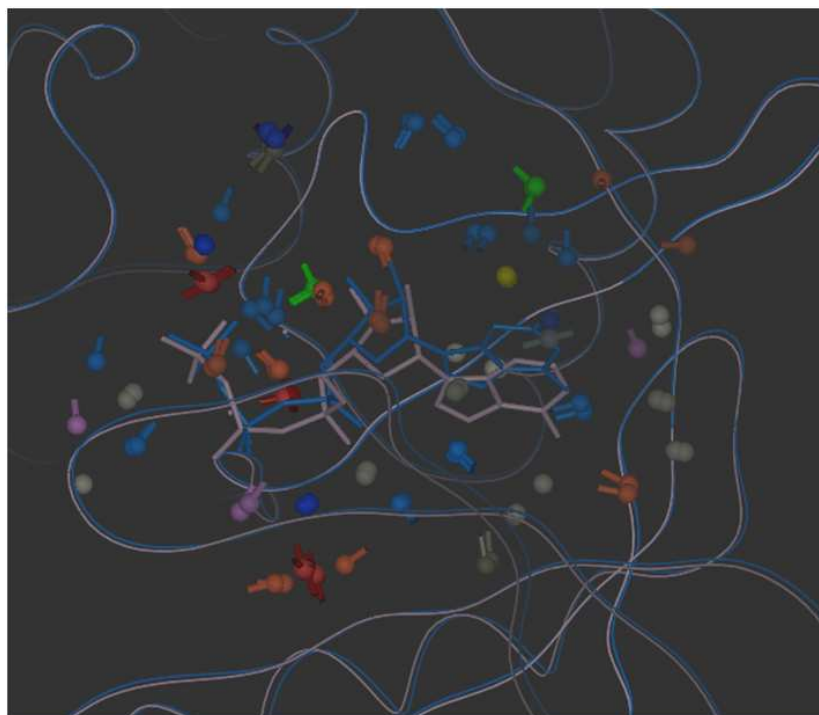
## Discussion

Pour cette étude, MED-SMA est appliquée à une famille de protéines particulières, appartenant au purinome. AXP, NAD et GXP sont des ligands simples, composés de nucléotides de même type et de groupements phosphates. Cependant, ces ligands sont flexibles et peuvent adopter différentes conformations dans les sites de liaison [187]. En effet,



---

malgré de fortes similitudes, certains sites de liaisons sont très sélectifs, alors que d'autres peuvent se fixer à différents ligands. Un exemple est représenté dans la Figure 44 où deux CK2 très semblables sont superposées avec leurs ligands respectifs. Le GTP est superposé sur le ligand ANP. Des SCFs sont visibles tout autour des ligands, ce qui montre de manière explicite que les sites de liaison sont globalement très proches.



**Figure 44 : Superposition des ligands GTP et ANP dans deux sites de liaison très similaires.**

Les deux ligands sont extraits des structures de CK2 (code PDB 1DAY en bleu et 1LP4 en rose). Les structures sont très proches mais elles sont co-cristallisées avec des ligands différents.

D'un point de vue méthodologique, il est difficile de comparer les résultats de notre classification à ceux d'autres méthodes. Nebel et ses collaborateurs [188] ont reporté une méthode qui génère automatiquement des motifs 3D issus de sites de liaison de structures protéiques qui se basent sur un consensus des positions des atomes. Ils évaluent leur méthode sur un jeu de ligands ayant une adénine et la valide en créant 18 motifs différents. Notre étude englobe un nombre beaucoup plus grand de protéines. Les différentes classes présentées dans leur étude se retrouvent par MED-SMA. Néanmoins, la classification présente des différences. Concernant le motif ADP4 (cf. figure 3 de l'article [188] et le texte associé), trois protéines sur cinq (1EHI, 1E4E, 1GSA, 1KJQ and 1IAH) sont associées au même groupe, le MED-cluster 5 (1EHI, 1E4E, 1KJQ). Les deux autres protéines (1GSA et 1IAH) sont associées à d'autres MED-clusters. Cette comparaison ne peut être étendue car le reste des deux jeux de données est très différent.

---

La classification des 2322 sites de liaison forme 247 groupes. La méthode est particulièrement robuste. En effet, chaque *multipatch* contient en moyenne 48 SCFs qui sont principalement impliqués dans des liaisons hydrogène et en moyenne 80% des SCFs des sites de liaisons sont présents dans les *multipatches*. De manière intéressante, un nombre conséquent de groupes sont composés de sites liés à différents types de ligands. Toutefois, même si GXP et AXP ne diffèrent que par deux groupements chimiques, et que NAD et AXP ont une partie strictement similaire. Les groupes sont assez spécifiques.

L'annotation manuelle des structures du jeu de données basée sur le champ MOLECULE des fichiers PDB a montré que la moitié des groupes ne sont associés qu'à une fonction et 12 % à plus de deux fonctions. Les sites à purine ne sont généralement pas associés aux fonctions mais aux mécanismes d'activation ou d'inhibition. Les sites de liaison rassemblés dans les mêmes groupes peuvent donc avoir une affinité pour les mêmes types de molécules actives. Nous avons analysé et présenté plusieurs groupes avec un  $N_{eq}$  élevé. Le MED-cluster 40 a la valeur de  $N_{eq}$  la plus élevée, 53. Il contient les sites de liaisons des 279 petites protéines G et il est associé à plus de 100 autres annotations fonctionnelles. Cependant, il rassemble toutes les petites protéines G du jeu de données, toutes activées et inhibées par le même mécanisme d'activation/inhibition. La présentation des liens entre les groupes souligne la complexité de l'approche. En effet, un lien est représenté par un site dont au moins deux sous-sites se retrouvent dans deux groupes distincts. Cette propriété n'est pas une simple distinction due à la flexibilité des protéines, mais une réelle distinction dans les sites de liaison des protéines, qui est considérée comme des similitudes de sous-poches.

Un résultat pertinent de cette étude est l'analyse de la distribution de motifs PROSITE au sein de MED-clusters. PROSITE est une base de données de motifs de résidus basée sur la séquence, particulièrement utilisée par la communauté scientifique. Même si un nombre important de groupes est associé à plusieurs motifs, ils sont assez homogènes car les *patterns* sont souvent redondants ou apparentés. Notre étude consolide le travail créatif de Kasuya et Thornton qui analysent une représentation en 3D des motifs PROSITE sur les structures des protéines [189]. Ils montrent que de nombreux motifs présentent des caractéristiques structurales similaires qui pourraient être utilisées pour créer des modèles de motifs fonctionnels tridimensionnels. Wu et ses collaborateurs [190] ont aussi récemment amélioré une étude [191] en montrant que lorsque l'information de la structure est disponible, elle est plus pertinente que les motifs PROSITE simples. Notre travail suggère que des caractéristiques communes et distinctes peuvent être associées à un motif donné, et que différents motifs peuvent partager des similitudes locales. De même, un tiers des protéines

---

inclues au jeu de données ne sont associées à aucun motif PROSITE, notre approche souligne l'intérêt d'enrichir ces annotations avec celles de structures ayant des similitudes structurales et fonctionnelles.

Ce type de relation entre les familles de protéines, mise en avant par l'approche MED-SMA, est très intéressant. Dans cette classification, nous avons choisi de fixer un score MED-SuMo minimal élevé (5,5 qui correspond à une détection d'au moins 10 SCFs communs) dans le but d'obtenir des groupes fonctionnellement spécifiques et de limiter le nombre de faux-positifs dans les groupes. Cette classification permet de proposer des mécanismes enzymatiques pour des protéines peu étudiées ou des protéines résolues récemment, ou encore la déduction de la fonction de protéines. En outre, il est possible de proposer une fonction aux protéines inconnues en détectant les groupes/clusters de sites ayant de fortes similitudes aux niveaux de leurs sites fonctionnels. En associant une structure à un groupe plutôt qu'à une seule structure, le nombre des faux positifs d'annotation fonctionnelle devrait être limité. Cette approche permet d'orienter de manière supervisée une démarche expérimentale qui viendrait confirmer les hypothèses issues de cette approche.

## **Conclusion**

MED-SMA est une approche de génomique structurale. Elle est rapide et comme noté par Ferrè et ses collaborateurs [192], les *patches* fonctionnels associés à une grande collection de cavités sur les surfaces des protéines peuvent être utilisés pour fournir des indices fonctionnels aux protéines dont la fonction n'est pas connue. Cette longue observation s'applique à notre étude. MED-SMA est une approche qui pourrait améliorer l'efficacité et le rendement des premières étapes des voies de la recherche de médicaments, comme le choix des premières pistes ou en améliorant les pistes pauvres dans la sélection de cible au profil favorable. Cette étude met en évidence l'utilité de MED-SuMo pour à la fois annoter les structures des protéines mais aussi permettre la classification fonctionnelle et structurale d'un jeu de sites de liaison.

---

## 4. Adaptation de la méthode pour la classification de gros jeux de données

POPS (PétaOpérations Par Seconde) est un projet du pôle de compétitivité SYSTEM@TIC [145] dont le but est de concevoir des systèmes informatiques hardwares adaptés à un très large spectre de besoins en calcul haute performance, vers le milliard de milliard d'opérations flottantes par seconde. Ce projet est dirigé par Bull et implique 18 partenaires. Des industriels tels EDF ou Eurodecision, des entreprises comme MEDIT et des laboratoires académiques tels le CEA LIST (Laboratoire d'Intégration des Systèmes et des Technologies) ou l'IBBMC (Institut de Biochimie et Biophysique Moléculaire et Cellulaire) y participent. Pour MEDIT, l'objectif est de classer tous les sites de liaisons de la PDB, soit 90000 sites environ, avec l'approche MED-SMA en utilisant des machines HPC (Calculateur à haute performance ou *High Performance Computing*) très puissantes mises à notre disposition par la société Bull.

Après les classifications des protéines du *fold* GHKL (146 sites) et des protéines liant les purines (2322 sites), des tentatives pour classer des jeux de sites plus grands (3500 sites de lectines) avaient été testées sur des machines plus ou moins puissantes comme une machine *AMD bi-Opteron Dual Core 270, 6GB RAM* et une *Intel bi-Xeon Quad Core 5335, 16 GB RAM*. Cependant, l'implémentation de la méthode n'était pas particulièrement adaptée aux très grands jeux de données. En effet, la première étape de comparaison multiple des sites de liaison s'effectue exclusivement dans la mémoire physique de la machine sans jamais construire de fichier intermédiaire. Les résultats étant retournés déjà ordonnés sous forme d'un fichier binaire global lisible seulement par MED-SuMo. Sur un jeu de 3500 sites, plus de six millions de comparaisons sont nécessaires ( $3500 * 3500 / 2$ ). La mémoire vive des machines utilisées se sature au bout de quelques heures de calcul sans qu'aucun résultat ne soit retourné, le *garbage collector* d'Ocaml n'arrivant plus à libérer suffisamment d'espace mémoire. Afin de pouvoir traiter les 90000 sites de la PDB, il était nécessaire de chercher une solution technique soit au niveau des options de compilation, soit au niveau de la distribution des calculs. Cette première observation nous a permis de réfléchir sur ce problème et, aussi d'anticiper d'autres problèmes d'implémentation concernant la classification de grands ensembles de sites.

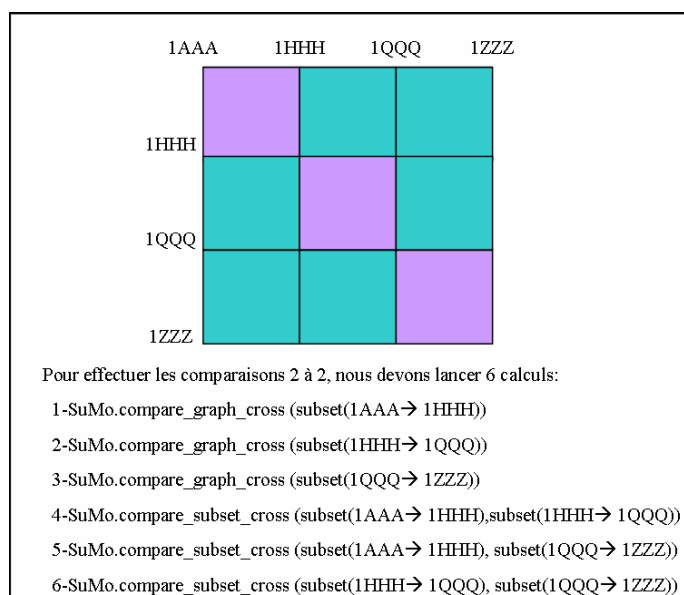
Comme décrit dans la partie III.iv, deux fonctions sont nécessaires pour le lancement de MED-SMA: *SuMo.compare\_graph\_cross* et *SuMo.multi*. Il faut remarquer que d'une part ces

---

deux fonctions communiquent par l'intermédiaire d'un seul fichier binaire et qu'ensuite la fonction *SuMo.multi* gère une longue liste de tâche en mémoire, et sans aucune étape intermédiaire. Jusqu'à maintenant, la procédure globale de classification des jeux de sites d'interactions ne posait pas de problème de performance, l'étape de comparaison deux à deux pour le jeu de purines a nécessité en environ 3h30 sur notre serveur bi-processeur (*bi-Xeon Quad Core 5335, 16 GB RAM*) et le reste de la procédure (création du graphe et classification par MCL) ne prenait que 30 minutes (sur une machine). Cependant, les performances globales de MED-SMA ne sont pas linéaires mais plutôt quadratique par rapport au nombre de sites à comparer et à classer. Ainsi, pour des jeux de données plus importants que ceux déjà utilisés, plusieurs jours de calculs peuvent être nécessaires. Il faut donc anticiper des arrêts inopinés de machines sans que cela implique obligatoirement de relancer la totalité des calculs. Ainsi, des étapes intermédiaires, et un autre mode stockage des résultats intermédiaires et finaux sont indispensables pour que MED-SMA soit lancé sur des grands jeux de données.

#### i. *MED-Distribute*

La comparaison multiple se lance avec un script MED-sumo-lua. Lorsque le noyau (*kernel*) de la machine utilisée est doté de plusieurs CPU et cœur de calcul, la fonction *SuMo.compare\_graph\_cross* permet de distribuer les calculs des comparaisons 2 à 2 en autant de processus séparés. Malgré cette première optimisation, la comparaison multiple d'un jeu de 3500 sites n'est pas possible, le calcul finit par saturer la mémoire et s'interrompt. Il est donc difficile d'envisager la comparaison multiple des 90000 sites de la PDB.



**Figure 45 : Exemple de découpage du jeu de données pour optimiser l'étape de comparaison multiple de MED-SMA.**

Les résultats des comparaisons deux à deux des sites de liaisons sont indépendants les uns des autres. Il n'est donc pas indispensable que toutes les comparaisons soient réalisées par le même script, l'essentiel étant la fusion des résultats finaux. La Figure 45 illustre la manière dont nous découpons le jeu de données afin de passer outre la saturation de la mémoire physique évoquée précédemment. Pour ce faire, la liste des graphes composant la base est récupérée. Pour l'exemple de la Figure 45, les graphes ont des identifiants allant de 1AAA à 1ZZZ. Cette liste est découpée en ensembles de taille fixe (par exemple des carrés de 2000 graphes). La nouvelle implémentation de la comparaison multiple intègre les deux fonctions suivantes:

- la fonction **SuMo.compare\_graph\_cross** qui lance les comparaisons multiples au sein de chaque bloc diagonal de la matrice (cases en violettes dans la Figure 45).
- la nouvelle fonction **SuMo.compare\_subset\_cross** qui permet de comparer tous les graphes d'un ensemble à tous les graphes d'un autre ensemble (cases vertes dans Figure 45), soit les blocs hors-diagonales.

Alors qu'une seule fonction était utilisée pour lancer les comparaisons multiples du jeu de sites (*SuMo.compare\_graph\_cross* (1AAA→1ZZZ)), le programme est découpé en plusieurs appels de fonctions différentes (cf. Figure 45). Les résultats de chaque appel sont stockés dans un fichier binaire et interprétés dans les étapes suivantes. Dans le cadre du projet POPS, la société Bull a mis à notre disposition une ferme de PC (cluster) HPC Xeon 64bits, composé de 36 nœuds de calculs, chacun disposant d'au moins huit processeurs. Afin de

prendre avantage de cette puissance distribuée sur les 36 noyaux linux associés, le programme *MED-Distribute* a été développé au sein de MEDIT-SA. Il permet de distribuer l'exécution d'un programme sur plusieurs nœuds de calcul et de récupérer les résultats générés. L'utilisateur découpe la requête principale en un nombre fixé de requêtes et *MED-Distribute* lance chacune d'elles sur un nœud de calcul différent. Chaque requête est un appel à une des fonctions de comparaisons multiples. Ces fonctions permettent aussi la distribution des calculs sur les CPUs disponibles, toute la puissance des nœuds de calcul est ainsi utilisée.

Pour chaque requête, un fichier binaire contenant les résultats des comparaisons effectuées est généré. Les informations sont ensuite directement extraites et stockées dans la base de données *multi.db* décrite dans la partie suivante.

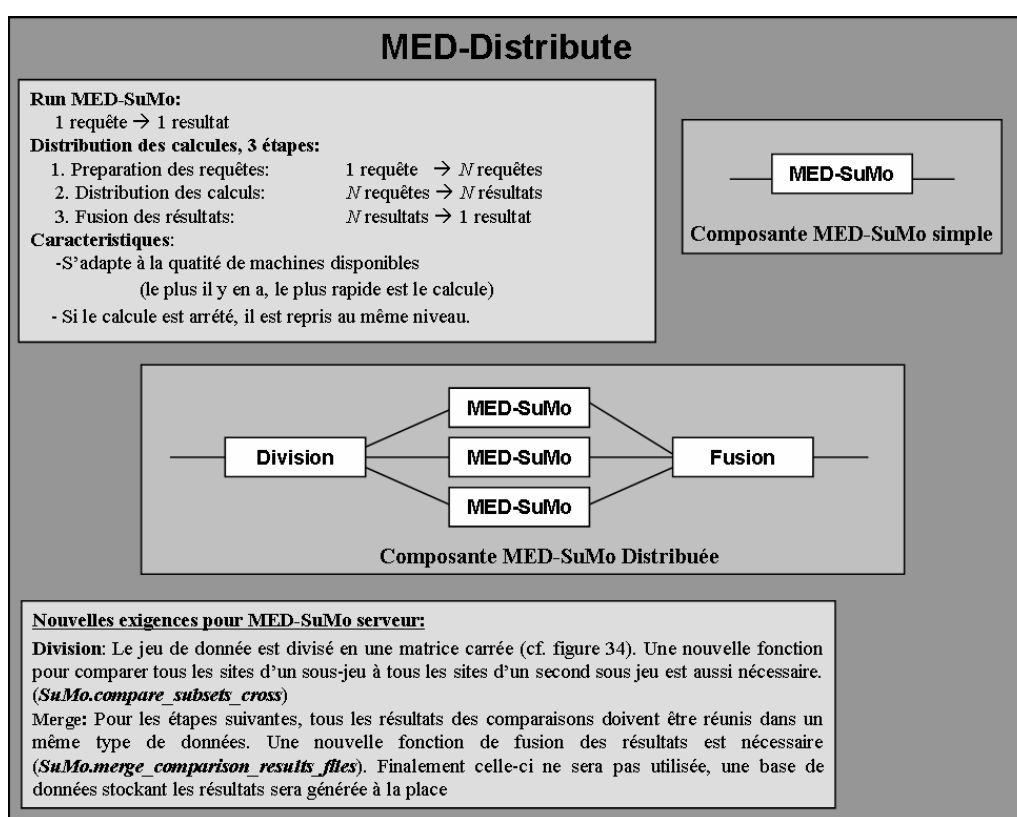
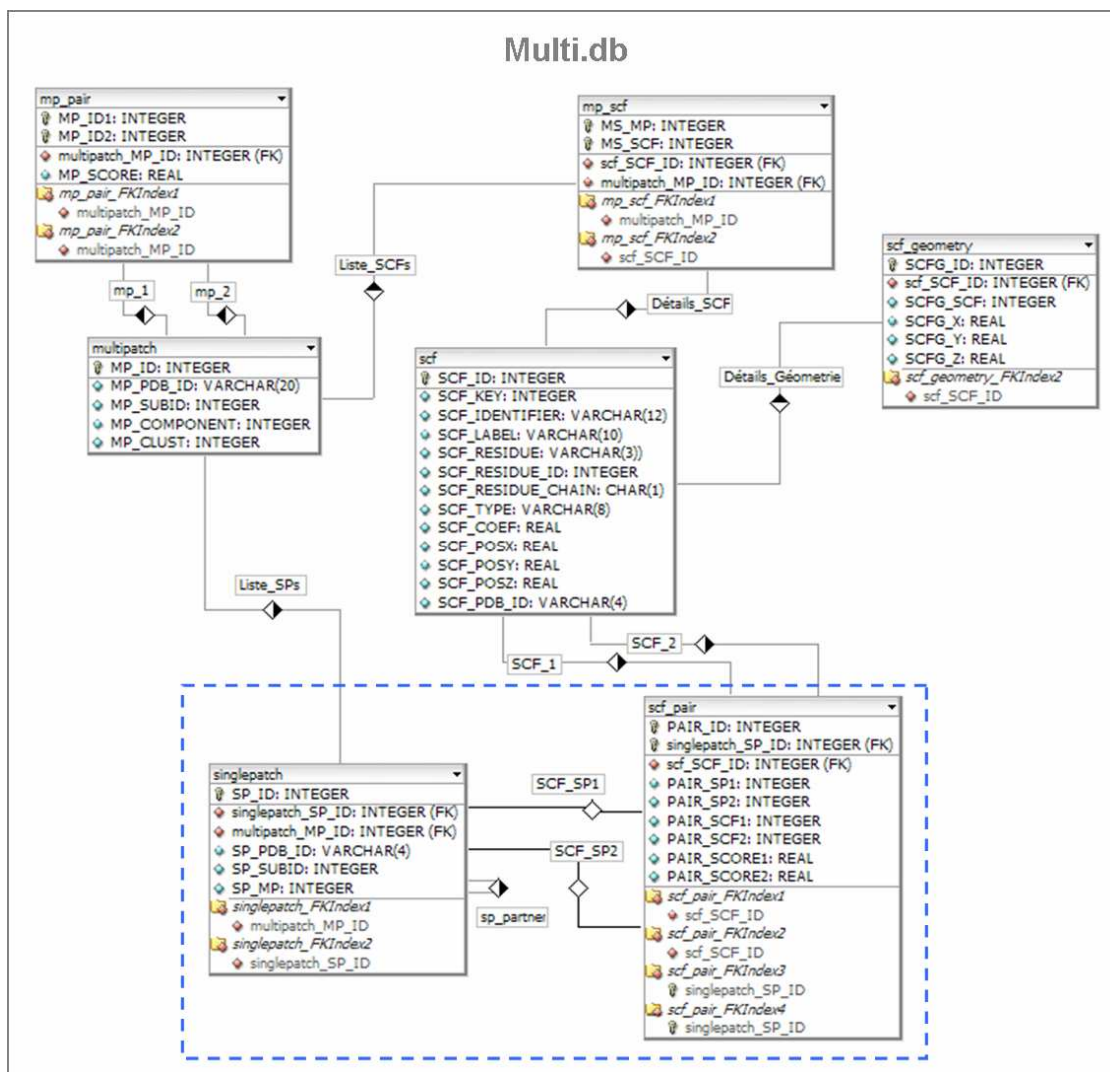


Figure 46 : Fonctionnement du programme *MED-Distribute*.

## ii. Stockage des résultats

Les résultats de comparaisons multiples contiennent différents types de données. MED-SMA considère à la fois le score MED-SuMo entre les sites mais surtout les similarités structurales 3D pour former les *singlepatches*. Les positions, les orientations, ainsi que les types des SCFs sont indispensables pour les étapes suivant les comparaisons multiples. La

Figure 47 représente le schéma global de la base de données *multi.db*. Elle stocke les résultats intermédiaires des comparaisons multiples ainsi que les résultats finaux d'une classification.



**Figure 47 : Architecture de la base *multi.db*.**

Sept tables répertorient d'une part l'ensemble des informations nécessaires à la classification des sites de liaisons et d'autre part les résultats finaux de la classification.

La base *multi.db* se compose de sept tables. Elles répertorient les éléments nécessaires à la classification des sites de liaisons et contiennent aussi les résultats finaux. La base est construite au moment de la lecture des résultats des comparaisons multiples générées par le programme *MED-Distribute*, elle est remplie au fur et à mesure du protocole :

- Les tables *singlepatch*, *scf\_pair*, *scf* et *scf\_geometry* sont remplies au moment de la création des *singlepatches* (cf. Figure 24c). La table *singlepatch* répertorie les informations sur les *singlepatch* : par exemple, la structure dont chaque *singlepatch* est issu, leurs partenaires détectés par les



---

comparaisons multiples. La table *scf\_pair* répertorie les SCFs en commun entre paire de *singlepatches*. Les tables *scf* et *scf\_geometry* contiennent des informations sur les SCFs des *singlepatches* tels que leurs identifiants, types, positions, et leurs géométries, permettant ainsi la représentation des SCFs dans un visualisateur 3D. Les tables *scf\_pair* et *singlepatch* (encadrées en bleu dans la Figure 47) ne servent qu'aux étapes intermédiaires, et sont inutiles pour les résultats finaux de la classification. Elles sont cependant conservées pour que la classification puisse être mise à jour régulièrement en fonction des nouvelles structures de la PDB.

- Les tables *multipatch* et *mp\_scf* sont remplies au moment de la construction des *multipatches*. Certains champs restent vides tel *MP\_CLUSTER* qui correspond à l'identifiant du groupe auquel le *multipatch* appartient. Il n'est précisé qu'une fois la classification terminée.
- La table *mp\_pair* fait le lien entre les *multipatches*. Pour chaque paire, le score MED-SuMo est calculé et stocké. La formation du graphe de similarité se fait par une requête sur cette table.

Ces modifications permettent maintenant de lancer la classification de tous les sites de la PDB. Nous avons pu en effet éliminer les limites techniques du protocole. Le travail n'est pourtant pas terminé, une longue phase d'analyse commence. Le fait que les résultats intermédiaires soient disponibles dans *multi.db* nous permet de lancer plusieurs classifications avec des matrices de similarité normalisées de différentes manières afin de choisir les meilleurs paramètres possibles. Les valeurs  $N_{eq}$ , spécificité et sensibilité seront utilisées pour évaluer la classification (cf. partie III.B.2.v.). Certains résultats intéressants des applications décrites précédemment seront aussi recherchés afin d'évaluer les clusters. Le projet POPS prévoit aussi le développement de deux interfaces graphiques permettant la visualisation globale de la classification et une analyse spécifique des groupes: une interface propriétaire utilisant l'architecture de MED-SuMo GUI et dont les spécifications sont en cours et une interface web dont le développement est déjà commencé et qui sera disponible pour la communauté scientifique. La base *multi.db* sera au cœur des deux interfaces et permettra par exemple d'accéder au contenu des groupes ou de superposer les *multipatches*. Le but de ces interfaces est de permettre une analyse et exploration complète de la classification et surtout de comprendre les similarités fonctionnelles détectées par la méthode MED-SMA dans les sites de liaison de la PDB.

---

## **C. Une nouvelle méthode de conception de novo de molécules actives**

### **1. Idée Générale**

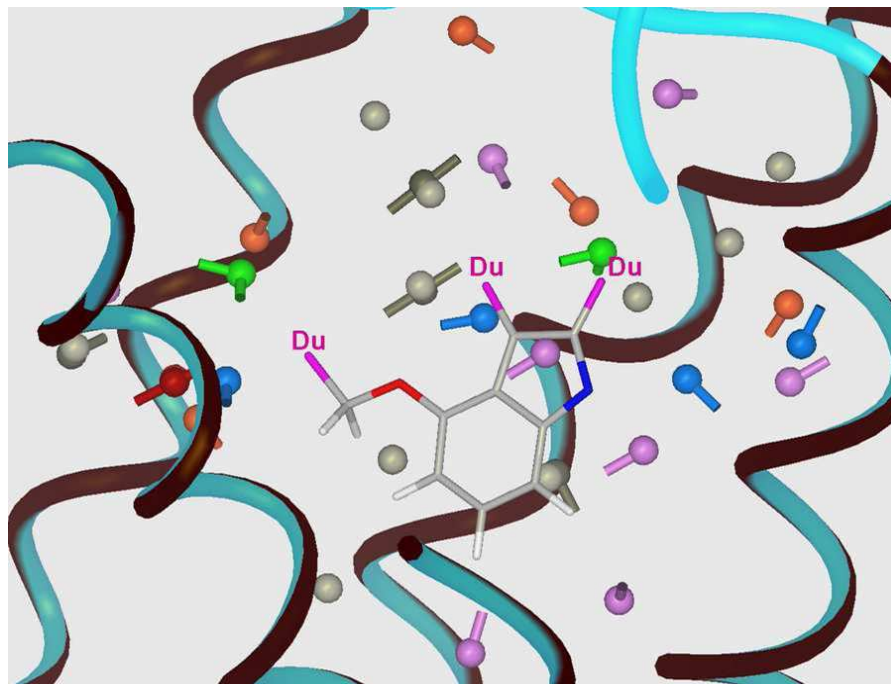
La PDB est une source majeure d'informations structurales expérimentales sur les interactions entre les ligands et les protéines. En dépassant en 2008 les 50000 enregistrements, elle constitue incontestablement une réelle richesse pour le chimiste médicinal. Les méthodes prédictives *Target-Based drug design* se basent sur ces informations biostructurales et proposent principalement les protocoles suivants: (1) le criblage virtuel qui positionne et calcule des scores pour un grand ensemble de molécules candidates en optimisant les interactions dans le site actif, (2) la génération *de novo* de candidat médicament en allongeant itérativement une molécule dans le site actif par optimisation des interactions avec la protéine, (3) le *fragment-based drug design* qui va prédire le positionnement de fragments dans et autour du site actif, fragments qui pourront servir de *hit* et par association générer des molécules candidates complètes. Au croisement des 2 dernières catégories, BREED [193] est une méthode de conception de molécules *de novo* qui se base sur les alignements structuraux de protéines d'une même superfamille pour superposer les ligands qui y sont fixés. Ces ligands, qui sont dans leurs conformations actives, sont ensuite recombinaisonnés en 3D par permutation de liaisons covalentes superposées pour former de nouvelles molécules. Ramensky et ses collaborateurs [194] ont été les premiers à remarquer que les similitudes locales des protéines détectées à l'échelle de la PDB avaient une influence positive sur les résultats de ces protocoles. En effet, si des similarités locales sont détectées entre paires de sites de liaison, la partie du ligand du second site (ou le fragment) se liant à la partie commune peut être transférée sur le premier ligand. Autant de combinaisons possibles qu'il y a de similitudes locales détectées dans les structures de la PDB. Ramensky et ses collaborateurs utilisent le concept de remplissage d'un site de liaison avec un nuage d'atomes. Une fonction de score basée sur ces nuages d'atomes permet ensuite de favoriser le positionnement de certains ligands dans des processus d'amarrage (*docking*), ou sert comme source d'information pour optimiser des molécules actives en les modifiant avec les atomes du nuage. Leur méthode présente toutefois deux limites : d'une part, des atomes et non des fragments de molécule sont utilisés, aucune information sur leur valence n'est intégrée. Il est donc difficile de les combiner pour former de nouvelles molécules, et la génération de molécule complète nécessite beaucoup plus de combinaisons que si des fragments sont

---

utilisés. Dans ce cas, 2 à 3 hybridations suffisent pour construire un nouveau ligand. D'autre part, la méthode de comparaison de l'environnement des ligands est basée sur des graphes d'atomes ce qui est beaucoup moins performant que les graphes de triplets de SCFs utilisés par MED-SuMo.

Dans ce cadre, MED-SuMo est un des rares logiciels [12,138] qui détecte les similitudes fonctionnelles de sites de liaison, et qui permet l'alignement des protéines cibles mais aussi des ligands dans les poches. Un de ces avantages majeurs est la possibilité d'utilisation d'un nouveau type de base ; une base MED-SuMo de sous-poches de liaison. L'analyse des méthodes de conception de nouvelles molécules combiné à l'approche originale de MED-SuMo ont mené au développement d'une nouvelle méthode de conception de molécules actives combinant la détection de similitudes locales des surfaces des protéines et une approche fragmentale [17]. Ce protocole propose trois applications avec (1) la génération d'une base de données de fragments ciblée sur une protéine donnée, (2) la conception *de novo* de molécules actives, et aussi (3) la possibilité d'optimiser des molécules déjà actives sur des protéines cibles précises. De plus, il se base sur l'ensemble des informations biostructurales de la PDB, ce qui maximise les probabilités de détecter des similitudes locales. Cette approche est particulièrement efficace à deux égards. D'une part, les bases MED-SuMo peuvent contenir des milliers de poches voire des centaines de milliers de sous-poches définies par des complexes protéine-fragment, ainsi que des centaines de milliers de descriptions de surfaces d'interaction locale et, d'autre part des requêtes de surfaces entières de protéines sont possibles.

Deux autres avantages sont également à relever. Le premier concerne la méthode de fragmentation appelée MED-Fragmentor qui génère de multiples motifs protéine-fragment à partir d'un ensemble de complexes protéine-ligand. La détection des structures 2D exactes issue d'une grande librairie de petite molécule sur les structures de ligand de la PDB permet d'identifier ces fragments. Ces motifs 3D, appelés MED-Portions, sont donc des sous-structures chimiques à la fois présentes dans une base de données de petites molécules actives **et** détectées dans une sous-structure de ligand fixé à une protéine. Les MED-Portions sont définis par plusieurs critères: (1) une sous-structure chimique dont les atomes ont la même topologie qu'une molécule issue d'une bibliothèque de molécules, soient des molécules synthétisable, (2) les atomes des extrémités des fragments pour marquer les connexions vers le ligand original: les anciens atomes sont remplacés par des *dummy* atomes qui n'ont pas de types mais qui conservent leurs positions, finalement (3) par la surface d'interaction de la protéine à proximité spatiale du fragment, donc décrite par les SCFs de MED-SuMo.



**Figure 48: Représentation d'un exemple de MED-Portions.**

Comprenant une sous-structure chimique, les atomes *dummy* aux extrémités qui étaient connectées, et les SCFs détectés autour du fragment sur la surface de la protéine.

La base de MED-Portions est construite en croisant donc tous les complexes protéine-ligand de la PDB avec un sous-groupe de petite molécule extraite de la chimiothèque *PubChem*. Son exhaustivité constitue le second avantage du protocole. D'avantage de détails sur sa construction seront décrits dans la partie suivante. Une fois générée, cette base de surface d'interaction protéine-fragment s'utilise de la même manière qu'une base MED-SuMo. Des surfaces d'interaction requêtes sont comparées à tous les éléments de la base. Les résultats (*hits*) détectés sont des MED-Portions ayant des SCFs communs avec la requête et formés d'une sous-structure chimique correctement positionnée pour interagir localement avec la surface requête. L'ensemble des MED-Portions récupérés est filtré pour garder les meilleurs candidats et les MED-Portions résultants sont ensuite combinés en 3D pour former de nouvelles molécules hybrides avec le programme MED-Hybridise. Cette approche permet donc le remplissage de n'importe quelle surface d'interaction par des MED-Portions qui, hybridés peuvent constituer de nouveaux ligands potentiels. Après la description complète du protocole, une application sur le récepteur protéine kinase VEGFR-2 illustrera l'utilité et l'efficacité de la méthode.

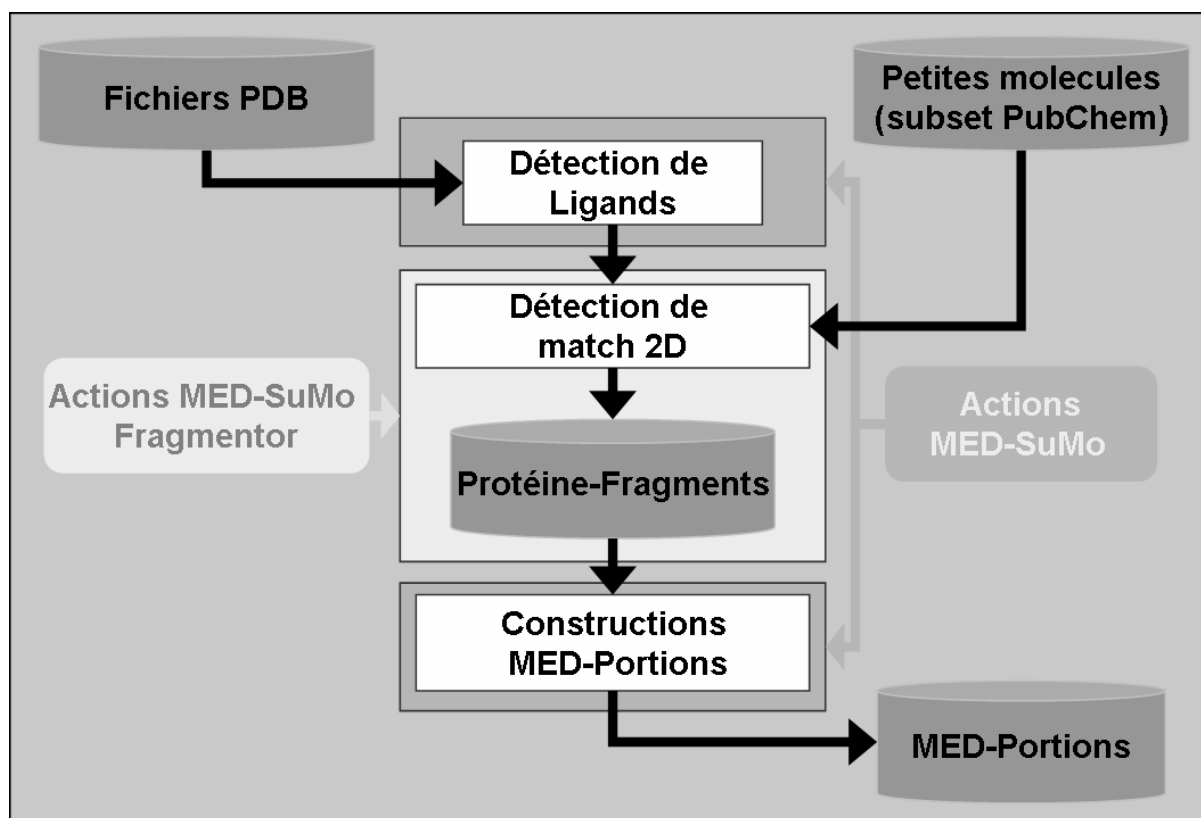
---

## 2. Mise en place du protocole

L'approche fragmentale est composée de trois parties essentielles. La première partie concerne la création du nouveau type de base MED-SuMo, la base *fragment*, composée de MED-Portions (cf. Figure 49). La seconde partie porte sur le lancement de *runs* MED-SuMo sur cette nouvelle base et la troisième partie permet la combinaison des fragments des hits détectés avec MED-Hybridise. Les paragraphes suivants contiennent une description de ces trois étapes. Les figures de cette partie sont extraites de la publication dont je suis co-auteur (article 3 [17]).

### i. Construction de la base *fragments*

De la même manière que les bases de sites de liaison (base *sites*) et de surfaces entières (base *full*), il faut avoir préalablement la PDB localement et référencée par l'intermédiaire de la base *pdb\_index.db* (cf. II.i). Pour chaque entrée de la PDB, un protocole unique est utilisé (cf. Figure 49): (i) détection des ligands, (ii) détection de sous-structures 2D dans ces molécules, fragmentation des molécules et stockage des informations dans une nouvelle base *frag\_info.db*, et (iii) construction de la base MED-SuMo. Trois sous-étapes sont nécessaires pour construire la base *fragments*:



**Figure 49 : Protocole de l'approche fragmentale pour la construction de la base de MED-Portions.**

Les rectangles représentent les traitements algorithmiques et les cylindres, les bases de données utilisées dans le protocole. La base « Fichiers PDB » est le *pdb\_index.db*, la base « Protéine-Fragments » est la base *frag\_info.db*, la base « Petites molécules » est la base *PubChem Compounds* et la base de « MED-Portions », la base *fragments*.

### Première étape : Détection des Ligands

Un ligand se définit par trois caractéristiques : (i) une molécule chimique ayant moins de 100 atomes lourds caractérisés par le champ *HETATM* dans le fichier PDB, (ii) une chaîne homo-peptidique ou hétéro-peptidique ayant moins de 10 résidus (acides aminés ou autres), (iii) un ligand lié de manière covalente à la protéine, défini par le champ *HETATM*. La détection des ligands par MED-SuMo serveur se fait avec la commande `MED-sumo-clui` (cf. Annexe 2):

```
# sumo pdb frag_lig_info ID_PDB
```

Elle transmet une liste d'informations sur les ligands et les atomes les constituant (tels que les noms de ligands, ou les positions des atomes) au programme MED-Fragmentor par l'intermédiaire d'un fichier texte au format PDB.

---

## Deuxième étape : MED-Fragmentor : Détections des sous-structures 2D et stockage des fragments

Le rôle de MED-Fragmentor est de construire l'ensemble des MED-Portions détectés sur chaque ligand de complexe. Un MED-Portion est la représentation de MED-SuMo de motifs fragment-protéine, ses caractéristiques ont été définies dans la partie III.C.1. Les ordres de liaison sont attribués aux ligands en fonction de leur géométrie avec le programme OpenBabel 2.2.0 [143,195] et les informations sont stockées sous la forme d'un fichier SD. A cette étape, aucune correction des ordres de liaison n'est effectuée.

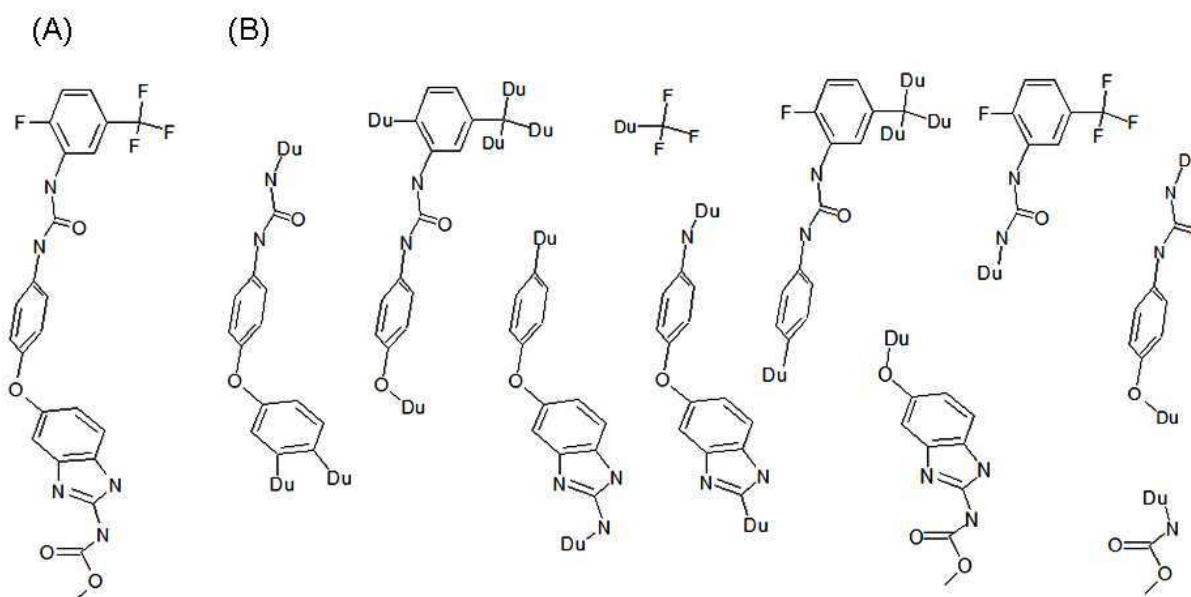
Pour qu'un ligand issu de la PDB soit sélectionné, il faut qu'il soit composé d'au moins 6 atomes. Ainsi le glycérol ( $C_3H_5(OH)_3$ ) est analysé alors que d'autres molécule comme l'ion phosphate ( $PO^{3-}$ ) ou l'imidazole ( $HC_3H_3N_2$ ) sont directement considérés comme le fragment d'un MED-Portion. L'identification des fragments s'effectue au travers d'une détection topologique de sous-structures 2D entre chaque ligand de la PDB et chaque fragment importé à partir d'une chimiothèque donnée.

*PubChem Compounds* [196] est une base de petites molécules annotées d'informations sur leur origine (fournisseur) et éventuellement certaines activités biologiques. Un filtre préalable permet d'éliminer les molécules considérées comme indésirables. Les molécules : (1) contenant d'autres atomes que [H, B, C, N, O, F, P, S, Cl, Se, Br, I, As, Te, Si], ou (2) ayant un poids moléculaire supérieur à 250 Da ou (3) ayant un poids moléculaire inférieur à 70 Da sauf celle contenant un cycle, sont éliminées. Finalement, 2 112 444 molécules sur 19 202 121 sont sélectionnées ; 1 577 071 en éliminant les doublons topologiques.

La génération des MED-Portions se fait par un algorithme de détection de sous-structures basé sur la théorie des graphes, en recherchant la plus grande sous-structure possible dans un ligand à partir de la structure 2D complète d'un fragment provenant d'une chimiothèque. Cet algorithme utilise les états d'hybridation des atomes similaires détectés et néglige les ordres de liaison explicites, la chiralité (comparaison 2D) ainsi que les atomes d'hydrogène. Les sous-structures détectées sont classées par ordre décroissant de leur nombre d'atome lourd. Lorsque des matches similaires sont détectés, les doublons sont éliminés afin de limiter la redondance (cf. règle dans [17]). Les coordonnées cartésiennes des atomes de chacune des sous-structures détectées sont conservées. L'extraction d'une sous-structure d'un ligand implique le découpage de certaines liaisons et la génération de liaison avec une valence libre. Lorsqu'une liaison est coupée, un atome factice (*dummy atom*) est ajouté à la sous-structure sur l'atome pointé par la liaison coupée, permettant ainsi la conservation de la trace où le ligand original était connecté. Les coordonnées cartésiennes du *dummy* atome sont

celles de l'atome découpé. Chaque correspondance est stockée dans la nouvelle base de données *frag\_info.db* (représentée par la base « Protéine-Fragments » dans la Figure 49), et servira pour la génération de la base MED-SuMo de MED-Portions.

Ce protocole permet la fragmentation de 81% des 11 011 ligands uniques de la PDB. Leurs atomes sont décrits par 13.8 MED-Portions en moyenne (18% par un MED-Portions, 46% par cinq MED-Portions et 65% par un à dix MED-Portions). Un exemple de la fragmentation du ligand GIG de la protéine kinase VEGFR-2 (Code PDB 2OH4 [197]) est montré Figure 50.



**Figure 50 : Description du ligand GIG par 10 MED-Portions.**

(A) structure du ligand GIG. (B) structures des MED-Portions représentés sans les SCFs afin que la figure reste lisible.

Les deux premières étapes aboutissent à la création de la base *frag\_info.db*, qui contient les informations nécessaires à la création de la base MED-SuMo de MED-Portions. La commande suivante permet le lancement du processus de fragmentation :

```
# sumo frag3d init
```

Comme cette étape nécessite plusieurs jours de calcul, une autre commande permet la reprise du processus afin que la fragmentation puisse reprendre si la machine est « malencontreusement » interrompue :



---

```
# sumo frag3d recover
```

### Troisième étape : Construction des MED-Portions et stockage dans la base *fragments*

Une base de sites MED-SuMo garde les références aux ligands des structures dans le *pdb\_index.db*. Pour chaque ligand, les SCFs sont détectés et un graphe MED-SuMo est généré et stocké. La base de fragments est construite en utilisant les sous-structures 2D détectées, exprimées avec leurs coordonnées 3D et stockées par le programme MED-Fragmentor. Pour chaque fichier PDB, une liste de fragments 3D est stockée dans la base *frag\_info.db* avec la position de chaque atome. Leurs positions permettent la détection des SCFs dans un rayon de 4.5 Å sur la surface de la protéine concernée. Pour chaque fragment, un graphe MED-SuMo est généré et stocké dans la base. La commande MED-sumo-clui suivante est utilisée :

```
# sumo graphdb update frags
```

L'enchaînement de ces trois étapes aboutit à la génération de la base MED-SuMo de MED-Portions. La base finale contient 395 360 MED-Portions.

### ii. *Cartographie des surfaces d'interactions avec des MED-Portions*

#### Première partie : Lancement d'un calcul MED-SuMo

La base de MED-Portions s'utilise de la même manière que les autres bases MED-SuMo, principalement par le MED-SuMo GUI ou avec MED-sumo-client batch. La surface d'une protéine est utilisée comme requête et MED-SuMo détecte les MED-Portions adoptant des surfaces d'interactions similaires. Les fragments issus des MED-Portions sont alors positionnés dans le repère de la protéine requête. Comme précisé dans la partie I.D.1, l'utilisateur est libre de construire toute sorte de requêtes.

Un MED-Portion détecté par MED-SuMo partage donc au moins 3 SCFs communs avec la requête. Comme les SCFs sont dans un rayon de 4.5 Å autour du fragment, les groupements chimiques qu'ils représentent, contribuent fortement aux interactions protéine-fragment locales. La liste des 50 000 meilleurs MED-Portions classés par score MED-SuMo décroissant peut être exportée de l'interface graphique dans un format XML ou dans un format SD annoté [198] en éliminant au préalable les doublons 3D. Un *run* dure en moyenne

---

vingt minutes si la requête est un site de liaison et une heure lorsque la requête est une surface entière de protéine (sur la machine *bi-Xeon Quad Core 5335, 16 GB RAM*).

### Seconde Partie : Sélection et analyse des MED-Portions

Les MED-Portions peuvent être filtrés, sélectionnés et analysés de différentes manières en fonction (1) des descripteurs de *hits* tels le MED-SuMo score ou le RMSD des SCFs communs détectés, (2) des descripteurs bioinformatiques tel le domaine PFAM de la chaîne où le fragment se fixe, (3) des atomes dont, il est possible de calculer les propriétés physico-chimiques ou de détecter si ils sont en contact avec les atomes de la protéine requête, (4) et finalement des types d'SCFs communs avec la requête.

Un score MED-SuMo minimal de 3.1 permet l'élimination des hits ayant moins de 4 SCFs.

#### iii. *Génération de molécules réelles avec MED-Hybridise*

##### MED-Hybridise

MED-Hybridise permet la recombinaison 3D des fragments issus des MED-Portions pour former des molécules putatives significatives pour la conception de médicament. Cet outil utilise en entrée la liste de MED-Portions exportée du MED-SuMo GUI dans le repère 3D de la protéine requête. Il fournit des outils standards de chémoinformatique pour analyser à la fois les fragments en entrée et les molécules finales hybridées. Différents types d'algorithmes de combinaison de sous-structures sont disponibles. MED-Hybridise utilise la méthode de combinaison de chaînes (« *Chain Combine* ») qui s'inspire de l'approche BREED [193] originellement implémentée pour la combinaison de ligands. Pour BREED, il faut tout d'abord fournir en entrée un ensemble de ligand préalablement superposé dans le site actif de la requête. Ensuite ce logiciel combine ces ligands pour former d'autres molécules. Le principe de la méthode de combinaison de chaînes repose sur la détection de liaisons qui se superposent entre paires de MED-Portions alignés. Deux liaisons sont semblables si leurs « ordres de liaison » sont les mêmes, si les distances entre les paires d'atomes sont comprises dans une certaine valeur et si les directions des vecteurs des deux liaisons sont comprises dans un certain angle. Par défaut, ces valeurs sont de 1 Å et 15°.

L'utilisation d'une protéine requête pour détecter les similitudes locales et les MED-Portions présentent certains avantages par rapport à BREED. D'une part, les hybrides formés s'ajustent sans heurt intermoléculaire sur la surface d'interaction de la requête. D'autre part,

MED-SuMo permet la détection de *hits* entre familles, dite interfamille, des fragments extraits de ligands co-cristallisés avec des protéines de familles différentes de celle de la requête alors que BREED n'utilise que des alignements de structures de même superfamille). Cette démarche va permettre la génération d'hybrides pouvant cibler certaines protéines sans ligand connu. Ensuite, BREED requiert d'avoir préalablement superposé les ligands dans le site actif, ce qu'un *run* standard MED-SuMo site *versus* la base sites peut d'ailleurs facilement générer. De plus, le MED-sumo-client en « batch » a permis le développement d'une composante MED-SuMo pour le programme de pipeline Scitegic Pipeline Pilot™ 7.0 [144]. MED-SuMo peut être inclus dans un enchaînement de programmes, ce qui permet de totalement automatiser l'enchaînement des trois étapes décrites:

- Lancement d'un *run* MED-SuMo sur la base fragments.
- Exportation de la liste des fragments et leurs coordonnées.
- Analyse et hybridation des fragments pour créer de nouvelles molécules.

La Figure 51 est un exemple d'enchaînement des programmes aboutissant à la génération de nouvelles molécules. En outre, cette nouvelle approche favorise la création de molécules plus facilement synthétisables. En effet, les MED-Portions contiennent des sous-structures chimiques issues de molécules de la base PubChem Compounds qui sont synthétisables. Une hybridation arbitraire de ces sous-structures est donc plus certainement synthétisable qu'une combinaison de sous-structures aléatoires ou encore d'un ensemble d'atomes. Enfin, les atomes *dummy* permettent la génération de liaison entre sous-structures chimiques où des substitutions peuvent réellement avoir lieu. Ces liaisons existent donc aussi plus certainement dans des molécules synthétisables.

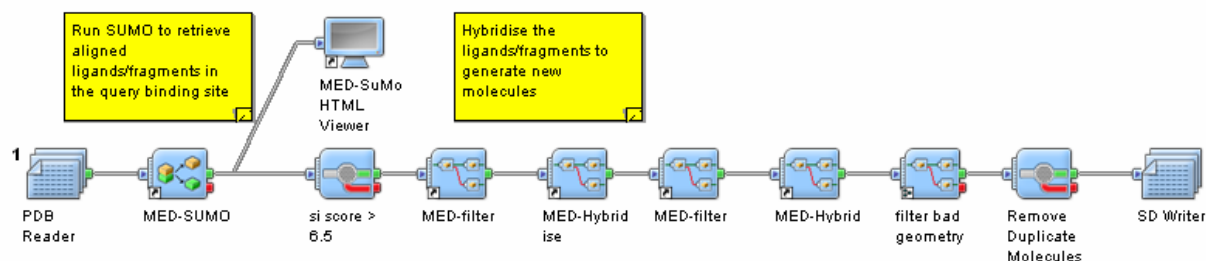
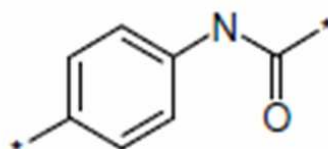


Figure 51 : Exemple d'enchaînement des programmes pour l'approche fragmentale avec le *framework* Scitegic Pipeline Pilot™ 7.0 [144].

### Combinaison de MED-Portions en ligands potentiels

---

Pour chaque cible, la première étape d'hybridation se fait entre un ensemble de sous-structures potentiellement intéressantes (cf. Figure 52) qui sont sélectionnées en fonction des cibles. Chaque étape d'hybridation commence avec tous les MED-Portions et inclut les hybrides des étapes précédentes. Entre chaque étape, différents filtres sont applicables. Dans un premier temps, les hybrides sont filtrés pour garder les molécules possédant une sous-structure 2D ou 3D particulière reconnue expérimentalement comme importante. Par exemple, pour les protéines kinases, la présence d'un groupement phenylamide est favorisée car il se lie à la partie charnière des sites de liaisons des ligands DFG-out. Finalement, les hybrides ayant plus de cinq heurts intermoléculaires avec la requête sont éliminés à la seconde étape.



**Figure 52 : Sous-structure intéressante pour la protéine kinase VEGFR-2 contenant le groupement phenylamide, extrait du ligand GIG de la structure 2OH4**

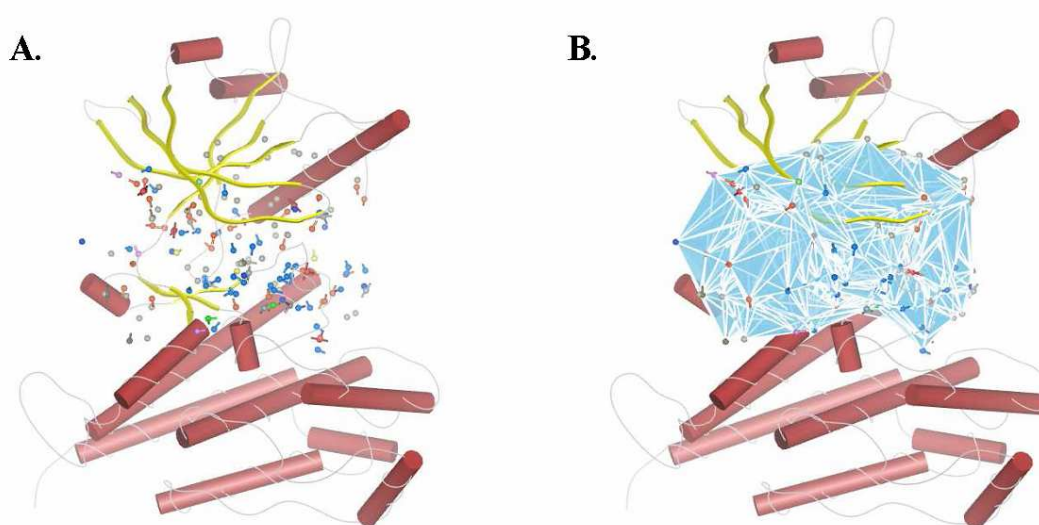
Nous allons maintenant décrire une application du protocole de conception de molécule active *de novo* sur un récepteur kinase VEGFR-2.

### **3. Application de l'approche fragmentale sur le récepteur protéine kinase VEGFR-2**

La superfamille des protéines kinase est une cible intéressante pour l'industrie pharmaceutique. En effet, la fonction centrale de ces protéines concerne la transduction des signaux cellulaires chez tous les organismes et son dysfonctionnement est à l'origine de pathologies graves comme le cancer, le diabète ou des maladies inflammatoires comme l'arthrite. Le site de liaison à l'ATP est localisé à l'interface entre deux lobes formant le repliement spécifique des protéines kinases. En 2001, le premier inhibiteur de protéine kinase s'administrant oralement, l'imatinib [199], a été approuvé contre la leucémie myéloïde chronique, prouvant ainsi qu'une poche à ATP conservée pouvait présenter une affinité particulière et être sélective envers une petite molécule. La structure de la protéine a été résolue et a révélé que la fixation de l'imatinib dans la poche à ATP provoque un

déplacement important de la phénylalanine dans le motif DFG qui développe une nouvelle poche. Alors que dans la conformation active, les deux poches sont connectées par une région charnière composée de résidus ayant un rôle crucial dans la liaison avec les ligands DFG-out. Les modifications conformationnelles induites par l'imatinib rendent la structure inactive.

Dans cette étude, nous avons sélectionné la structure d'une protéine kinase DFG-out particulièrement étudiée, le récepteur VEGFR-2. La requête MED-SuMo soumise est représentée Figure 53)



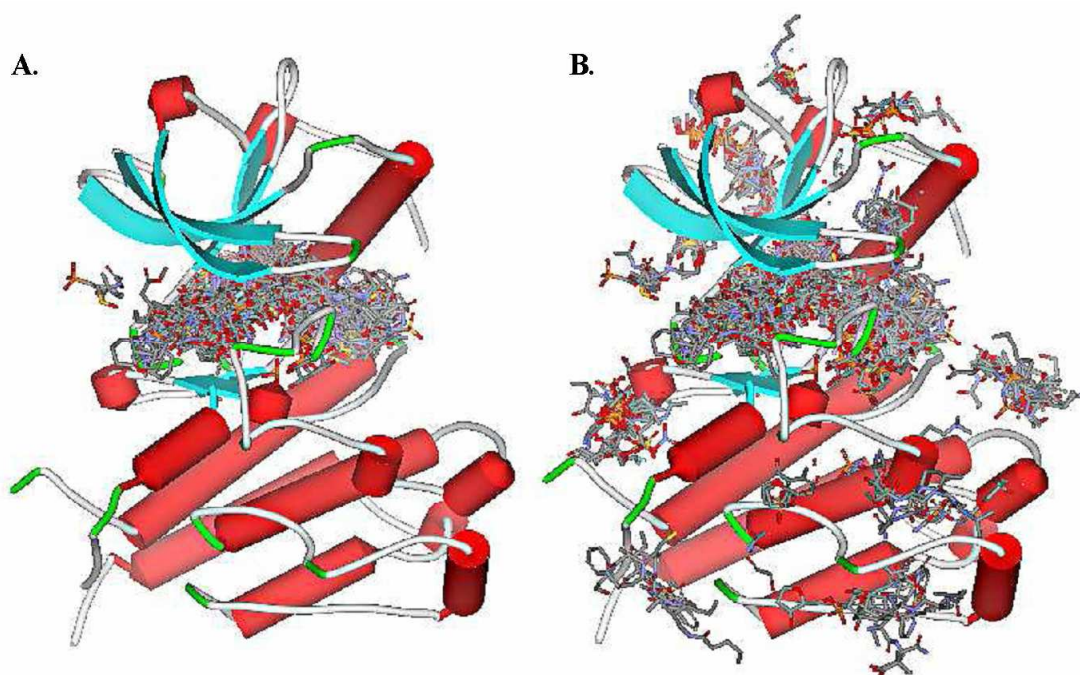
**Figure 53 : Représentation de la requête soumise à la base de MED-Portions.**

Le site de liaison de la structure de la protéine kinase VEGFR-2 (code PDB : 2OH4). **A.** Représentation de la requête contenant les SCFs détectés dans un rayon de 10 Å autour du ligand. **B.** Représentation des triplets de SCFs qui forment le graphe de la requête qui est comparé aux graphes de MED-Portions pré-compilés.

Le protocole décrit précédemment est utilisé pour générer des inhibiteurs DFG-out potentiels. Des hybrides 3D sont donc générés dans la poche DFG-out et seuls ceux contenant la sous-structure phénylamide du ligand GIG ont été gardés. Ce choix est important pour la génération de ligands DFG-out car ce groupement proche de la région charnière est uniquement observé dans les ligands se fixant au DFG-out. Il va donc induire un biais intéressant pour les hybrides générés sachant que nous voulons éviter de générer des ligands se fixant aux structures DFG-in. De plus, la base de MED-Portions contient à la fois les structures DFG-out et DFG-in ce qui est profitable car la région charnière est commune à ces

deux structures. Les MED-Portions extraites de DFG-in peuvent donc contribuer à la génération des hybrides DFG-out.

Après le filtrage des hits, 1474 MED-Portions sont conservés (cf. Figure 54), parmi lesquels 25% proviennent de familles PFAM différentes des protéines kinases. Ce sont des *hits* interfamilles. Ce pourcentage est important et ces *hits* influent sur les résultats de l'hybridation.



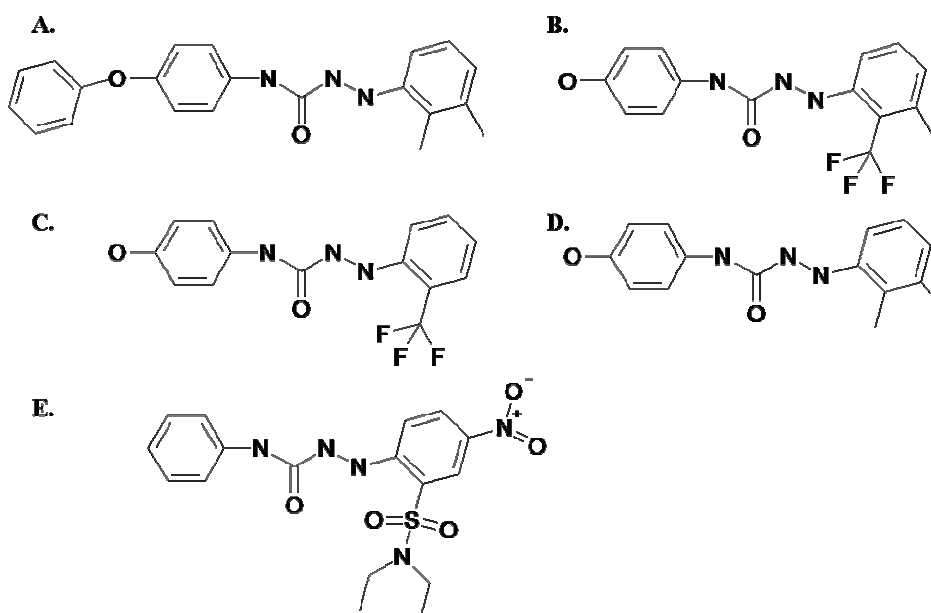
**Figure 54 : MED-Portions détectés par MED-SuMo.**

En A. représentation des 1474 MED-Portions détectés par MED-SuMo. En B. MED-Portions détectés pour une requête comprenant toute la surface de la protéine.

L'analyse des châssis moléculaires (*scaffold*) des hybrides permet d'extraire 9000 molécules uniques dont 175 correspondent aux ligands de la PDB (19 ligands uniques). Le protocole mis en place permet donc la génération de ligands connus présents dans la PDB et plus particulièrement de 8000 nouveaux *scaffolds* non observés dans la PDB. 83 des 175 matches sont des ligands de protéine kinase tels GIG, L09, BMU, L10, 1PP, G2G, BAX, 2RL. La comparaison à une base de molécules (*PubMed Compounds*) détecte 549 matches (294 uniques) qui sont des molécules candidates intéressantes pour inhiber en conformation DFG-out. En parallèle, nous avons aussi recherché les hybrides générés dans une base de données des faux-positifs pour le VEGFR-2, la base DUD [200], seuls 2 *scaffolds* sont retrouvés (ceux des molécules : ZINC00341936 et ZINC00570337). Un faible taux de faux positifs est donc généré par notre approche. De manière intéressante, 27 matches (dont 13

uniques) sont annotés comme actifs sur les protéines kinases, correspondant à 735 ligands des tests d'activité des protéines kinases de la *PubChem*. Ces ligands pourraient donc potentiellement être testés sur les protéines kinases connues pour se lier aux sites DFG-out.

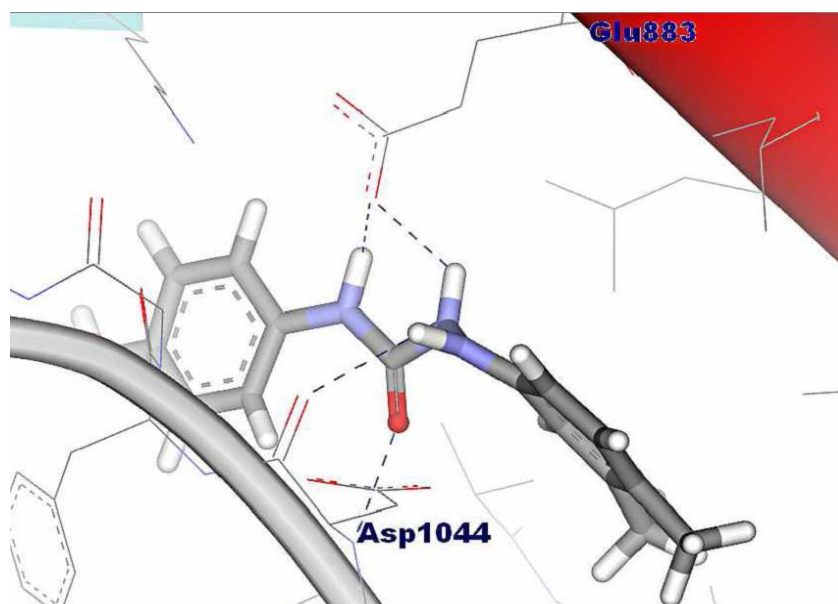
Certains *scaffolds* sont retrouvés dans les tests d'activité des protéines kinases de la *PubChem* et pas dans la PDB. La plupart ont des différences mineures avec les ligands de la PDB tels qu'un O à la place d'un N dans une région liante, ou d'un N à la place d'un C dans un cycle aromatique. Le ligand correspondant au composé *PubChem* ayant l'identifiant CID=3527591 est particulièrement intéressant. Il est annoté comme inhibiteur de la protéine kinase FAK et contient une sous-structure benzohydrazide originale. Nos résultats suggèrent que cette molécule est un ligand DFG-out potentiel (cf. Figure 55 et Figure 56) ce qui nécessiterait naturellement d'être validé expérimentalement. De manière surprenante, une structure récemment déposée dans la PDB (3C1X) est une protéine kinase complexée avec le ligand DFG-out contenant un groupement malonamide [201]. Ce ligand est proche de notre hybride car il est différent de la fraction amide ou urée que nous retrouvons habituellement dans cette position dans les complexes issus de la PDB.



**Figure 55 : Exemple d'hybrides obtenus avec le programme MED-Hybridize.**

A-E : Structure de cinq hybrides contenant des scaffolds de la base « *PubChem Compounds* » comme par exemple la sous-structure ayant l'identifiant CID=3527591 représentée dans E.





**Figure 56 : Molécule D de la Figure 55 représentée selon ses coordonnées 3D dans la requête 2OH4.**

Cette molécule est une hybride de quatre MED-Portions extraits des fichiers PDB : 2HZ0, 2OH4, 1WBN, 3CTQ (toutes étant des structures de kinases DFG-out) combiné avec la sous-structure phénylamide initiale. Une minimisation énergétique sur la géométrie de l'hybride a permis la relaxation du groupement benzohydrazide. Une liaison hydrogène est détectée entre le groupement carbonyle de l'aspartate 1044 et le groupement benzohydrazide.

### Conclusion

Nous avons développé un protocole complet et automatique pour le *fragment-based drug design* permettant de générer des molécules *de novo* basées sur des informations structurales de la PDB. Cette nouvelle méthode est une combinaison des trois techniques MED-Portions, MED-SuMo et MED-Hybridise. Elle se base sur la soumission d'une surface d'interaction requête afin de superposer des motifs protéine-fragment issus de la PDB. Les sous-structures chimiques sont ensuite exportés en gardant leurs coordonnées 3D dans le repère 3D de la protéine requête, et ensuite combinées pour former de nouveaux composés hybrides. Le programme MED-SuMo Fragmentor permet de générer un grand nombre de MED-portions, des complexes protéine-fragment extraits à partir des complexes protéine-ligand de la PDB se basant sur de réelles petites molécules chimiques. Le protocole peut être enrichi avec des données biostructurales propriétaires au format PDB. Toute autre chimiothèque peut aussi servir au protocole de fragmentation. Nous avons choisi la *PubChem* car d'une part elle regroupe un grand nombre de molécules réelles de fournisseurs différents et d'autre part certaines de ces molécules disposent d'annotations biologiques. L'objet MED-Portion est un concept nouveau qui contient une sous-structure chimique, des *dummy* atomes représentant les valences libres sur les fragments et les SCFs de MED-SuMo représentant les groupements chimiques de la surface de la protéine en interaction potentielle avec les atomes



---

du fragment. Combiné avec l'approche de MED-SuMo, ce protocole permet de remplir des surfaces d'interactions de protéines (sites de liaison, surfaces entières ou autres) par un ensemble de MED-Portions. L'application sur la protéine kinase VEGFR-2 permet de récupérer des *scaffolds* de ligands actifs connus. De plus, la structure d'une sous-structure potentiellement intéressante (phénylamide) est étendue par hybridation avec les MED-Portions obtenus jusqu'à la formation d'une molécule candidate. En détectant parmi les molécules hybrides celles directement disponibles dans une chimiothèque donnée, il devient possible d'acheter directement ces molécules candidates. Une application sur les GPCR a aussi été réalisée et a été publié récemment (Article 3, [17]).

## Conclusion générale

---

## IV. Conclusion générale

La fonction d'une protéine est étroitement liée à son repliement tridimensionnel. Cette affirmation qui semblait évidente doit être considérée avec précaution car de plus en plus de contre-exemples ont été mis en évidence. Si la détection des repliements adoptés permet de proposer une classification de la PDB sur des critères thermodynamiques fondamentaux, la ou les fonctions d'une protéine s'expriment au travers des interactions intermoléculaires avec d'autres partenaires biologiques. Les interactions entre macromolécules biologiques et partenaire moléculaire reposent sur un ensemble d'interactions de faible énergie: liaisons hydrogènes, interactions électrostatiques, contacts Van der Waals et effets hydrophobes, qui interviennent entre groupements chimiques spécifiques. La surface de la protéine entre donc en contact avec ses partenaires, le cœur de sa structure lui conditionnant le repliement observé. Étudier cette surface d'interaction permet de comprendre les mécanismes impliqués dans la fonction de la protéine. L'affirmation proposée pourrait donc être remplacée par « la fonction d'une protéine est étroitement liée à la surface d'interaction exposée de sa structure tridimensionnelle », sans pour autant minimiser l'importance des mécanismes de flexibilité/adaptabilité. MED-SuMo repose sur l'assertion selon laquelle des surfaces d'interaction similaires peuvent se lier aux mêmes molécules. Bien qu'il s'inspire d'autres méthodes basées sur l'étude de groupements chimiques à la surface des protéines, son approche complète basée sur la théorie des graphes est innovante et rapide.

Mon travail de thèse s'est divisé en deux parties distinctes. La première a été de participer au développement de méthode dans la plateforme logicielle MED-SuMo dans le cadre du savoir-faire mis en œuvre par MEDIT, en ajoutant certaines fonctionnalités ou en améliorant les communications client-serveur. L'intérêt de MED-SuMo dans différentes applications de modélisation moléculaires basées sur la structure des protéines a été mis en évidence dans une revue [14], notamment pour l'annotation fonctionnelle de protéines hypothétiques. En effet, alors que les approches d'études classiques des protéines ne donnent aucun résultat exploitable pour la protéine TM1012 [13], MED-SuMo propose des fonctions biochimiques en détectant des surfaces d'interaction similaires à d'autres sites de liaison dont la fonction est définie et les mécanismes d'action connus. Pour la protéine YBL036C, annotée comme hypothétique, la situation est différente. PSI-BLAST détecte des similitudes avec des

---

alanines racémases avec un taux d'identité de séquence proche de 40 %. Profunc, avec la méthode de comparaison DALI, détecte également des similitudes avec des structures d'alanines racémases. En ce cas, MED-SuMo confirme de manière précise ce que d'autres approches détectent en localisant quatre alanines racémases dont les sites de liaison se superposent parfaitement avec celui de la requête (cf. Figure 22).

La seconde partie de mon travail a été d'étendre MED-SuMo vers deux autres approches: la classification de surfaces d'interactions avec MED-SMA et la conception *de novo* de molécules actives avec l'approche fragmentale.

MED-SMA intègre d'une part une comparaison 2 à 2 de tous les sites d'interaction d'un jeu de protéines donné, par la technologie MED-SuMo, et d'autre part la classification globale de ces sites. Les deux applications décrites sont réalisées sur des ensembles de sites de liaison connus et pour chacune d'elle, MED-SMA rassemble dans les mêmes groupes des sites connus pour fixer des types équivalents de molécules. Dans la classification des sites puriques, deux regroupements de protéines kinases, issues de familles différentes (selon le travail de Manning et ses collaborateurs [185]) sont particulièrement intéressantes. La première concerne les sites d'une CDK2 et d'une GSK3 $\beta$  sont rassemblés, ce qui est corroboré expérimentalement par une étude des relations structure-activité (SAR) des structures de protéines kinases résolues [184]. Ces protéines ont des activités comparables et sont inhibées par les mêmes types de molécules. La seconde concerne le site d'une CDK2 et celui d'une protéine tyrosine kinase Aurora-A sont aussi rassemblés par la méthode, ces deux sites s'avèrent être même plus proche que celui de la protéine 1B38 et d'autres CDK2 du jeu de données. Alors que ces deux protéines kinases sont aussi issues de familles différentes, une autre étude montre qu'elles sont inhibées par les mêmes classes de molécules. L'application sur les protéines du repliement GHKL montre aussi le même type de regroupement, les protéines se liant à la molécule active, radicicol, sont rassemblées dans le même groupe.

Alors que la taille des sites de liaison dépend fortement de la taille des ligands fixés, les MED-clusters ne sont pas influencés par le type des ligands des sites. En effet, les ligands sont mélangés dans un grand nombre de groupes formés. MED-SMA rassemble des sites de liaisons qui adoptent des modes de liaison proches, et qui par extension fixent des ligands proches. Ce type de relation entre les familles de protéines est très intéressant et leur identification est une application directe de MED-SMA. Si des protéines peu étudiées sont détectées dans un groupe, cette classification peut aussi permettre de déduire des mécanismes

---

enzymatiques ou encore de déterminer la fonction de protéines. MED-SMA pourrait donc permettre d'améliorer l'efficacité et le rendement de la recherche relative à la conception de médicaments, notamment en facilitant le choix des premières pistes à suivre. Cette étude illustre l'utilité de MED-SuMo pour à la fois annoter les structures des protéines mais aussi permettre la classification fonctionnelle et structurale d'un jeu de sites de liaison. Cette nouvelle méthode s'inclut dans un projet en cours de plus grande envergure qui concerne la classification de tous les sites de liaisons de la PDB. Une fois achevée et validée, elle sera mise à disposition de la communauté scientifique au travers d'une interface web actuellement en développement à MEDIT.

La seconde extension de MED-SuMo est le protocole de conception *de novo* de molécules actives par *fragment-based drug design* (FBDD). Cette nouvelle méthode est une combinaison des trois techniques MED-Portions, MED-SuMo et MED-Hybridise. Elle se base sur la comparaison d'une surface d'interaction requête, à une base d'interactions protéine-fragment extraits à partir des complexes protéine-ligand de la PDB, dans le but de superposer des fragments dans le site actif de la requête. Ces sous-structures chimiques sont exportées en conservant leurs positions et combinées 3D pour former de nouveaux composés hybrides. Pour cette nouvelle approche, mon rôle principal avec le savoir-faire mis en œuvre par MEDIT, a été l'adaptation de MED-SuMo serveur pour (1) qu'il communique avec le programme MED-Fragmentor, et (2) qu'il puisse gérer le nouveau type de base MED-SuMo de surfaces d'interactions définie par les MED-Portions. L'efficacité de la méthode est illustré par l'application sur la protéine kinase VEGFR-2. Elle permet de récupérer des *scaffolds* de ligands actifs connus. Une extension de la méthode qui inclura les molécules d'eau des fichiers PDB a été initiée.

MED-SuMo est un logiciel en permanente évolution, sa capacité à détecter des surfaces d'interaction similaires est puissante et ouvre de multiples nouvelles applications car applicable à tous les types de surfaces d'interaction. Mon activité future au sein de MEDIT sera de gérer et d'optimiser les protocoles de MED-SuMo existants et de l'adapter pour d'autres applications. Une extension de la méthode pour détecter et comparer les interfaces protéine-protéine a déjà pu être initiée et des premiers tests satisfaisants ont été réalisés. Cependant, la taille de ces interfaces étant supérieure à celle des sites de liaison classique, des modifications techniques seront donc nécessaires pour optimiser les résultats obtenus.

---

L'interface graphique est aussi en cours d'adaptation afin de profiter pleinement d'une visualisation en 3D des similitudes détectées. Ce projet est un nouveau challenge captivant.

Enfin, d'autres évolutions sont prévues comme la prise en compte des facteurs B des structures des protéines dans le calcul du score MED-SuMo. Un nouveau projet d'annotations des structures est en outre envisagé. Pour l'heure, l'index PDB permet d'importer toute nouvelle annotation dans les bases de données MED-SuMo. Les annotations comme PFAM ou PROSITE ne concernent souvent qu'une petite partie de la protéine (une chaîne, ou quelques résidus). La possibilité de faire le lien entre les similitudes détectées par MED-SuMo et des annotations précises (de l'ordre du résidu) permettrait d'exploiter pleinement les résultats détectés.

## BIBLIOGRAPHIE

## BIBLIOGRAPHIE

- [1] Liolios, K., Mavromatis, K., Tavernarakis, N. and Kyrpides, N.C. (2008). The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res* 36, D475-9.
- [2] Bairoch, A. and Apweiler, R. (1997). The SWISS-PROT protein sequence database: its relevance to human molecular medical research. *J Mol Med* 75, 312-6.
- [3] Godzik, A., Jaume Canaves, Slawomir Grzechnik, Lukasz Jaroszewski, and Andrew Morse, J.O., Xianhong Wang, Bill West and John Wooley. (2003). Challenges of structural genomics: Expectations and Outcomes. *Biosilico* 1, 36-41.
- [4] Matthews, B.W. (2007). Protein Structure Initiative: getting into gear. *Nature Structural & Molecular Biology* 14, 459-460.
- [5] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000). The Protein Data Bank. *Nucleic Acids Res* 28, 235-42.
- [6] Hassell, A.M., An, G., Bledsoe, R. K., Bynum, J. M., Carter, H. L., 3rd, Deng, S. J., Gampe, R. T., Grisard, T. E., Madauss, K. P., Nolte, R. T., Rocque, W. J., Wang, L., Weaver, K. L., Williams, S. P., Wisely, G. B., Xu, R. and Shewchuk, L. M. (2007). Crystallization of protein-ligand complexes. *Acta Crystallogr D Biol Crystallogr* 63, 72-9.
- [7] Deléage, G., Geourjon, C. and Jambon, M. (2002). Process for identifying similar 3D substructures onto 3D atomic structures and its applications. 02291407.1. CNRS.
- [8] Jambon, M. (2003). Un système bioinformatique de recherche de similitudes fonctionnelles dans les structures 3D de protéines. June, 20th, 2003, Institut de Biologie et Chimie des Protéines (IBCP). Université Claude Bernard, Lyon I
- [9] Jambon, M., Imberty, A., Deleage, G. and Geourjon, C. (2003). A new bioinformatic approach to detect common 3D sites in protein structures. *Proteins* 52, 137-45.
- [10] Jambon, M., Andrieu, O., Combet, C., Deleage, G., Delfaud, F. and Geourjon, C. (2005). The SuMo server: 3D search for protein functional sites. *Bioinformatics* 21, 3929-30.



- [11] Rosen, M., Lin, S.L., Wolfson, H. and Nussinov, R. (1998). Molecular shape comparisons in searches for active sites and functional similarity. *Protein Eng* 11, 263-77.
- [12] Schmitt, S., Kuhn, D. and Klebe, G. (2002). A new method to detect related function among proteins independent of sequence and fold homology. *J Mol Biol* 323, 387-406.
- [13] Doppelt, O., Moriaud, F., Bornot, A. and de Brevern, A.G. (2007). Functional annotation strategy for protein structures. *Bioinformatics* 1, 357-9.
- [14] Doppelt-Azeroual, O., Moriaud, F., Delfaud, F. (2009). MED-SuMo Applications. *Infectious Disorders-Drug Targets*
- [15] Doppelt-Azeroual, O., Moriaud, F., Delfaud, F., de Brevern, A.G. (2009). Analysis of HSP90 related folds with MED-SuMo classification approach. *Drug Design Development and Therapy*
- [16] Doppelt-Azeroual, O., Koch, K., Delfaud, F., Moriaud, F. de Brevern, A.G. (2009). Fast and Automated Functional Classification with MED-SuMo: An Application on the Purinome.
- [17] Moriaud, F., Doppelt-Azeroual, O., Martin, L., Oguievetskaia, K., Koch, K., Vorotyntsev, A., Adcock, S.A. and Delfaud, F. (2009). Computational Fragment-Based Approach at PDB Scale by Protein Local Similarity. *J Chem Inf Model* 49, 280-294.
- [18] Oguievetskaia, K., Martin-Chanas L., Vorotyntsev, A., Doppelt-Azeroual, O., Brotel,X., Adcock,S.A., de Brevern,A.G., Delfaud,F. and Moriaud, F. . (2009 (submitted)). Mitotic Kinesin Eg5 Inhibitors Generation By Computational MED-Portion Based Drug Design At PDB Scale.
- [19] Consortium, I.H.G.S. (2004). Finishing the euchromatic sequence of the human genome. *Nature* 431, 931-45.
- [20] Jones, D.T. (2000). Protein structure prediction in the postgenomic era.
- [21] Altschul, S.F. and Koonin, E.V. (1998). Iterated profile searches with PSI-BLAST--a tool for discovery in protein databases. *Trends Biochem Sci* 23, 444-7.
- [22] Eddy, S.R. (1996). Hidden Markov models. *Curr Opin Struct Biol* 6, 361-5.

- [23] Fischer, D. and Eisenberg, D. (1999). Finding families for genomic ORFans. *Bioinformatics* 15, 759-62.
- [24] Congreve, M., Murray, C.W. and Blundell, T.L. (2005). Structural biology and drug discovery. *Drug Discov Today* 10, 895-907.
- [25] (2000) Crystallography Made Crystal Clear
- [26] Ochsenbein, F., Gilquin, Bernard. (2007). La RMN pour comprendre les protéines. *Clefs CEA* 56
- [27] Rieping, W., Habeck, M., Bardiaux, B., Bernard, A., Malliavin, T.E. and Nilges, M. (2007). ARIA2: automated NOE assignment and data integration in NMR structure calculation. *Bioinformatics* 23, 381-2.
- [28] Agez, M., Chen, J., Guerois, R., van Heijenoort, C., Thuret, J.Y., Mann, C. and Ochsenbein, F. (2007). Structure of the histone chaperone ASF1 bound to the histone H3 C-terminal helix and functional insights. *Structure* 15, 191-9.
- [29] Lacapere, J.J., Pebay-Peyroula, E., Neumann, J.M. and Etchebest, C. (2007). Determining membrane protein structures: still a challenge! *Trends Biochem Sci* 32, 259-70.
- [30] Rouiller, I., Xu, X. P., Amann, K. J., Egile, C., Nickell, S., Nicastro, D., Li, R., Pollard, T. D., Volkman, N. and Hanein, D. (2008). The structural basis of actin filament branching by the Arp2/3 complex. *J Cell Biol* 180, 887-95.
- [31] Chandonia, J.M. and Brenner, S.E. (2006). The impact of structural genomics: expectations and outcomes. *Science* 311, 347-51.
- [32] Weigelt, J., McBroom-Cerajewski, L.D., Schapira, M., Zhao, Y. and Arrowsmith, C.H. (2008). Structural genomics and drug discovery: all in the family. *Curr Opin Chem Biol* 12, 32-9.
- [33] Bernstein, F.C. et al. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *J Mol Biol* 112, 535-42.
- [34] Hooft, R.W., Vriend, G., Sander, C. and Abola, E.E. (1996). Errors in protein structures. *Nature* 381, 272.
- [35] Bourne, P.E., and Weissig, H. (2003) *Structural Bioinformatics*. San Diego, CA, USA.
- [36] Ramachandran, G.N. and Sasisekharan, V. (1968). Conformation of polypeptides and proteins. *Adv Protein Chem* 23, 283-438.

- [37] Kleywegt, G.J. and Jones, T.A. (1997). Model building and refinement practice. *Methods Enzymol* 277, 208-30.
- [38] Brunger, A.T. (1992). Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature* 355, 472-5.
- [39] Yuan, Z., Bailey, T.L. and Teasdale, R.D. (2005). Prediction of protein B-factor profiles. *Proteins* 58, 905-12.
- [40] Bartlett, G.J., Porter, C.T., Borkakoti, N. and Thornton, J.M. (2002). Analysis of catalytic residues in enzyme active sites. *J Mol Biol* 324, 105-21.
- [41] Bornot A., O.B.d.B.A.G. (2007). How flexible protein structures are? New questions on the protein structure plasticity. *BIOFORUM Europe* 11, 24-25.
- [42] Shapovalov, M.V. and Dunbrack, R.L., Jr. (2007). Statistical and conformational analysis of the electron density of protein side chains. *Proteins* 66, 279-303.
- [43] DeLano, W.L.T. (2002) The PyMOL Molecular Graphics System DeLano Scientific, <http://www.pymol.org>, San Carlos, CA, USA.
- [44] Murzin, A.G., Brenner, S.E., Hubbard, T. and Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 247, 536-40.
- [45] Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B. and Thornton, J.M. (1997). CATH--a hierarchic classification of protein domain structures. *Structure* 5, 1093-108.
- [46] Holm, L. and Sander, C. (1996). Mapping the protein universe. *Science* 273, 595-603.
- [47] Lathrop, R.H. (1994). The protein threading problem with sequence amino acid interaction preferences is NP-complete. *Protein Eng* 7, 1059-68.
- [48] Escalier, V., Pothier, J., Soldano, H. and Viari, A. (1998). Pairwise and multiple identification of three-dimensional common substructures in proteins. *J Comput Biol* 5, 41-56.
- [49] Holm, L. and Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *J Mol Biol* 233, 123-38.
- [50] Shindyalov, I.N. and Bourne, P.E. (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng* 11, 739-47.

- [51] Holm, L. and Sander, C. (1996). The FSSP database: fold classification based on structure-structure alignment of proteins. *Nucleic Acids Res* 24, 206-9.
- [52] Needleman, S.B. and Wunsch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* 48, 443-53.
- [53] Taylor, W.R. and Orengo, C.A. (1989). Protein structure alignment. *J Mol Biol* 208, 1-22.
- [54] Levine, M., Stuart, D., and Williams, J. (1984). A method for the systematic comparison of the three-dimensional structures of proteins and some results. *Acta Crystallographica Section A* 40, 600-610.
- [55] Usha, R. and Murthy, M.R. (1986). Protein structural homology: a metric approach. *Int J Pept Protein Res* 28, 364-9.
- [56] Carpentier, M., Brouillet, S. and Pothier, J. (2005). YAKUSA: a fast structural database scanning method. *Proteins* 61, 137-51.
- [57] Kabsch, W. and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22, 2577-637.
- [58] Krissinel, E. and Henrick, K. (2004). Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr D Biol Crystallogr* 60, 2256-68.
- [59] Hutchinson, E.G. and Thornton, J.M. (1996). PROMOTIF--a program to identify and analyze structural motifs in proteins. *Protein Sci* 5, 212-20.
- [60] Gibrat, J.F., Madej, T. and Bryant, S.H. (1996). Surprising similarities in structure comparison. *Curr Opin Struct Biol* 6, 377-85.
- [61] Harrison, A., Pearl, F., Mott, R., Thornton, J. and Orengo, C. (2002). Quantifying the similarities within fold space. *J Mol Biol* 323, 909-26.
- [62] Singh, A.P. and Brutlag, D.L. (1997). Hierarchical protein structure superposition using both secondary structure and atomic representations. *Proc Int Conf Intell Syst Mol Biol* 5, 284-93.
- [63] Shapiro, J. and Brutlag, D. (2004). FoldMiner and LOCK 2: protein structure comparison and motif discovery on the web. *Nucleic Acids Res* 32, W536-41.
- [64] Shapiro, J. and Brutlag, D. (2004). FoldMiner: structural motif discovery using an improved superposition algorithm. *Protein Sci* 13, 278-94.

- [65] Lu, G. (2000). TOP: a new method for protein structure comparisons and similarity searches. *J. Appl. Cryst.* 33, 176-183.
- [66] Vriend, G. (1990). WHAT IF: a molecular modeling and drug design program. *J Mol Graph* 8, 52-6, 29.
- [67] Shatsky, M., Nussinov, R. and Wolfson, H.J. (2002). Flexible protein alignment and hinge detection. *Proteins* 48, 242-56.
- [68] Shatsky, M., Nussinov, R. and Wolfson, H.J. (2004). FlexProt: alignment of flexible protein structures without a predefinition of hinge regions. *J Comput Biol* 11, 83-106.
- [69] Pearson, W.R. and Lipman, D.J. (1988). Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A* 85, 2444-8.
- [70] Ye, Y. and Godzik, A. (2003). Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics* 19 Suppl 2, ii246-55.
- [71] Offmann, B., Tyagi, M., de Brevern, A.G. . (2007). Local Protein structures. *Current Bioinformatics* 2, 165-202.
- [72] Tyagi, M., de Brevern, A.G., Srinivasan, N. and Offmann, B. (2008). Protein structure mining using a structural alphabet. *Proteins* 71, 920-37.
- [73] Yang, J.M. and Tung, C.H. (2006). Protein structure database search and evolutionary classification. *Nucleic Acids Res* 34, 3646-59.
- [74] de Brevern, A.G. (2005). New assessment of a structural alphabet. *In Silico Biol* 5, 283-9.
- [75] de Brevern, A.G., Etchebest, C. and Hazout, S. (2000). Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins* 41, 271-87.
- [76] Etchebest, C., Benros, C., Hazout, S. and de Brevern, A.G. (2005). A structural alphabet for local protein structures: improved prediction methods. *Proteins* 59, 810-27.
- [77] Levitt, M. and Chothia, C. (1976). Structural patterns in globular proteins. *Nature* 261, 552-8.
- [78] Dutta, R. and Inouye, M. (2000). GHKL, an emergent ATPase/kinase superfamily. *Trends Biochem Sci* 25, 24-8.

- [79] Cuff, A.L., Sillitoe, I., Lewis, T., Redfern, O.C., Garratt, R., Thornton, J. and Orengo, C.A. (2008). The CATH classification revisited--architectures reviewed and new ways to characterize structural divergence in superfamilies. *Nucleic Acids Res*
- [80] Michie, A.D., Orengo, C.A. and Thornton, J.M. (1996). Analysis of domain structural class using an automated class assignment protocol. *J Mol Biol* 262, 168-85.
- [81] Greene, L.H. et al. (2007). The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucleic Acids Res* 35, D291-7.
- [82] Redfern, O.C., Harrison, A., Dallman, T., Pearl, F.M. and Orengo, C.A. (2007). CATHEDRAL: a fast and effective algorithm to predict folds and domain boundaries from multidomain protein structures. *PLoS Comput Biol* 3, e232.
- [83] Whisstock, J.C. and Lesk, A.M. (2003). Prediction of protein function from protein sequence and structure. *Q Rev Biophys* 36, 307-40.
- [84] Redfern, O.C., Dessailly, B. and Orengo, C.A. (2008). Exploring the structure and function paradigm. *Curr Opin Struct Biol* 18, 394-402.
- [85] Wierenga, R.K. (2001). The TIM-barrel fold: a versatile framework for efficient enzymes. *FEBS Lett* 492, 193-8.
- [86] Hegyi, H. and Gerstein, M. (1999). The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J Mol Biol* 288, 147-64.
- [87] Chothia, C. (1974). Hydrophobic bonding and accessible surface area in proteins. *Nature* 248, 338-9.
- [88] Chothia, C. and Janin, J. (1975). Principles of protein-protein recognition. *Nature* 256, 705-8.
- [89] Lo Conte, L., Chothia, C. and Janin, J. (1999). The atomic structure of protein-protein recognition sites. *J Mol Biol* 285, 2177-98.
- [90] Jones, S. and Thornton, J.M. (1996). Principles of protein-protein interactions. *Proc Natl Acad Sci U S A* 93, 13-20.
- [91] Neuvirth, H., Raz, R. and Schreiber, G. (2004). ProMate: a structure based prediction program to identify the location of protein-protein binding sites. *J Mol Biol* 338, 181-99.

- [92] Janin, J. and Chothia, C. (1990). The structure of protein-protein recognition sites. *J Biol Chem* 265, 16027-30.
- [93] Xu, D., Tsai, C.J. and Nussinov, R. (1997). Hydrogen bonds and salt bridges across protein-protein interfaces. *Protein Eng* 10, 999-1012.
- [94] Bhat, T.N. et al. (1994). Bound water molecules and conformational stabilization help mediate an antigen-antibody association. *Proc Natl Acad Sci U S A* 91, 1089-93.
- [95] Rodier, F., Bahadur, R.P., Chakrabarti, P. and Janin, J. (2005). Hydration of protein-protein interfaces. *Proteins* 60, 36-45.
- [96] Chakrabarti, P. and Janin, J. (2002). Dissecting protein-protein recognition sites. *Proteins* 47, 334-43.
- [97] Bogan, A.A. and Thorn, K.S. (1998). Anatomy of hot spots in protein interfaces. *J Mol Biol* 280, 1-9.
- [98] Janin, J., Bahadur, R.P. and Chakrabarti, P. (2008). Protein-protein interaction and quaternary structure. *Q Rev Biophys* 41, 133-80.
- [99] Janin, J. (1999). Wet and dry interfaces: the role of solvent in protein-protein and protein-DNA recognition. *Structure* 7, R277-9.
- [100] Nadassy, K., Wodak, S.J. and Janin, J. (1999). Structural features of protein-nucleic acid recognition sites. *Biochemistry* 38, 1999-2017.
- [101] Schwabe, J.W. (1997). The role of water in protein-DNA interactions. *Curr Opin Struct Biol* 7, 126-34.
- [102] Lavery, R., and Zakrzewska, K. (1997) Base and base pair morphologies, helical parameters, and definitions. In *Oxford University Press* (Neidle, E.)
- [103] Jones, S., van Heyningen, P., Berman, H.M. and Thornton, J.M. (1999). Protein-DNA interactions: A structural analysis. *J Mol Biol* 287, 877-96.
- [104] Nikolov, D.B., Chen, H., Halay, E.D., Hoffman, A., Roeder, R.G. and Burley, S.K. (1996). Crystal structure of a human TATA box-binding protein/TATA element complex. *Proc Natl Acad Sci U S A* 93, 4862-7.
- [105] Stawiski, E.W., Gregoret, L.M. and Mandel-Gutfreund, Y. (2003). Annotating nucleic acid-binding function based on protein structure. *J Mol Biol* 326, 1065-79.

- [106] Ahmad, S., Gromiha, M.M. and Sarai, A. (2004). Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. *Bioinformatics* 20, 477-86.
- [107] Shanahan, H.P., Garcia, M.A., Jones, S. and Thornton, J.M. (2004). Identifying DNA-binding proteins using structural motifs and the electrostatic potential. *Nucleic Acids Res* 32, 4732-41.
- [108] Guenot, J., Fletterick, R.J. and Kollman, P.A. (1994). A negative electrostatic determinant mediates the association between the *Escherichia coli* trp repressor and its operator DNA. *Protein Sci* 3, 1276-85.
- [109] Laskowski, R.A., Luscombe, N.M., Swindells, M.B. and Thornton, J.M. (1996). Protein clefts in molecular recognition and function. *Protein Sci* 5, 2438-52.
- [110] Goodford, P.J. (1985). A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J Med Chem* 28, 849-57.
- [111] Lichtarge, O., Bourne, H.R. and Cohen, F.E. (1996). An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol* 257, 342-58.
- [112] Bork, P., Jensen, L.J., von Mering, C., Ramani, A.K., Lee, I. and Marcotte, E.M. (2004). Protein interaction networks from yeast to human. *Curr Opin Struct Biol* 14, 292-9.
- [113] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25, 3389-402.
- [114] Pearson, W.R. (1990). Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol* 183, 63-98.
- [115] Luthy, R., Xenarios, I. and Bucher, P. (1994). Improving the sensitivity of the sequence profile method. *Protein Sci* 3, 139-46.
- [116] Bairoch, A. (1991). PROSITE: a dictionary of sites and patterns in proteins. *Nucleic Acids Res* 19 Suppl, 2241-5.
- [117] Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Howe, K.L. and Sonnhammer, E.L. (2000). The Pfam protein families database. *Nucleic Acids Res* 28, 263-6.



- [118] Porter, C.T., Bartlett, G.J. and Thornton, J.M. (2004). The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res* 32, D129-33.
- [119] Gribskov, M., Luthy, R. and Eisenberg, D. (1990). Profile analysis. *Methods Enzymol* 183, 146-59.
- [120] Sigrist, C.J., De Castro, E., Langendijk-Genevaux, P.S., Le Saux, V., Bairoch, A. and Hulo, N. (2005). ProRule: a new database containing functional and structural information on PROSITE profiles. *Bioinformatics* 21, 4060-6.
- [121] de Castro, E., Sigrist, C.J., Gattiker, A., Bulliard, V., Langendijk-Genevaux, P.S., Gasteiger, E., Bairoch, A. and Hulo, N. (2006). ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Res* 34, W362-5.
- [122] Bairoch, A. (2000). The ENZYME database in 2000. *Nucleic Acids Res* 28, 304-5.
- [123] Sottriffer, C. and Klebe, G. (2002). Identification and mapping of small-molecule binding sites in proteins: computational tools for structure-based drug design. *Farmaco* 57, 243-51.
- [124] Hendlich, M., Rippmann, F. and Barnickel, G. (1997). LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *J Mol Graph Model* 15, 359-63, 389.
- [125] Brady, G.P., Jr. and Stouten, P.F. (2000). Fast prediction and visualization of protein binding pockets with PASS. *J Comput Aided Mol Des* 14, 383-401.
- [126] An, J., Totrov, M. and Abagyan, R. (2005). Pocketome via comprehensive identification and classification of ligand binding envelopes. *Mol Cell Proteomics* 4, 752-61.
- [127] Huang, B. and Schroeder, M. (2006). LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Struct Biol* 6, 19.
- [128] Momany, F.A., McGuire, R.F., Burgess, A.W., Scheraga, H.A. (1975). *J. Phys. Chem* 79, 2361-2381.
- [129] Connolly, M.L. (1983). Solvent-accessible surfaces of proteins and nucleic acids. *Science* 221, 709-13.

- [130] Beadle, B.M. and Shoichet, B.K. (2002). Structural bases of stability-function tradeoffs in enzymes. *J Mol Biol* 321, 285-96.
- [131] Brylinski, M., Prymula, K., Jurkowski, W., Kochanczyk, M., Stawowczyk, E., Konieczny, L. and Roterman, I. (2007). Prediction of functional sites based on the fuzzy oil drop model. *PLoS Comput Biol* 3, e94.
- [132] Brylinski, M., Konieczny, L., Czerwonko, P., Jurkowski, W. and Roterman, I. (2005). Early-stage folding in proteins (in silico) sequence-to-structure relation. *J Biomed Biotechnol* 2005, 65-79.
- [133] Laskowski, R.A., Watson, J.D. and Thornton, J.M. (2005). ProFunc: a server for predicting protein function from 3D structure. *Nucleic Acids Res* 33, W89-93.
- [134] Baroni, M., Cruciani, G., Sciabola, S., Perruccio, F. and Mason, J.S. (2007). A common reference framework for analyzing/comparing proteins and ligands. Fingerprints for Ligands and Proteins (FLAP): theory and application. *J Chem Inf Model* 47, 279-94.
- [135] Kinoshita, K. and Nakamura, H. (2003). Identification of protein biochemical functions by similarity search using the molecular surface database eF-site. *Protein Sci* 12, 1589-95.
- [136] Shulman-Peleg, A., Nussinov, R. and Wolfson, H.J. (2004). Recognition of functional sites in protein structures. *J Mol Biol* 339, 607-33.
- [137] Mintz, S., Shulman-Peleg, A., Wolfson, H.J. and Nussinov, R. (2005). Generation and analysis of a protein-protein interface data set with similar chemical and spatial patterns of interactions. *Proteins* 61, 6-20.
- [138] Shulman-Peleg, A., Nussinov, R. and Wolfson, H.J. (2005). SiteEngines: recognition and comparison of binding sites and protein-protein interfaces. *Nucleic Acids Res* 33, W337-41.
- [139] Hendlich, M., Bergner, A., Gunther, J. and Klebe, G. (2003). Relibase: design and development of a database for comprehensive analysis of protein-ligand interactions. *J Mol Biol* 326, 607-20.
- [140] Powers, R., Copeland, J.C., Germer, K., Mercier, K.A., Ramanathan, V. and Revesz, P. (2006). Comparison of protein active site structures for functional annotation of proteins and drug design. *Proteins* 65, 124-35.

- [141] OCaml, <http://caml.inria.fr/>.
- [142] Andrieu, O., <http://oandrieu.nerim.net/ocaml/mlsqlite/>.
- [143] OpenBabel, <http://openbabel.sourceforge.net/>.
- [144] Scitegic Pipeline Pilot™ 7.0; Accelrys Software Inc. 10188 Telesis Court, S.S.D., CA 92121, USA.
- [145] <http://www.pops-systematic.org/>
- [146] <http://www.systematic-paris-region.org>
- [147] Banner, D.W. and Hadvary, P. (1991). Crystallographic analysis at 3.0-Å resolution of the binding to human thrombin of four active site-directed inhibitors. *J Biol Chem* 266, 20085-93.
- [148] Natchus, M.G. et al. (2001). Development of new carboxylic acid-based MMP inhibitors derived from functionalized propargylglycines. *J Med Chem* 44, 1060-71.
- [149] Lua-ML, <http://caml.inria.fr/cgi-bin/hump.en.cgi?contrib=321>.
- [150] Loewenstein, Y. et al. (2009). Protein function annotation by homology-based inference. *Genome Biol* 10, 207.
- [151] Kuhn, D., Weskamp, N., Schmitt, S., Hullermeier, E. and Klebe, G. (2006). From the similarity analysis of protein cavities to the functional classification of protein families using cavbase. *J Mol Biol* 359, 1023-44.
- [152] Kuhn, D., Weskamp, N., Hullermeier, E. and Klebe, G. (2007). Functional classification of protein kinase binding sites using Cavbase. *ChemMedChem* 2, 1432-47.
- [153] Van Dongen, S. (2000). Graph Clustering by Flow Simulation  
University of Utrecht
- [154] Enright, A.J. and Ouzounis, C.A. (2001). BioLayout--an automatic graph layout algorithm for similarity visualization. *Bioinformatics* 17, 853-4.
- [155] Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal* 27, 379-423.
- [156] Hazout, S. (2007) Entropy-derived measures for assessing the accuracy of N-state prediction algorithms. In *Recent Advances in Structural Bioinformatics*. (de Brevern, A.G.), pp. 395-417. Research signpost, Trivandrum, India.

- [157] [http://en.wikipedia.org/wiki/Specificity\\_\(tests\)](http://en.wikipedia.org/wiki/Specificity_(tests))
- [158] Picard, D. (2002). Heat-shock protein 90, a chaperone for folding and regulation. *Cell Mol Life Sci* 59, 1640-8.
- [159] Whitesell, L. and Lindquist, S.L. (2005). HSP90 and the chaperoning of cancer. *Nat Rev Cancer* 5, 761-72.
- [160] Goetz, M.P., Toft, D.O., Ames, M.M. and Erlichman, C. (2003). The Hsp90 chaperone complex as a novel target for cancer therapy. *Ann Oncol* 14, 1169-76.
- [161] Zhang, T., Hamza, A., Cao, X., Wang, B., Yu, S., Zhan, C.G. and Sun, D. (2008). A novel Hsp90 inhibitor to disrupt Hsp90/Cdc37 complex against pancreatic cancer cells. *Mol Cancer Ther* 7, 162-70.
- [162] Roe, S.M., Prodromou, C., O'Brien, R., Ladbury, J.E., Piper, P.W. and Pearl, L.H. (1999). Structural basis for inhibition of the Hsp90 molecular chaperone by the antitumor antibiotics radicicol and geldanamycin. *J Med Chem* 42, 260-6.
- [163] Prodromou, C., Roe, S.M., O'Brien, R., Ladbury, J.E., Piper, P.W. and Pearl, L.H. (1997). Identification and structural characterization of the ATP/ADP-binding site in the Hsp90 molecular chaperone. *Cell* 90, 65-75.
- [164] Guarnieri, M.T., Zhang, L., Shen, J. and Zhao, R. (2008). The Hsp90 inhibitor radicicol interacts with the ATP-binding pocket of bacterial sensor kinase PhoQ. *J Mol Biol* 379, 82-93.
- [165] Corbett, K.D. and Berger, J.M. (2006). Structural basis for topoisomerase VI inhibition by the anti-Hsp90 drug radicicol. *Nucleic Acids Res* 34, 4269-77.
- [166] Bellon, S. et al. (2004). Crystal structures of Escherichia coli topoisomerase IV ParE subunit (24 and 43 kilodaltons): a single residue dictates differences in novobiocin potency against topoisomerase IV and DNA gyrase. *Antimicrob Agents Chemother* 48, 1856-64.
- [167] Haystead, T.A. (2006). The purinome, a complex mix of drug and toxicity targets. *Curr Top Med Chem* 6, 1117-27.
- [168] Kabsch, W., Mannherz, H.G., Suck, D., Pai, E.F. and Holmes, K.C. (1990). Atomic structure of the actin:DNase I complex. *Nature* 347, 37-44.

- [169] Westover, K.D., Bushnell, D.A. and Kornberg, R.D. (2004). Structural basis of transcription: nucleotide selection by rotation in the RNA polymerase II active center. *Cell* 119, 481-9.
- [170] Gatzeva-Topalova, P.Z., May, A.P., Sousa, M.C. . (2005). Structure and Mechanism of ArnA: Conformational Change Implies Ordered Dehydrogenase Mechanism in Key Enzyme for Polymyxin Resistance *Structure* 13, 929-942.
- [171] Allard, S.T. et al. (2002). Toward a structural understanding of the dehydratase mechanism. *Structure* 10, 81-92.
- [172] Webb, N.A., Mulichak, A.M., Lam, J.S., Rocchetta, H.L., Garavito, R.M. . (2004). Crystal structure of a tetrameric GDP-D-mannose 4,6-dehydratase from a bacterial GDP-D-rhamnose biosynthetic pathway. *Protein Sci* 13
- [173] Hung, L.W., Wang, I.X., Nikaido, K., Liu, P.Q., Ames, G.F., Kim, S.H. (1998). Crystal structure of the ATP-binding subunit of an ABC transporter. *Nature* 396, 703-707.
- [174] Chen, J., Lu, G., Lin, J., Davidson, A.L., Quioco, F.A. (2003). A tweezers-like motion of the ATP-binding cassette dimer in an ABC transport cycle. *Mol. Cell* 12, 651-661
- [175] Lewis, H.A., Buchanan, S.G., Burley, S.K., Connors, K., Dickey, M., Dorwart, M., Fowler, R., Gao, X., Guggino, W.B., Hendrickson, W.A., Hunt, J.F., Kearins, M.C., Lorimer, D., Maloney, P.C., Post, K.W., Rajashankar, K.R., Rutter, M.E., Sauder, J.M., Shriver, S., Thibodeau, P.H., Thomas, P.J., Zhang, M., Zhao, X., Emtage, S. . (2004). Structure of nucleotide-binding domain 1 of the cystic fibrosis transmembrane conductance regulator. *Embo J.* 23, 282-293
- [176] Ramaen, O., Leulliot, N., Sizun, C., Ulryck, N., Pamlard, O., Lallemand, J.Y., Tilbeurgh, H., Jacquet, E. (2006). Structure of the human multidrug resistance protein 1 nucleotide binding domain 1 bound to Mg<sup>2+</sup>/ATP reveals a non-productive catalytic site. *J.Mol.Biol.* 359, 940-949
- [177] Zaitseva, J., Oswald, C., Jumpertz, T., Jenewein, S., Wiedenmann, A., Holland, I.B., Schmitt, L. (2006 ). A structural analysis of asymmetry required for catalytic activity of an ABC-ATPase domain dimer. *Embo J.* 25, 3432-3443.
- [178] Gaudet, R., Wiley, D.C. (2001). Structure of the ABC ATPase domain of human TAP1, the transporter associated with antigen processing. *EMBO J.* 20, 4964-4972.

- [179] Wild, K., Grafmuller, R., Wagner, E. and Schulz, G.E. (1997). Structure, catalysis and supramolecular assembly of adenylate kinase from maize. *Eur J Biochem* 250, 326-31.
- [180] Iyidogan, P. and Lutz, S. (2008). Systematic exploration of active site mutations on human deoxycytidine kinase substrate specificity. *Biochemistry* 47, 4711-20.
- [181] Reyes, C.L., Rutenber, E., Walter, P. and Stroud, R.M. (2007). X-ray structures of the signal recognition particle receptor reveal targeting cycle intermediates. *PLoS ONE* 2, e607.
- [182] Podobnik, M., Weitze, T.F., O'Donnell, M. and Kuriyan, J. (2003). Nucleotide-induced conformational changes in an isolated *Escherichia coli* DNA polymerase III clamp loader subunit. *Structure* 11, 253-63.
- [183] Manly, C.J., Chandrasekhar, J., Ochterski, J.W., Hammer, J.D. and Warfield, B.B. (2008). Strategies and tactics for optimizing the Hit-to-Lead process and beyond--a computational chemistry perspective. *Drug Discov Today* 13, 99-109.
- [184] Vieth, M., Higgs, R.E., Robertson, D.H., Shapiro, M., Gragg, E.A. and Hemmerle, H. (2004). Kinomics-structural biology and chemogenomics of kinase inhibitors and targets. *Biochim Biophys Acta* 1697, 243-57.
- [185] Manning, G., Whyte, D.B., Martinez, R., Hunter, T. and Sudarsanam, S. (2002). The protein kinase complement of the human genome. *Science* 298, 1912-34.
- [186] Pevarello, P. et al. (2006). 3-Amino-1,4,5,6-tetrahydropyrrolo[3,4-c]pyrazoles: a new class of CDK2 inhibitors. *Bioorg Med Chem Lett* 16, 1084-90.
- [187] Wei, L. and Altman, R.B. (1998). Recognizing protein binding sites using statistical descriptions of their 3D environments. *Pac Symp Biocomput*, 497-508.
- [188] Nebel, J.C., Herzyk, P. and Gilbert, D.R. (2007). Automatic generation of 3D motifs for classification of protein binding sites. *BMC Bioinformatics* 8, 321.
- [189] Kasuya, A. and Thornton, J.M. (1999). Three-dimensional structure analysis of PROSITE patterns. *J Mol Biol* 286, 1673-91.
- [190] Wu, S., Liang, M.P. and Altman, R.B. (2008). The SeqFEATURE library of 3D functional site models: comparison to existing methods and applications to protein function annotation. *Genome Biol* 9, R8.

- [191] Halperin, I., Glazer, D.S., Wu, S. and Altman, R.B. (2008). The FEATURE framework for protein function annotation: modeling new functions, improving performance, and extending to novel applications. *BMC Genomics* 9 Suppl 2, S2.
- [192] Ferre, F., Ausiello, G., Zanzoni, A. and Helmer-Citterich, M. (2005). Functional annotation by identification of local surface similarities: a novel tool for structural genomics. *BMC Bioinformatics* 6, 194.
- [193] Pierce, A.C., Rao, G. and Bemis, G.W. (2004). BREED: Generating novel inhibitors through hybridization of known ligands. Application to CDK2, p38, and HIV protease. *J Med Chem* 47, 2768-75.
- [194] Ramensky, V., Sobol, A., Zaitseva, N., Rubinov, A. and Zosimov, V. (2007). A novel approach to local similarity of protein binding sites substantially improves computational drug design results. *Proteins* 69, 349-57.
- [195] Guha, R., Howard, M.T., Hutchison, G.R., Murray-Rust, P., Rzepa, H., Steinbeck, C., Wegner, J. and Willighagen, E.L. (2006). The Blue Obelisk-interopability in chemical informatics. *J Chem Inf Model* 46, 991-8.
- [196] <http://pubchem.ncbi.nlm.nih.gov>.
- [197] Hasegawa, M. et al. (2007). Discovery of novel benzimidazoles as potent inhibitors of TIE-2 and VEGFR-2 tyrosine kinase receptors. *J Med Chem* 50, 4453-70.
- [198] Formats, C. [www.mdl.com/downloads/public/ctfile/ctfile.pdf](http://www.mdl.com/downloads/public/ctfile/ctfile.pdf).
- [199] Novartis. (2006). 021588s009Ibl.
- [200] Huang, N., Shoichet, B.K. and Irwin, J.J. (2006). Benchmarking sets for molecular docking. *J Med Chem* 49, 6789-801.
- [201] Schroeder, G.M. et al. (2008). Identification of pyrrolo[2,1-f][1,2,4]triazine-based inhibitors of Met kinase. *Bioorg Med Chem Lett* 18, 1945-51.

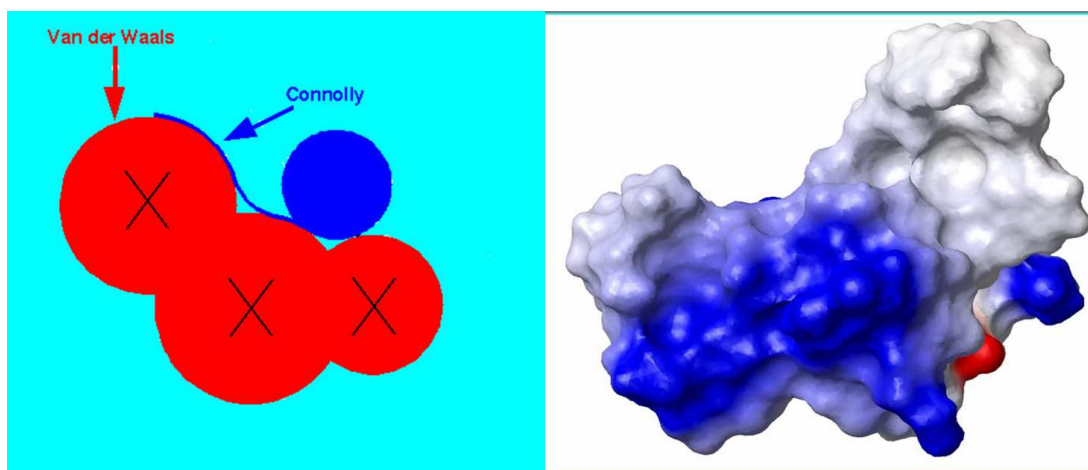
# ANNEXES



---

## **Annexe 1 : Surface de Connolly**

La surface de Connolly (ou surface moléculaire) est définie en faisant rouler une sphère sonde d'un rayon égal à celui d'une molécule d'eau sur la surface de Van der Waals. Elle est a été définie par Michael Connolly [129]. La Figure 57 illustre la description précédente et montre en exemple la surface de la protéine  $\delta$ -ACTX-Hv1a.



**Figure 57 : Exemples de surface de Connolly.**

À gauche, les sphères rouges représentent les rayons de Van der Waals des atomes, la sphère bleue représente la sonde, et la surface bleue qui se dessine est la surface de Connolly. A droite est représentée la surface de Connolly (ou surface moléculaire) de la protéine  $\delta$ -ACTX-Hv1a. Figure extraite de [143].

## **Annexe 2 : Description complète des lignes de commandes MED-sumo-clui**

La Figure 20 est une représentation de l'architecture globale de MED-SuMo. Trois interfaces permettant l'utilisation des fonctionnalités de MED-SuMo sont disponibles. Les deux premières, MED-sumo-clui et MED-sumo-lua sont locale. Cette annexe décrit les fonctionnalités des commandes disponibles via l'interface MED-sumo-clui. Trois types de commandes existent : 'sumo pdb' gère la base *pdb\_index.db* (cf. partie II.i), 'sumo graphdb' gère les bases de graphes *sites*, *full* et *frags* (cf. parties II.ii, II.iii et III.i), et 'sumo frag3d' gère la base de fragment générée par le programme MED-Fragmentor (cf. partie III.i). Une description des lignes de commande disponibles suit:

```
sumo pdb ...  
    Manipule l'index PDB.  
    pdb where  
    Affiche l'emplacement de l'index PDB.  
    pdb info
```

---

Affiche les statistiques de l'index PDB.

```
pdb get ID
```

Affiche le contenu du fichier PDB ID.

```
pdb list [RE]
```

Liste les ID PDB et les noms de fichiers (matchant l'optionnelle expression régulière RE)

```
pdb query QUERY
```

Liste les ID PDB et les HEADER matchant la requête QUERY.

```
pdb check
```

Vérifie l'index PDB et affiche les fichiers nouveaux, mis à jour et ceux manquants.

```
pdb update
```

Met à jour l'index PDB

```
pdb update_fragments
```

Met à jour l'index PDB et la base de données frag\_info.db.

```
pdb add PATH...
```

Ajoute des fichiers à l'index PDB.

```
pdb drop [ID | PATH]
```

Retire des entrées de l'index PDB; l'argument est soit un ID, un nom de fichier ou un nom de répertoire

```
pdb annot PATH
```

Ajoute les annotations d'un fichier.

```
pdb get_annots
```

Affiche la liste d'annotations existant dans l'index PDB.

```
pdb blacklist [add ID... | drop ID... | clear | show]
```

Gère la liste d'entrées de la PDB blacklistées.

```
pdb ligands ID
```

Affiche la liste des ligands de l'entrée PDB ID

```
pdb frag_lig_info ID
```

Affiche la listes des atomes des ligands de l'entrée PDB ID

```
sumo graphdb ...
```

Manipule les bases MED-SuMo

```
graphdb where
```

Donne l'emplacement des bases de graphes (graphdb)

```
graphdb info
```

Affiche les statistiques des graphdb

```
graphdb update [sites|full|frags]
```

Lance la compilation des graphdb en fonction de l'index PDB

```
graphdb stats [sites|full|frags] [IDS...]
```

---

Affiche les statistiques de certains graphes

sumo frag3d ...

Manipule la base *frag\_info.db*

frag3d where

Donne l'emplacement de la base *frag\_info.db*

frag3d init

Lis l'index PDB et lance la fragmentation des ligands par le programme MED-Fragmentor

frag3d recover

Reprend la fragmentation où elle s'est arrêtée si le calcul est interrompu

frag3d info

Affiche le nombre de fragments stockés dans *frag\_info.db*

frag3d list ID

Affiche la liste des noms des fragments correspondant au fichier PDB ID

### **Annexe 3 : Description des requêtes MED-sumo-client.**

L'interfaçage client-serveur de MED-SuMo se fait par l'intermédiaire d'une liste de requête permettant d'accéder aux fonctionnalités de MED-SuMo serveur. Cette annexe contient une brève description des requêtes disponibles et qui n'ont pas été décrites dans le manuscrit. Une requête se lance comme illustré dans la partie haute de la Figure 15 soit :

```
sumo-client -address address -port #####- u user -p passwd -command  
"{requête}"
```

available-functions

Permet de récupérer la liste de requête disponible pour le sumo-client

available-profiles

cf. partie III.ii

available-annotations

cf. partie III.ii

job-cancel

Permet d'annuler un *run* MED-SuMo en utilisant son identifiant

job-status

Vérifie le statut d'un *run* lancé. Il peut être en cours, fini, ou infructueux

pdb-headers-get

Permet de récupérer les annotations stockées dans l'index PDB pour une structure

pdb-headers-query

Permet de récupérer une liste de structure PDB ayant des annotations précises

---

`sumo-classify`

Classe les résultats d'un run MED-SuMo en fonction de leurs signatures en SCFs, permet la création du dendrogram dans MED-SuMo GUI

`sumo-delete-results`

Efface les résultats des runs MED-SuMo stockés coté serveur

`sumo-full-results`

Permet de récupérer les resultants d'un *run*, différent mode sont accessible : mode CS (pour MED-SuMo GUI), mode XML, mode TEXT

`sumo-select`

Permet la construction de la requête qui sera soumise, par exemple, le site de liaison n°1, ou toute la surface d'une protéine

`sumo-check`

Vérifie si la requête soumise est valide

`sumo-groups`

Permet de récupérer les informations 3D des SCFs et dans les triplets formés (n'est utilisé que pas MED-SuMo GUI)

`sumo-molecule`

Permet de récupérer les informations 3D des SCFs, ainsi que les triplets formés (n'est utilisé que pas MED-SuMo GUI)

`sumo-ligands-results`

Récupère les informations des ligands des n meilleurs hits

`sumo-list-results`

Liste les résultats des *runs* disponible sur le serveur

`sumo-run`

Permet de lancer un *run* MED-SuMo (cf. Figure 15)

**Annexe 4 : Résultats de la classification des protéines de la superfamille SCOP GHKL :**

MED-SMA Cl	PDB_LIG_Id	Ligand	SCOP Family	MED-SMA Cl	PDB_LIG_Id	Ligand	SCOP Family	MED-SMA Cl	PDB_LIG_Id	Ligand	SCOP Family	MED-SMA Cl	PDB_LIG_Id	Ligand	SCOP Family
Cl.1	HEI_132	AMP	DNA_Gyrase_B_Ecoli	Cl.4	IB2_116	ADP	Mill	Cl.4	2BRE_119	K12	HSP90_Yeast	Cl.4	2BRE_119	K12	HSP90_Yeast
Cl.1	HEI_30	AMP	DNA_Gyrase_B_Ecoli	Cl.4	IB3_131	AMP	Mill	Cl.4	2BRE_218	K12	HSP90_Yeast	Cl.4	2BRE_218	K12	HSP90_Yeast
Cl.1	IKX0_131	AMP	TOPO_VI	Cl.4	IB50_155	RDC	HSP90_Yeast	Cl.4	2BSM_177	BSM	HSP90_Human	Cl.4	2BSM_177	BSM	HSP90_Human
Cl.1	IKX0_229	AMP	TOPO_VI	Cl.4	IB10_139	ADP	HSP90_Human	Cl.4	2B10_181	C15	HSP90_Human	Cl.4	2B10_181	C15	HSP90_Human
Cl.1	IKX0_328	AMP	TOPO_VI	Cl.4	IEA6_110	ADP	PMS2	Cl.4	2B10_279	C15	HSP90_Human	Cl.4	2B10_279	C15	HSP90_Human
Cl.1	IKX0_425	AMP	TOPO_VI	Cl.4	IEA6_2109	ADP	PMS2	Cl.4	2B1H_189	2D7	HSP90_Human	Cl.4	2B1H_189	2D7	HSP90_Human
Cl.1	IKX0_522	AMP	TOPO_VI	Cl.4	HNJ1_46	ATG	PMS2	Cl.4	2B1V_174	2D0	HSP90_Human	Cl.4	2B1V_174	2D0	HSP90_Human
Cl.1	IKX0_621	AMP	TOPO_VI	Cl.4	HNJ1_244	ATG	PMS2	Cl.4	2B25_170	AB4	HSP90_Human	Cl.4	2B25_170	AB4	HSP90_Human
Cl.1	IPV6_11	AMP	DNA_TOPO_II_B yeast	Cl.4	LM6_183	ADP	Pyruvate_dehydrogenase_Kinase	Cl.4	2B25_285	AB4	HSP90_Human	Cl.4	2B25_285	AB4	HSP90_Human
Cl.1	IPV6_20	AMP	DNA_TOPO_II_B yeast	Cl.4	MMH1_143	AMP	Mill	Cl.4	2CCS_188	4B4	HSP90_Human	Cl.4	2CCS_188	4B4	HSP90_Human
Cl.1	IOZR_117	COX	DNA_TOPO_II_B yeast	Cl.4	MHI_108	AMP	Mill	Cl.4	2CCU_130	"2ET"	HSP90_Human	Cl.4	2CCU_130	"2ET"	HSP90_Human
Cl.1	IOZR_2113	AMP	DNA_TOPO_II_B yeast	Cl.4	MHI_142	AMP	Mill	Cl.4	2CCU_197	2D9	HSP90_Human	Cl.4	2CCU_197	2D9	HSP90_Human
Cl.1	IOZR_3111	AMP	DNA_TOPO_II_B yeast	Cl.4	IOF_183	KOS	HSP90_Human	Cl.4	2CDD_196	C15	HSP90_Human	Cl.4	2CDD_196	C15	HSP90_Human
Cl.1	IS6_1102	AMP	TOPO_IV	Cl.4	IOV_113	NEC	HSP90_Dog	Cl.4	2CDD_294	C15	HSP90_Human	Cl.4	2CDD_294	C15	HSP90_Human
Cl.1	IS6_2100	AMP	TOPO_IV	Cl.4	IOV_187	RDI	HSP90_Dog	Cl.4	2EXL_141	GMV	HSP90_Dog	Cl.4	2EXL_141	GMV	HSP90_Dog
Cl.1	IS6_117	ADP	TOPO_VI	Cl.4	IOVE_143	CDY	HSP90_Dog	Cl.4	2FV_240	H84	HSP90_Dog	Cl.4	2FV_240	H84	HSP90_Dog
Cl.1	IS6_111	ADP	TOPO_VI	Cl.4	ITBW_1107	AMP	HSP90_Dog	Cl.4	2FV_212	H84	HSP90_Dog	Cl.4	2FV_212	H84	HSP90_Dog
Cl.1	IS6_28	ADP	TOPO_VI	Cl.4	ITBW_2106	AMP	HSP90_Dog	Cl.4	2FV_218	H71	HSP90_Dog	Cl.4	2FV_218	H71	HSP90_Dog
Cl.1	IS6_188	ADP	TOPO_VI	Cl.4	ITCO_125	ATP	HSP90_Dog	Cl.4	2FYP_180	RDE	HSP90_Dog	Cl.4	2FYP_180	RDE	HSP90_Dog
Cl.1	IS6_284	ADP	TOPO_VI	Cl.4	ITCO_2121	ATP	HSP90_Dog	Cl.4	2FYP_239	RDE	HSP90_Dog	Cl.4	2FYP_239	RDE	HSP90_Dog
Cl.1	IS6_19	ADP	TOPO_VI	Cl.4	ITC6_116	ADP	HSP90_Dog	Cl.4	2GFD_172	RDA	HSP90_Dog	Cl.4	2GFD_172	RDA	HSP90_Dog
Cl.1	IS6_25	SAP	TOPO_VI	Cl.4	ITC6_2114	ADP	HSP90_Dog	Cl.4	2GFD_267	RDA	HSP90_Dog	Cl.4	2GFD_267	RDA	HSP90_Dog
Cl.2	IS6_112	SAP	alpha-ketoadid_dehydrogenase_kinase	Cl.4	IU07_126	PA7	HSP90_Dog	Cl.4	2GMP_130	PA7	HSP90_Dog	Cl.4	2GMP_130	PA7	HSP90_Dog
Cl.2	IS6_133	ADP	alpha-ketoadid_dehydrogenase_kinase	Cl.4	IU2_195	RDC	HSP90_Dog	Cl.4	2GMP_2128	PA7	HSP90_Dog	Cl.4	2GMP_2128	PA7	HSP90_Dog
Cl.2	IS6_118	"128"	Histidine_Kinase_CheA	Cl.4	IU2_633	RDC	HSP90_Dog	Cl.4	2H51_122	D28	HSP90_Human	Cl.4	2H51_122	D28	HSP90_Human
Cl.2	IS6_110	AMP	Histidine_Kinase_PhoQ	Cl.4	IU2_13	NEC	HSP90_Dog	Cl.4	2H81_139	NEI	HSP90_Dog	Cl.4	2H81_139	NEI	HSP90_Dog
Cl.2	IS6_261	ADP	Pyruvate_dehydrogenase_Kinase	Cl.4	IU2_22	NEC	HSP90_Dog	Cl.4	2H81_2138	NEI	HSP90_Dog	Cl.4	2H81_2138	NEI	HSP90_Dog
Cl.2	LU0_175	ADP	Anti-sigma_factor_spoIIab	Cl.4	IU6_188	PUS	HSP90_Human	Cl.4	2HCH_142	NSA	HSP90_Dog	Cl.4	2HCH_142	NSA	HSP90_Dog
Cl.2	LU0_273	ADP	Anti-sigma_factor_spoIIab	Cl.4	IUV_16	PUS	HSP90_Human	Cl.4	2HCH_240	NSA	HSP90_Dog	Cl.4	2HCH_240	NSA	HSP90_Dog
Cl.2	TH8_123	ADP	Anti-sigma_factor_spoIIab	Cl.4	IUV_182	PUS	HSP90_Human	Cl.4	2HG1_136	NSO	HSP90_Dog	Cl.4	2HG1_136	NSO	HSP90_Dog
Cl.2	TH8_124	ADP	Anti-sigma_factor_spoIIab	Cl.4	IUV_144	PUS	HSP90_Human	Cl.4	2HG1_234	NSO	HSP90_Dog	Cl.4	2HG1_234	NSO	HSP90_Dog
Cl.2	TH1_104	ADP	Anti-sigma_factor_spoIIab	Cl.4	IUV_174	PUS	HSP90_Human	Cl.4	2HKJ_138	RDC	HSP90_Dog	Cl.4	2HKJ_138	RDC	HSP90_Dog
Cl.2	TH1_2101	ADP	Anti-sigma_factor_spoIIab	Cl.4	IUV_141	PUS	HSP90_Human	Cl.4	2WS_148	NP4	HSP90_Yeast	Cl.4	2WS_148	NP4	HSP90_Yeast
Cl.2	TH1_135	ATP	Anti-sigma_factor_spoIIab	Cl.4	IUV_171	PUS	HSP90_Human	Cl.4	2WU_145	NP5	HSP90_Yeast	Cl.4	2WU_145	NP5	HSP90_Yeast
Cl.2	TH1_232	ATP	Anti-sigma_factor_spoIIab	Cl.4	IUV_108	PUS	HSP90_Human	Cl.4	2WV_100	MIS	HSP90_Yeast	Cl.4	2WV_100	MIS	HSP90_Yeast
Cl.2	TH1_27	ATP	Anti-sigma_factor_spoIIab	Cl.4	IUV_168	PUS	HSP90_Human	Cl.4	2WV_128	2GG	HSP90_Human	Cl.4	2WV_128	2GG	HSP90_Human
Cl.2	TH1_24	ATP	Anti-sigma_factor_spoIIab	Cl.4	IUV_1135	PUS	HSP90_Human	Cl.5	IB8_129	ADP	Histidine_Kinase_CheA	Cl.5	IB8_129	ADP	Histidine_Kinase_CheA
Cl.2	TH1_323	ATP	Anti-sigma_factor_spoIIab	Cl.4	IUV_1133	PUS	HSP90_Human	Cl.5	IB8_2127	ADP	Histidine_Kinase_CheA	Cl.5	IB8_2127	ADP	Histidine_Kinase_CheA
Cl.2	2CZA_120	ADP	Sensor_histidine_kinase_TM0883	Cl.4	IUV_132	PUS	HSP90_Human	Cl.5	IB9_157	AMP	Histidine_Kinase_CheA	Cl.5	IB9_157	AMP	Histidine_Kinase_CheA
Cl.2	2CH4_156	AMP	Histidine_Kinase_CheA	Cl.4	IYCI_115	4BC	HSP90_Human	Cl.5	IB9_254	ADP	Histidine_Kinase_CheA	Cl.5	IB9_254	ADP	Histidine_Kinase_CheA
Cl.3	IAD_176	NOV	DNA_GYRASE_B_Ecoli	Cl.4	IYCI_145	4BC	HSP90_Human	Cl.5	IB6_151	ADP	Histidine_Kinase_CheA	Cl.5	IB6_151	ADP	Histidine_Kinase_CheA
Cl.3	IAD_186	NOV	DNA_GYRASE_B_Ecoli	Cl.4	IYCI_115	4BC	HSP90_Human	Cl.5	IB6_249	ADP	Histidine_Kinase_CheA	Cl.5	IB6_249	ADP	Histidine_Kinase_CheA
Cl.3	IAD_184	NOV	DNA_GYRASE_B_TT	Cl.4	IYCI_189	43P	HSP90_Human	Cl.5	IB6_119	AMP	Histidine_Kinase_CheA	Cl.5	IB6_119	AMP	Histidine_Kinase_CheA
Cl.3	IKZ_152	CBN	DNA_GYRASE_B_Ecoli	Cl.4	IYET_139	GDM	HSP90_Human	Cl.5	IB6_215	AMP	Histidine_Kinase_CheA	Cl.5	IB6_215	AMP	Histidine_Kinase_CheA
Cl.3	ISM_1105	NOV	TOPO_IV	Cl.4	IYSL_131	NEC	HSP90_Dog	Cl.5	IB6_150	ADP	Histidine_Kinase_CheA	Cl.5	IB6_150	ADP	Histidine_Kinase_CheA
Cl.3	ISM_2103	NOV	TOPO_IV	Cl.4	IYTO_180	ADP	HSP90_Dog	Cl.5	IB6_247	ADP	Histidine_Kinase_CheA	Cl.5	IB6_247	ADP	Histidine_Kinase_CheA
Cl.4	IAD_162	GMV	HSP90_Yeast	Cl.4	IYVH_137	H84	HSP90_Yeast	Cl.5	IB6_247	ADP	Histidine_Kinase_CheA	Cl.5	IB6_247	ADP	Histidine_Kinase_CheA
Cl.4	IAD_137	ADP	HSP90_Yeast	Cl.4	IYVH_158	RDE	HSP90_Yeast	Cl.5	IB6_247	ADP	Histidine_Kinase_CheA	Cl.5	IB6_247	ADP	Histidine_Kinase_CheA
Cl.4	IAD_17	ADP	HSP90_Yeast	Cl.4	ZBR_120	C15	HSP90_Yeast	Cl.5	IB6_247	ADP	Histidine_Kinase_CheA	Cl.5	IB6_247	ADP	Histidine_Kinase_CheA

## Articles

---

## ARTICLES

### Article 1 :

#### **Functional annotation strategy for protein structures**

Olivia Doppelt, Fabrice Moriaud, Aurélie Bornot and Alexandre G. de Brevern

### Abstract

Whole-genome sequencing projects are a major source of unknown function proteins. However, as predicting protein function from sequence remains a difficult task, research groups recently started to use 3D protein structures and structural models to bypass it. MED-SuMo compares protein surfaces analyzing the composition and spatial distribution of specific chemical groups (hydrogen bond donor, acceptor, positive, negative, aromatic, hydrophobic, guanidinium, hydroxyl, acyl and glycine). It is able to recognize proteins that have similar binding sites and thus, may perform similar functions. We present here a fine example which points out the interest of MED-SuMo approach for functional structural annotation.

### Article 2 :

#### **Analysis of HSP90 related folds with MED-SuMo classification approach.**

Olivia Doppelt-Azeroual, Fabrice Moriaud, François Delfaud and Alexandre G. de Brevern

### Abstract

Three-dimensional structural information is critical for understanding functional protein properties and the precise mechanisms of protein functions implicated in physiological and pathological processes. Comparison and detection of protein binding sites are key steps for annotating structures with functional predictions and are extremely valuable steps in a drug design process. In this research area, MED-SuMo is a powerful technology to detect and characterise similar local regions on protein surfaces. Each amino acid residue's potential chemical interactions are represented by specific Surface Chemical Features (SCFs). The MED-SuMo heuristic is based on the representation of binding sites by a graph structure suitable for exploration by an efficient comparison algorithm. We use this approach to analyze one particular SCOP superfamily which includes HSP90 chaperone, MutL/DNA topoisomerase, histidine kinases and  $\alpha$ -ketoacid dehydrogenase kinase C (BCK). They share a common fold and a common region for ATP-binding. To analyze both similar and differing features of this fold, we use a novel classification method, the MED-SuMo Multi approach

---

(MED-SMA). We highlight common and distinct features of these proteins. The different clusters created by MED-SMA yield interesting observations. For instance, one cluster gathers three types of proteins (HSP90, topoisomerase VI and BCK) which all bind the drug radicicol.

### **Article 3**

#### **Computational Fragment-Based Approach at PDB Scale by Protein Local Similarity**

Fabrice Moriaud, Olivia Doppelt-Azeroual, Laetitia Martin, Ksenia Oguievetskaia,  
Kerstin Koch, Artem Vorotyntsev, Stewart A. Adcock and François Delfaud

#### **Abstract**

The large volume of protein-ligand structures now available enables innovative and efficient protocols in computational FBDD (Fragment-Based Drug Design) to be proposed based on experimental data. In this work, we build a database of MED-Portions, where a MED-Portion is a new structural object encoding protein-fragment binding sites. MED-Portions are derived from mining all available protein-ligand structures with any library of small molecules. Combined with the MED-SuMo software to superpose similar protein interaction surfaces, pools of matching MED-Portions can be retrieved from any binding surface query. The rapidity of this technology allows its application to a diverse set of 107 protein binding sites. The selectivity of the protocol is shown by a qualitative correlation between the average hydrophobicity of the pools of MED-Portions and those of the binding sites. To generate hitlike molecules, MED-Portions are combined in 3D with the MED-Hybridise toolkit. Our MED-Portion/MED-SuMo/MED-Hybridise protocol is applied to two targets that represent important protein superfamilies in drug design: a protein kinase and a G-Protein Coupled Receptor (GPCR). We retrieved actives molecules of PubChem bioassays for the two targets. The results show the potential for finding relevant leads from any protein 3D structure since the occurrence of interfamily MED-Portions is 25% for protein kinase and almost 100% for the GPCR.

### **Article 4**

#### **MED-SuMo applications**

Olivia Doppelt-Azeroual, Fabrice Moriaud, Stewart Adcock and François Delfaud

#### **Abstract**

Three-dimensional protein structures are a major source of information to help understand functional protein properties which is highly influencing drug design



---

computational methods. Target-based drug design approaches assist in designing and optimizing compounds that bind to specific targets. MED-SuMo is a powerful technology to localize similar local regions on protein surfaces. It is a target-based and uses the concept of exploiting all macromolecule structures available in the PDB. Therefore, it is in contrast with widely used methods such as docking/scoring, de novo design or map-based methods. MED-SuMo is covering and contributing to a large panel of drug discovery applications. Here, we describe applications which are specific to this kind of methods. For example, functional annotation, pocket profiling, structural superpositions, and automated functional binding site classification. Others applications assists in an innovative way, the medicinal chemist and the molecular modeller in more frequent task like lead discovery and lead optimization like drug repurposing and Fragment-Based drug design.

## **Article 5**

### **Fast and Automated Functional Classification with MED-SuMo: An Application on the Purinome**

Olivia Doppelt-Azeroual, Kerstin Koch, François Delfaud, Fabrice Moriaud and  
Alexandre G. de Brevern

#### **Abstract**

##### **Background**

Ligand-protein interactions are essential at biological processes. The precise characterization of protein binding sites is crucial to understand protein function. MED-SuMo is a powerful technology to localize similar local regions on protein surfaces. Its heuristic is based on a 3D representation of macromolecules using precise Surface Chemical Features associating chemical characteristics with geometrical properties.

Here we present a new automated and fast classification method (MED-SMA). It uses MED-SuMo ability to detect and characterize local similarities to classify binding sites. After the comparison of all the binding sites, a similarity graph is built and is classified with Markov Clustering algorithm.

##### **Results**

In this work, we classify a dataset of 2322 purine binding sites. Indeed, purine binding sites are particularly studied druggable sites as they are more catalytic than functional. So, instead of gathering binding sites with the same functions, we regroup binding sites with similar binding modes *i.e.* they can be inhibited or activated by the same molecules. 247

---

clusters are created. Results are analyzed in regard to PROSITE annotations and carefully refined functional annotations from the PDB. As expected, binding sites with related mechanism are gathered, *e.g.*, the Small GTPases. Nevertheless, protein kinases from different Kinome family can also be found in the same clusters, *e.g.*, Aurora-A and CDK2 that are inhibited by the same molecules. Representative examples of different clusters are presented.

### **Conclusion**

The efficiency of MED-SMA approach is shown as it gathers binding sites of proteins with similar Structure Activity Relation (SAR) (CDK2 and GSK3 $\beta$ ). Moreover, an efficient new enrichment protocol enables to associate apo-structures to the binding site classification.

The gathering of binding sites with similar binding modes can be used in target based drug design applications or to predict cross-reactivity and potential toxic side effects. MED-SMA can also be easily used at a proteome level.