# Algorithmes pour la prédiction in silico d'interactions par similarité entre macromolécules biologiques

Alice Voland

université PARIS-SACLAY

NNT : 2017SACLV014

Thèse de doctorat

UNIVERSITÉ DE VERSAILLES ST-QUENTIN-EN-YVELINES

université PARIS-SACLAY

# Algorithmes pour la prédiction in silico d'interactions par similarité entre macromolécules biologiques

Thèse de doctorat de l'Université Paris-Saclay
préparée à l'Université de Versailles-Saint-Quentin-en-Yvelines

École doctorale n°580 Sciences et Technologies de l'Information et de la communication (STIC) Spécialité de doctorat: Informatique

Thèse présentée et soutenue à Versailles, le 3 avril 2017, par

**Alice Voland**

Composition du Jury :

| | |
|---|---|
| Alain Denise<br>Professeur, Université Paris-Sud | Président du jury |
| Anne-Claude Camproux<br>Professeur, Université Paris Diderot | Rapporteur |
| Guillaume Fertin<br>Professeur, Université de Nantes | Rapporteur |
| Dominique Barth<br>Professeur, Université de Versailles-Saint-Quentin | Directeur de thèse |
| Sandrine Vial<br>Maître de conférence, Université de Versailles-Saint-Quentin | Co-encadrante |
| Benjamin Schwarz<br>Docteur directeur R&D, Bionext SA | Co-encadrant |

# Remerciements

Les trois années et demie qui m'ont conduit au terme de cette thèse constituent un investissement personnel et professionnel, que je n'aurais pu réussir sans les soutiens de mes proches. C'est pourquoi je tiens à remercier en tout premier lieu mon épouse, Cyrielle, pour une quantité de raisons que je n'aurais pas la place d'énumérer ici. Je résumerais simplement en tentant d'exprimer ma gratitude pour sa confiance inconditionnelle, et son soutien permanent. Merci.

Je souhaite remercier Dominique Barth pour avoir dirigé ma thèse. Je le remercie pour sa confiance dans certains de mes choix, autant que pour ses inestimables conseils et son encadrement guidés par son expérience. Je remercie ma co-encadrante Sandrine Vial pour sa disponibilité constante tout au long de ma thèse, pour ses efforts pour comprendre ma situation, et pour avoir su guider et accompagner mes travaux. Je remercie également mon co-encadrant Benjamin Schwarz pour avoir su intégrer le suivi scientifique de ma thèse au sein de la vie de l'entreprise.

Je remercie les rapporteurs de cette thèse, Anne-Claude Camproux et Guillaume Fertin, pour leurs analyses complémentaires et pertinentes de ma thèse fortement multi-disciplinaire. Je remercie également Alain Denise pour avoir accepté de présider ma soutenance.

Je veux remercier l'ensemble de l'équipe de Bionext. Ceux que j'ai eu le plaisir de voir rejoindre cette aventure et en particulier Pascal Muller qui a pris le temps de vulgariser son expertise en chimie. Et celui qui était déjà présent, Nicolas Gagnière notre CTO, qui m'a aidé à différentes étapes tant sur le plan des infrastructures informatiques que pour mieux comprendre la biologie structurale ou encore pour repérer les coquilles dans le manuscrit.

Je remercie également Serge Albou. Je n'oublie pas son soutien à la fois professionnel et personnel qu'il a toujours su concilier aux impératifs de la direction d'entreprise.

Je veux remercier tout particulièrement Jorge Ramirez qui a su, par l'encadrement de mon stage de master au CNRS, me donner le goût pour la recherche et qui m'a transmis l'offre de thèse qui m'a conduit ici. Je remercie aussi les autres enseignants et chercheurs de l'Université de Montpellier qui m'ont encouragé et soutenu à poursuivre dans un doctorat en informatique après un master en mathématiques, et en particulier Rodolphe Giroudeau.

à Victoria

# Table des matières

# Introduction

Le coût de développement d'un médicament n'a jamais été aussi élevé, et la durée aussi grande. Selon les différentes études ce coût est en général estimé entre 1 et 2 milliards de dollars américains, pour une durée entre 8 et 20 ans avant une mise sur le marché [Paul 2010, Khanna 2012]. De plus ces chiffres sont en constante croissance, et s'ils peuvent s'expliquer en partie par la rigueur des réglementations, notamment européennes et américaines, ils montrent principalement l'utilité et la nécessité de trouver des outils innovants et performants pour proposer de nouvelles thérapies.

Les avancées dans la biologie structurale permettent de disposer maintenant des structures tridimensionnelles de nombreuses molécules biologiques dont les interactions chimiques sont responsables de certaines pathologies, ainsi que les éventuelles thérapies. L'accès à ces données motive l'utilisation d'outils *in silico* (informatiques), moins chers et plus rapides que les expériences *in vitro* ou *in vivo* afin notamment de prédire les interactions entre les molécules. Ces outils peuvent s'intégrer dans le processus de développement d'un médicament d'une part en guidant le choix des composés ayant les meilleurs potentiels thérapeutiques et d'autre part en étudiant les mécanismes d'action au niveau moléculaire de médicaments déjà sur le marché. Certaines méthodes *in silico* sont déjà couramment employées pour optimiser le choix d'un composé comme médicament candidat, cependant l'espace des cibles, c'est-à-dire des molécules biologiques sur lesquelles les médicaments agissent, reste encore largement à explorer [Overington 2006]. En particulier l'action d'un même médicament sur plusieurs cibles peut expliquer certains effets secondaires indésirables et dangereux, ou éventuellement des effets positifs pour de nouvelles indications, ou encore un effet lié à l'action simultanée sur plusieurs cibles [Hopkins 2008]. Ce constat met en lumière le besoin de nouvelles méthodes *in silico* permettant une étude plus exhaustive de cet espace des cibles [Lavecchia 2015].

L'entreprise BIONEXT SA propose différents logiciels pour analyser, visualiser, et prédire les interactions entre molécules biologiques. La présente thèse, réalisée dans le cadre d'une collaboration avec le laboratoire DAVID de l'Université Versailles-Saint-Quentin, a pour objectif l'étude des méthodes *in silico* de prédiction d'interaction entre les molécules à partir des structures tridimensionnelles. Il s'agit plus précisément d'étudier les interactions entre des petites molécules, typiquement des médicaments, et les grandes molécules biologiques, ou macromolécules, telles que les protéines. L'approche choisie consiste à utiliser les similarités entre les structures des macromolécules biologiques pour déduire certaines similarités dans la capacité d'interaction avec d'autres molécules. Un algorithme est développé dans le cadre de la thèse, après une analyse d'un logiciel propriétaire pré-existant dans l'entreprise mais qui ne permet pas d'apporter de réponse satisfaisante à la problématique de la prédiction d'interactions. Une méthode de validation est également proposée, qui rend compte de la capacité de l'algorithme à apporter de nouvelles prédictions pertinentes du point de vue pharmacologique. Le mémoire présentant les travaux réalisés dans le cadre de la thèse est architecturé en quatre chapitres. Tout d'abord le problème informatique est défini dans le contexte de son application en biologie, puis les différentes approches existantes sont présentées, ensuite l'algorithme BIOBIND développé dans le cadre de la thèse est décrit, et enfin une validation de cet algorithme est proposée.

Le premier chapitre de la thèse présente la problématique pharmacologique de la prédiction de cibles pour une molécule à potentiel thérapeutique. Nous définissons la notion de molécules, en

distinguant les petites molécules qui sont appelées *ligands* et les grandes molécules appelées *cibles* sur lesquelles elles agissent en formant des *complexes ligand-cible*. Ces derniers complexes sont réalisés par des interactions chimiques qui permettent aux ligands de se fixer sur leurs cibles. Après avoir présenté les différents modèles moléculaires et décrit l'action des petites molécules au sein du système biologique que constitue les cellules, les enjeux pharmaceutiques et plus généralement les différentes problématiques biologiques liées à la prédiction de cibles seront énoncés. Les différentes approches *in silico* de prédiction d'interactions sont alors présentées : par évaluation de l'affinité d'un ligand avec une cible ou à partir d'une recherche de similarités entre molécules. En effet la similarité dans la structure tridimensionnelle des molécules peut traduire une similarité dans les propriétés chimiques et la capacité à lier les mêmes ligands.

Le second chapitre présente les différentes approches permettant de prédire de nouvelles cibles pour un ligand donné, par similarité avec une première cible connue. Il s'agit, à partir d'un premier *site de liaison* connu d'un ligand, c'est-à-dire la région d'une première cible capable de lier ce ligand, de rechercher dans un ensemble de macromolécules un motif similaire. Nous proposons une classification suivant quatre axes principaux : la représentation des molécules, la définition de la similarité, la méthode de recherche du motif, et l'alignement entre les motifs. Il existe de nombreuses approches, utilisant des modèles de la chimie et des techniques de recherches variées. Les avantages et inconvénients des différents points de vue sont étudiés avant de présenter les choix pour notre algorithme de recherche de similarité.

Le chapitre suivant détaille le fonctionnement de notre algorithme BioBind de recherche de similarités locales en surface des macromolécules. La surface des molécules est modélisée comme une triangulation où chaque sommet est associé à un atome accessible, c'est-à-dire susceptible d'être en interaction avec un atome d'une autre molécule. Cette construction est basée sur la théorie des formes alpha définissant un polytope à partir d'un ensemble de sphères. Elle permet de définir une notion de région de surface par sélection d'un ensemble quelconque de sommets. À partir de la définition d'une région requête de surface représentant le site de liaison connu, la recherche d'une région similaire à la surface d'une autre macromolécule est présentée comme un problème d'optimisation de recherche de la meilleure superposition. Pour résoudre l'exploration de ce vaste espace de recherche, une heuristique est proposée en superposant des régions circulaires, qui approximent un disque géodésique et sont définies de manière exhaustive sur notre représentation de la surface des macromolécules.

Enfin le dernier chapitre présente une validation de notre algorithme, notamment par une comparaison avec d'autres approches concurrentes. Nous proposons une méthode d'évaluation considérant la prédiction de cibles comme un problème de classification binaire. Étant données un ligand, une première cible connue, et un ensemble de cibles candidates, il s'agit de déterminer un sous-ensemble de ces cibles candidates prédites pour interagir également avec le ligand. Cette classification est paramétrée par un score, permettant d'utiliser la mesure définie par la caractéristique de performance ROC, ou l'aire AUC qu'elle délimite. Nous utilisons dans un premier temps deux jeux de données issus de la littérature, cependant ils ne sont pas tout à fait adaptés à notre problématique de prédiction de cibles notamment car les ligands choisis sont généralement peu représentatif des molécules à potentiel thérapeutique. Nous considérons alors un nouveau jeu de données développé par l'entreprise Bionext SA dans cet objectif. Sur l'ensemble des trois jeux de données, notre algorithme se compare favorablement à deux autres approches, un logiciel de *docking* et une approche par similarité.

Chapitre I

# Prédiction *in silico* d'interaction entre molécules biologiques

## Introduction

Les macromolécules biologiques, telles que les protéines, l'ADN, ou l'ARN, assurent le fonctionnement des cellules. Ces différentes fonctions sont régulées par les interactions chimiques qui se produisent entre elles ainsi qu'avec d'autres petites molécules, appelées ligands. En particulier les médicaments sont des ligands qui agissent en se liant à certaines macromolécules, appelées cibles. La compréhension de ce réseau d'interactions est nécessaire dans le processus de conception d'un médicament, on s'intéresse ici en particulier à la prédiction des cibles pour une molécule donnée.

Après une description d'un modèle des différents mécanismes chimiques et biologiques des cellules, ce premier chapitre présente la problématique de la prédiction *in silico* (informatique) d'interactions entre les molécules biologiques et plus précisément le problème de la prédiction des cibles à partir de la donnée des structures tridimensionnelles des molécules impliquées. Les différents enjeux pharmaceutiques et biologiques sont présentés, ainsi que les principales techniques permettant de résoudre le problème à partir des informations sur les structures tridimensionnelles des molécules impliquées. Il existe essentiellement deux types d'approches, la première consiste à prédire directement la possibilité d'une interaction à partir de la donnée seule des deux molécules concernées, et la seconde repose sur un principe d'inférence de l'interaction afin de traduire le problème en une recherche de similarité entre les molécules. Enfin, notre algorithme BioBind qui se situe dans cette seconde catégorie est présenté.

## Sommaire

# 1 Molécules biologiques et interactions chimique

Les molécules biologiques sont les molécules présentes dans les cellules des êtres vivants. Différentes interactions se produisent entre ces molécules, à l'origine du fonctionnement des cellules. Les concepts principaux sont présentés dans cette section, une introduction plus complète aux mécanismes de la biologie cellulaire peut être trouvée par exemple dans [Alberts 2013].

## 1.1 Molécules biologiques et modélisation

### 1.1.1 Molécules et macromolécules

Une *molécule* se définit comme un ensemble d'atomes et des liaisons covalentes entre ces atomes. Il s'agit de liaisons réalisées par un partage d'électrons entre deux ou plusieurs atomes qui définissent ainsi la géométrie des molécules. Ces liaisons covalentes sont à distinguer d'autres types de liaisons chimiques, plus faibles, qui peuvent exister entre des atomes d'une même molécule ou bien entre des atomes de différentes molécules.

Un *conformère* est un plongement d'une molécule dans l'espace tridimensionnel. Une même molécule peut ainsi avoir plusieurs conformères, principalement car certaines liaisons admettent un axe de rotation. On identifie la notion de conformère à la représentation tridimensionnelle d'une molécule, correspondant à la position et la nature de chaque atome qui compose la molécule. La figure 1 présente l'exemple d'une molécule ayant deux conformations distinctes observées.

Une *macromolécule* est définie comme une molécule « de grande taille », mais il n'existe pas de seuil canonique concernant cette taille que ce soit le volume, la masse, ou le nombre d'atomes. Cependant, dans le contexte des molécules biologiques, une définition naturelle consiste à appeler macromolécule toute molécule décomposable en une suite de petites molécules liées entre elles par des liaisons covalentes formant une chaîne, en suivant la définition IUPAC [1] [Gold 2014]. Il s'agit essentiellement des protéines, des acides ribonucléiques (ARN), et des acides désoxyribonucléiques (ADN), qui participent au fonctionnement des cellules de la plupart des êtres vivants. La figure 2 présente un exemple de macromolécule.

### 1.1.2 Structure des macromolécules biologiques

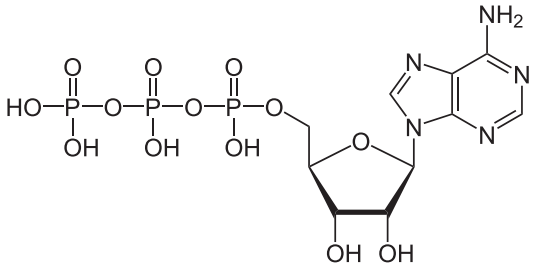
Les macromolécules biologiques peuvent être modélisées à plusieurs niveaux de granularité, appelés structures primaire, secondaire, tertiaire, et quaternaire. La figure 3 présente ces différents modèles pour un exemple de protéine.

La *structure primaire* correspond à la représentation de la séquence des petites molécules formant une chaîne dans une macromolécule, chaque petite molécule, appelée *résidu*, étant représentée par une lettre. La donnée de la séquence de ces résidus revient essentiellement à décrire la nature de la macromolécule, c'est-à-dire l'ensemble des atomes et liaisons covalentes qui la constituent, bien qu'il existe parfois également des liaisons covalentes entre résidus non consécutifs. Les résidus des protéines sont les acides aminés et sont liés par des liaisons peptidiques, pour l'ARN et l'ADN il s'agit de nucléotides liés par des liaisons phosphodiester.

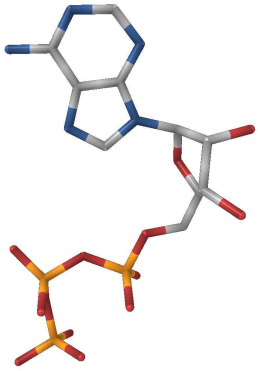
Bien qu'il n'y ait généralement pas d'interaction covalente entre des résidus non-consécutifs il existe d'autres interactions chimiques plus faibles qui déterminent la géométrie globale de la molécule. Certaines sous-séquences ont ainsi des propriétés chimiques propres qui confèrent notamment des caractéristiques géométriques locales. La donnée de la suite de ces sous-séquences est appelée *structure secondaire*.

1. http://goldbook.iupac.org/M03667.html

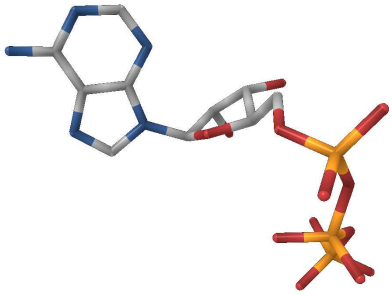
Adénosine triphosphate (ATP)




Représentation des liaisons covalentes



Première conformation



Seconde conformation

FIGURE 1 – Représentation de deux conformères de l'adénosine triphosphate (ATP). Les conformères sont issus des fichiers 1xdn et a4ff de la PDB, une base de données de molécules biologiques observées expérimentalement.

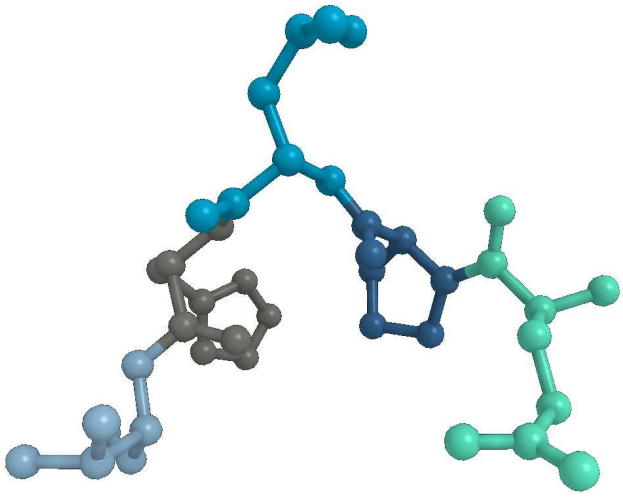



FIGURE 2 – Structure de chaîne des macromolécules biologiques. Ce sont 5 acides aminés extraits de la protéine *L-serine dehydratase* représentée dans le fichier 1p5j de la PDB, une base de données de molécules biologiques observées expérimentalement.

Représentations d'une partie de la protéine *L-serine dehydratase*

Structure primaire : T T P A L T I E R L K N E G A T C K V V G E L L D E A F E L A K A L A K N N P G

Structure secondaire : boucle hélice $\alpha$ boucle feuillet $\beta$ boucle hélice $\alpha$ boucle

Structure tertiaire : Représentation schématique à gauche, et complète à droite.

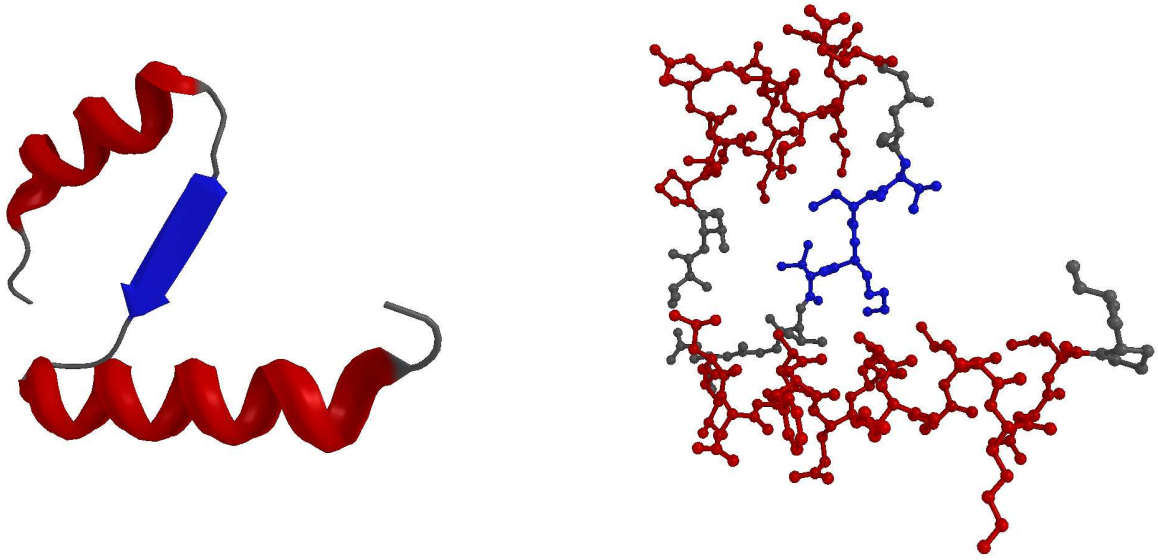


FIGURE 3 – Différents niveaux de structure de la protéine *L-serine dehydratase* (résidus 90 à 129, représentés dans le fichier 1p5j de la PDB). En haut, la séquence est modélisée par une suite de lettres, chacune représentant un résidu suivant le codage standard des acides aminés. Ensuite, la structure secondaire correspond à la suite des sous-séquences de résidus, les sous-séquences étant les hélices $\alpha$ et les feuillets $\beta$. Enfin la représentation tridimensionnelle à gauche correspond à un modèle appelé « cartoon » des sous-séquences de la structure secondaire plongées dans l'espace, et à droite la représentation complète décrit la position de chaque atome et les liaisons covalentes.

Enfin, la *structure tertiaire* correspond à la conformation de la molécule, c'est-à-dire à l'arrangement spatial de chacun des atomes qui la composent. En particulier la donnée de la nature et la position de chaque atome d'une macromolécule biologique permet de retrouver automatiquement les structures primaire et secondaire. La réciproque étant généralement fausse (voir par exemple une revue sur les méthodes de prédiction du repliement des protéines [Dill 2012]).

Il existe également une notion de *structure quaternaire* des macromolécules biologiques. Elle correspond à des groupements de plusieurs macromolécules liées entre elles par des liaisons chimiques non-covalentes. Ce groupement, appelé *assemblage* possède alors une fonction biologique propre. Dans le contexte de notre étude ces assemblages seront considérées comme une seule macromolécule, car nos modèles peuvent être appliqués de la même manière.

## 1.2 Interactions entre molécules et complexes ligand-cible

### 1.2.1 Propriétés chimiques et fonctions biologiques

Le dogme central de la biologie moléculaire énoncé par Francis Crick dès 1958 et précisé en 1970 [Crick 1958, Crick 1970] avait comme objectif de décrire les transferts d'information entre les différentes macromolécules. En particulier, l'ADN est vu comme le support de l'information

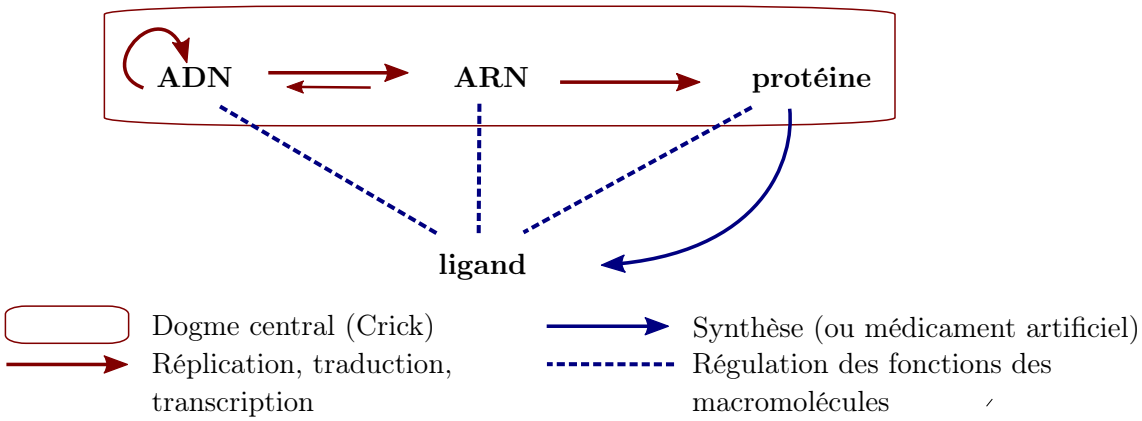



FIGURE 4 – Rôle des petites molécules dans le cadre du dogme central régissant les macromolécules. Adapté à partir de la figure 1 de [Schreiber 2005].

génétique codé par la séquence des nucléotides, les protéines assurent les fonctions biologiques de la cellule, et l'ARN a principalement pour rôle de transférer l'information de l'ADN afin de produire les protéines, et parfois une fonction biologique propre au même titre que les protéines.

Ce principe a depuis été partiellement corrigé pour tenir compte de nouvelles découvertes notamment concernant les fonctions de l'ARN, mais reste suffisant pour décrire la problématique de la prédiction d'interactions. Ce sont donc essentiellement l'ARN et les protéines qui assurent les fonctions biologiques de la cellule. Pour ce faire ces macromolécules interagissent entre elles ou avec d'autres molécules selon leurs propriétés géométriques et chimiques propres.

Les fonctions des protéines sont très variées, par exemple en assurant l'architecture d'une partie de la cellule, le transport de différentes substances, ou la transmission d'un signal. De nombreuses protéines ont un rôle enzymatique, c'est-à-dire qu'elle permettent à une réaction chimique de s'effectuer, sans en être un substrat ni un produit mais en diminuant l'énergie nécessaire pour que l'interaction s'effectue. La description des différents mécanismes sort du cadre de notre étude ; on retient simplement que les différentes fonctions biologiques des macromolécules sont déterminées par leur structure tertiaire.

La principale extension qu'on peut faire au dogme central de la biologie moléculaire qui soit nécessaire à la compréhension des mécanismes cellulaires est le rôle des petites molécules [Schreiber 2005], présenté en figure 4. En effet des interactions, généralement non covalentes, peuvent avoir lieu entre des petites molécules et les macromolécules, ce qui peut modifier la fonction biologique des macromolécules. Ces petites molécules peuvent être endogènes, c'est-à-dire produites par l'organisme, ou exogènes, issues du milieu extérieur et éventuellement artificielles comme dans le cas de certains médicaments. Ces petites molécules ont ainsi souvent un rôle de régulation sur la fonction des macromolécules en activant ou désactivant certaines fonctions.

### 1.2.2 Complexe ligand-cible et pharmacologie

Un *complexe ligand-cible* est défini comme l'interaction d'une petite molécule appelée *ligand* avec une macromolécule, appelée *cible*. Ces interactions entre des ligands et leurs cibles interviennent naturellement dans les cellules en permettant notamment la régulation des fonctions des macromolécules. La pharmacologie vise à étudier ces interactions afin d'agir sur ces mécanismes en introduisant de nouveaux ligands dans l'organisme.

La capacité d'un ligand à se lier à une cible a dans un premier temps été décrite par Emil Fischer en suivant le principe *clé-serrure* [Fischer 1894]. Il suppose que l'interaction est entière-

ment déterminée par la forme et la nature chimique d'une région (serrure) de la cible, qui doivent être complémentaires à celles du ligand (clé). Dans ce cadre, il serait possible d'étudier indépendamment le ligand et la cible afin d'en déterminer la compatibilité ; cependant des découvertes ultérieures ont permis de définir d'autres principes décrivant mieux les mécanismes réels, comme l'*ajustement induit* [Koshland 1995] qui explique la capacité d'une cible à s'adapter en fonction de chaque ligand.

Les mécanismes physiques et chimiques en jeu, s'ils peuvent être modélisés, restent très difficiles à simuler et donc à prédire. De plus la réponse à la question de l'interaction n'est pas dichotomique : la force de l'interaction ou l'*affinité du ligand avec la cible* est variable. La valeur de cette affinité est souvent décisive car généralement plusieurs ligands peuvent être en compétition et c'est leurs affinités respectives qui permettront de décider quelles interactions auront effectivement lieu.

La région de la macromolécule dont les atomes sont en interaction avec le ligand est appelée *site de liaison*. La notion de *mode de liaison* fait référence à la position du ligand et l'ensemble des liaisons chimiques intermoléculaires qui ont lieu avec le site de liaison.

# 2 Prédiction *in silico* d'interaction

La problématique de la prédiction d'interaction est très générale, et notre étude est centrée sur le problème défini ici de la prédiction de cibles pour un ligand. Nous présentons également les principaux enjeux pour la biologie notamment en pharmacologie, et les données accessibles qui permettent d'utiliser des approches *in silico* (informatiques).

## 2.1 Détermination de toutes les cibles d'un ligand

### 2.1.1 Prédiction *in silico* de l'affinité

L'objectif qui est suivi est de déterminer pour l'ensemble des macromolécules présentes dans une cellule l'affinité d'un ligand, qui sera référencé comme le *ligand requête*, avec chacune des macromolécules.

Différentes techniques expérimentales *in vitro* existent, mais sont généralement longues et coûteuses ce qui les rend non praticables pour un grand nombre de cibles. De plus l'effet observé expérimentalement n'est pas toujours un effet direct du ligand d'intérêt sur une cible donnée mais parfois un effet indirect difficile à préciser.

On suppose qu'on dispose des informations chimiques et géométriques sur ce ligand, en particulier son ou ses conformères les plus stables. On suppose également qu'on dispose des informations sur la structure tridimensionnelle des macromolécules. On se propose alors de fournir un algorithme permettant de prédire les affinités entre le ligand et chaque macromolécule, en particulier pour en déterminer le sous-ensemble susceptible d'interagir avec le ligand.

### 2.1.2 Sélection des cibles potentielles

*Étant donnée une petite molécule dont la structure est fournie, étant donnés un ensemble de macromolécules dont les structures sont également fournies, quel est le sous-ensemble de ces macromolécules qui peuvent former un complexe avec la petite molécule ?*

Cette formulation permet de voir le problème de la prédiction de cibles comme un problème de classification binaire : il s'agit de sélectionner un sous-ensemble des cibles, prédites pour interagir avec le ligand. La plupart des algorithmes fournissent une classification en fonction d'un

paramètre seuil, typiquement sous la forme d'un classement ou d'une valeur de score pour chaque macromolécule : un rang ou une valeur limite servant de délimitation entre les deux groupes.

## 2.2 Enjeux pour la biologie et la pharmacologie

La chémogénomique vise à comprendre l'effet des petites molécules au sein d'un système biologique complet [Bender 2007, Bredel 2004]. Pour cela il est nécessaire de comprendre le fonctionnement complet d'une cellule de l'information génétique au phénotype, ainsi que l'ensemble des interactions ayant lieu entre les différentes molécules, afin de comprendre l'action d'un ligand dans le réseau global.

Par exemple, un médicament peut avoir un ou plusieurs effets thérapeutiques, ainsi que des effets dits secondaires. L'effet thérapeutique constitue l'action souhaitée du médicament, alors que les effets secondaires sont l'ensemble des autres actions du médicament. Au niveau moléculaire cette dualité peut s'expliquer par l'affinité du médicament ligand d'une part avec sa ou ses cibles dites thérapeutiques, et d'autre part avec d'autres cibles secondaires connues ou non. Ces interactions avec les cibles secondaires peuvent avoir des effets très variés, parfois bénéfiques ou neutres et parfois plus graves ou même mortels. Ce sont ces derniers effets indésirables qui doivent être mieux compris, et que l'on souhaite éviter.

Le paradigme *un-médicament-une-cible* a notamment été utilisé pour adresser la problématique des effets secondaires. Il s'agit, à partir de la connaissance d'une cible thérapeutique connue pour son action dans le mécanisme cellulaire visé, de proposer un ligand ayant une affinité suffisamment grande avec cette cible. L'espoir étant que l'affinité du ligand avec toute autre cible serait inférieure, évitant ainsi la possibilité d'effets secondaires. Différentes études montrent cependant que ce paradigme ne permet pas en général d'arriver à une sélectivité suffisante, ce qui se traduit soit par une efficacité réduite soit par une toxicité accrue selon que les cibles secondaires ont un effet neutre ou négatif [Hopkins 2008].

Un nouveau paradigme, la *polypharmacologie* [Lavecchia 2015], part du constat que la plupart des médicaments ont plusieurs cibles [Paolini 2006]. L'objectif est alors de comprendre l'ensemble des cibles secondaires d'un ligand dans le cadre d'une étude plus globale du système biologique qui est visé, au lieu de ne considérer qu'une seule cible. Les cibles secondaires peuvent avoir un effet toxique, mais elles peuvent aussi avoir un effet thérapeutique pour la même indication ou éventuellement pour une autre pathologie que celle étudiée initialement. Cette dernière technique consistant à utiliser une molécule conçue dans un premier objectif pour une nouvelle thérapie est appelée *repositionnement* [Lavecchia 2015].

## 2.3 Données structurales et fonctionnelles accessibles

### 2.3.1 Modèle tridimensionnel des molécules

Les données structurales sont bien évidemment essentielles à l'utilisation d'approches basées sur la structure des macromolécules. On cite en particulier une base de données publique très riche, la PDB (en anglais *Protein Data Bank*, Banque de Données de Protéines) Il s'agit de la plus grande base de données publique de structures tridimensionnelles de molécules biologiques, déterminées expérimentalement. Chaque fichier de la PDB contient la représentation tridimensionnelle d'un ou plusieurs conformères d'une ou plusieurs molécules. En 2016, la PDB propose plus d'une centaine de milliers de fichiers structuraux représentant majoritairement des protéines et des complexes ligand-protéine, mais également des ARN et ADN. Le niveau de couverture du protéome de cette

base et sa croissance rapide permettent d'envisager des méthodes informatiques suffisamment exhaustives.

Chaque fichier structural, qu'il soit issu de la PDB ou d'une autre base publique ou privée, est annoté par des informations sur la fiabilité et la qualité des données reportées qui dépendent de la méthode de construction de la représentation tridimensionnelle. Les molécules résolues par diffraction des rayons X sont entre autres annotées par une valeur dite de résolution. Les molécules résolues par résonance magnétique nucléaire sont plus difficiles à exploiter car cette technique propose plusieurs modèles de la même molécule. Enfin d'autres méthodes *in silico* permettent parfois de prédire la structure tertiaire, notamment par homologie en utilisant une similarité avec d'autres macromolécules dont la structure primaire est proche, comme SWISS MODEL [Biasini 2014].

Quelle que soit la méthode expérimentale ou *in silico*, les fichiers structuraux permettent essentiellement d'avoir accès à la représentation tridimensionnelle et ainsi à la structure tertiaire de macromolécules, et de complexes ligand-cible. En particulier la position et la nature de chaque atome est fournie. En prenant l'exemple de la PDB, on remarque que certaines macromolécules sont représentées plusieurs fois, souvent dans des conformations différentes. Cette *redondance* peut être très importante : environ 20 000 protéines différentes sont représentées dans la PDB alors que plus de 250 000 conformations sont proposées. Certaines protéines, plus souvent étudiées sont sur-représentées dans la PDB (éventuellement dans des conformations distinctes), alors que pour un grand nombre de protéines il n'existe aucune représentation tridimensionnelle connue.

### 2.3.2 Bases de données biologiques

Les données structurales sont généralement croisées avec d'autres types de données complémentaires. On cite notamment concernant les protéines, la base de données Uniprot [Apweiler 2004] qui a pour objectif de centraliser les informations et de classer l'ensemble des protéines connues. La séquence de chaque protéine est stockée de même que ses fonctions connues et les liens avec d'autres sources d'informations sur la protéine. En particulier, lorsque qu'une ou plusieurs structures de la PDB sont disponibles, elles sont référencées.

L'utilisation des bases de données connexes, comme DrugBank [Wishart 2008] centralisant des information sur les actions connues de nombreux ligands sur leurs cibles, est primordiale pour intégrer les approches de prédictions structurales dans une étude d'un système biologique. Les différentes applications pouvant être faites d'une prédiction *in silico* de cibles par croisement avec d'autres sources de données sont très variées et ne peuvent être détaillées ici [Bender 2007, Keiser 2009, Koutsoukas 2011, Lounkine 2012] .

## 3 Approche par *docking inverse*

L'approche par *docking* (ou par amarrage) consiste à prédire directement la capacité de deux molécules à interagir, par une estimation de l'affinité. Les logiciels de *docking* sont traditionnellement développés pour le problème d'optimisation de ligands sur une seule cible fixée, dont le site de liaison est également déterminé. L'application de ces outils au problème complémentaire de la prédiction de cibles est appelée *docking inverse*.

## 3.1 Présentation du problème de *docking*

### 3.1.1 Problème d'optimisation

Le problème de *docking* entre deux molécules peut se définir de manière générale comme la recherche du complexe optimal. La notion d'optimalité est définie comme la minimisation d'une fonction évaluant l'énergie nécessaire à la formation du complexe, la recherche pouvant donc se voir comme un problème d'optimisation de cette fonction sur l'espace des conformations possibles.

On considère ici le problème de *docking* entre une petite molécule et une protéine, d'autres méthodes étant utilisées pour le *docking* entre macromolécules. Afin de respecter le vocabulaire généralement utilisé dans le contexte du docking, la petite molécule sera toujours appelée ligand, en revanche la macromolécule sera appelée *récepteur*. L'étude complète des algorithmes de *docking* sort du cadre de la thèse ; on présente ici le fonctionnement général ainsi que les principaux algorithmes utilisés. Une revue des approches les plus utilisées est proposée dans [Sousa 2013].

### 3.1.2 Espace de recherche et flexibilité

L'espace de recherche du problème de *docking* est l'ensemble des conformations du ligand, du récepteur et leurs positions relatives. Une première distinction peut ici être faite concernant la gestion ou non de la *flexibilité* du ligand ou de la cible. On dit qu'un logiciel de *docking* est *ligand-flexible* si dans la recherche du complexe la conformation du ligand peut être modifiée suivant des liaisons covalentes de la molécule où un axe de rotation est permis. De manière analogue le *docking* est dit *récepteur-flexible* si la conformation du récepteur peut également être modifiée. Dans ce cas le degré de liberté est généralement plus faible que pour le ligand, la flexibilité étant locale aux chaînes latérales des résidus sans que le squelette carboné assurant la structure entre les résidus ne soit modifié.

Une seconde distinction est faite sur la localisation explorée du ligand par rapport au récepteur. Lorsqu'une zone est déterminée *a priori* comme contenant un site de liaison, l'espace de recherche est réduit à cette zone. Dans le cas contraire où le ligand peut être placé à n'importe quel endroit du récepteur, on dit que le *docking* est *aveugle*. C'est en particulier le cas dans l'application au *docking inverse* où on cherche à prédire l'interaction ou non avec un ligand sans connaître la localisation du site de liaison éventuel, alors que dans l'application classique on suppose que l'interaction a lieu sur un site déterminé et on cherche à déterminer le meilleur ligand *pour ce site*. Ainsi le *docking aveugle* revient à poser le problème de la recherche du site en plus du problème de la détermination d'une affinité pour ce site.

Chaque implémentation utilise des représentations différentes, cependant elles sont généralement des variantes du même procédé. La position et l'orientation du récepteur étant fixées, un complexe est modélisé comme la donnée de la position et l'orientation du ligand, et éventuellement des angles de torsion pour les liaisons admettant un axe de rotation lorsque la flexibilité est prise en compte. Ainsi la flexibilité correspond à une augmentation dans le nombre de degrés de liberté dans l'espace de recherche alors que la connaissance ou non de la zone de *docking* correspond à l'amplitude des valeurs pouvant être prises pour la position du ligand.

### 3.1.3 Fonction de score et affinité

La fonction de score joue un rôle essentiel dans tout logiciel de *docking*, car c'est cette fonction qui est minimisée par l'algorithme de recherche. Cette fonction associe à chaque point de l'espace de recherche une valeur qui est typiquement une évaluation de l'énergie nécessaire à la formation

du complexe, cette énergie étant donc négative dès que la formation du complexe est énergétiquement favorable. D'autres fonctions de score existent, par exemple par comptage des interactions entre atomes prédites par la pose, toujours dans l'objectif de pouvoir comparer deux points de l'espace de recherche afin de décider lequel est le plus favorable. Certaines fonctions de scores sont différentiables sur l'espace de recherche, ce qui permet l'utilisation de méthodes de type descente de gradient dans les algorithmes de recherche en phase d'optimisation locale.

La méthode usuelle pour définir une fonction de score consiste à modéliser l'ensemble des interactions ayant lieu pour une conformation donnée du ligand et du récepteur. Il s'agit essentiellement des interactions dites de Van der Waals et électrostatiques, dont les modèles sont appelés *champs de force*. À titre d'exemple, le champ de force CHARMM [Steffen 2010] est utilisé par EADock DSS [Grosdidier 2011], AMBER [Cornell 1995, Wang 2004] est utilisé par DOCK [Ewing 2001] ou AutoDock/Vina [Trott 2010].

### 3.1.4 Algorithmes de recherche

Le rôle de l'algorithme de recherche est de converger sur l'espace de recherche vers le point minimisant la fonction de score décrite précédemment. Aucun algorithme ne permettant de résoudre le problème d'optimisation de manière exacte, différentes heuristiques sont développées. Le procédé consiste généralement en deux étapes utilisant des méthodes différentes : la recherche globale suivie d'une optimisation locale.

Les algorithmes génétiques sont souvent utilisés pour résoudre le problème de la recherche. Parmi les logiciels de *docking* les plus populaires et cités dans les publications scientifiques, AutoDock et son successeur AutoDock Vina [Trott 2010] utilisent une variante dite Lamarckienne d'un algorithme génétique [Fuhrmann 2010]. Une population initiale est générée aléatoirement dans l'espace de recherche, chaque individu correspondant à une conformation du ligand sur le récepteur, paramétré par ses coordonnées positionnelles et de torsion. La phase de sélection consiste à retenir les individus associés aux meilleurs scores qui engendrent une nouvelle génération par des mutations qui sont des sauts dans les différentes coordonnées. La variante Lamarckienne est une méthode d'optimisation applicable sur une fonction de score non différentiable, où une descente de gradient n'est donc pas possible.

Les algorithmes génétiques dépendent d'un grand nombre de paramètres, qui sont optimisés dans chaque implémentation de logiciel de *docking*. Il est généralement difficile de comparer ces algorithmes de recherche, notamment à cause des interdépendances entre l'espace de recherche, la fonction de score, et l'algorithme de recherche [Cole 2005].

## 3.2 Docking inverse et prédiction de cibles

### 3.2.1 Objectif du *docking inverse*

L'usage typique d'un logiciel de *docking* concerne la phase de sélection ou d'optimisation du ligand dans le processus de conception d'un médicament. Dans ce contexte, une cible d'intérêt est choisie et un site de liaison connu. L'objectif est alors de déterminer le meilleur ligand possible pour cibler ce site, ou d'optimiser un ligand en y apportant des modifications mineures susceptibles d'améliorer l'affinité.

Le principe du docking inverse est complémentaire, pour un ligand d'intérêt fixé l'objectif est de déterminer l'ensemble des cibles avec lesquelles il peut interagir. En particulier il s'agit d'un *docking aveugle* si on ne dispose *a priori* pas de l'information du site où pourrait se lier le ligand, alternativement une phase préalable de prédiction des sites potentiels peut être réalisée.

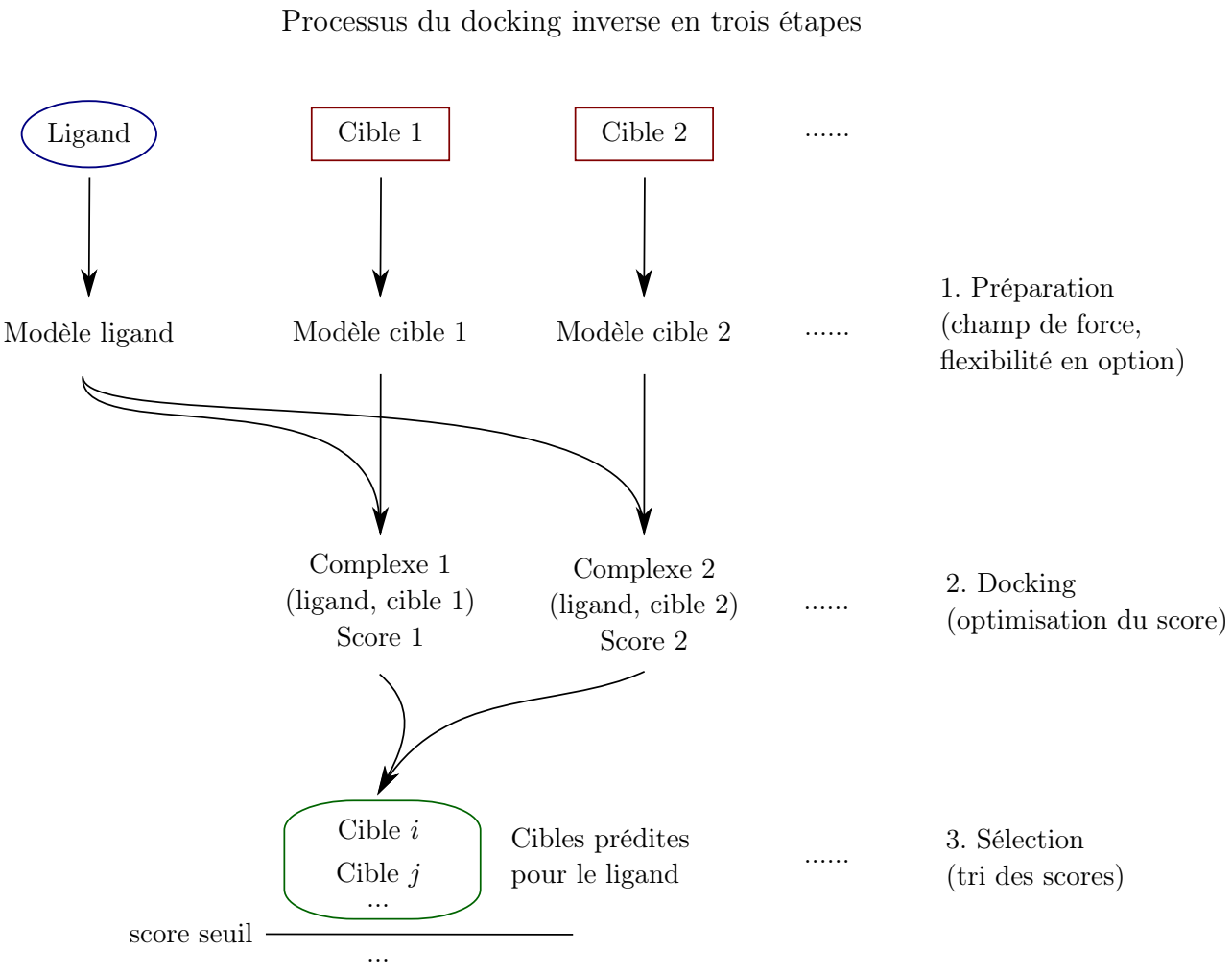



FIGURE 5 – Processus global du docking inverse.

Le processus général est présenté dans la figure 5. Une première phase de préparation consiste à déterminer une représentation adaptée des molécules impliquées. Ensuite le *docking* consiste à déterminer un ensemble de conformations, par optimisation d'un score. Enfin, une sélection des conformations est effectuée sur la base du score attribué à chaque résultat de la phase de *docking*.

### 3.2.2 Fonction de score et variabilité des cibles

La principale difficulté dans l'utilisation des approches de *docking* classiques pour une application de docking inverse semble se situer dans la fonction de score qui est trop optimisée, voire biaisée, pour être une valeur comparable entre ligands différents ou différentes poses d'un même ligand, mais en tout état de cause sur *une seule cible* et plus précisément *le même site de liaison* sur cette cible. Une étude réalisée avec le logiciel Glide [Friesner 2004, Halgren 2004] montre ainsi que les scores obtenus dans des sites de cibles différentes ne sont pas comparables [Wang 2012]. Cette même étude mentionne également une seconde difficulté pouvant apparaître pour une utilisation à grande échelle : contrairement aux jeux de données contrôlées utilisées dans les essais, les structures des macromolécules sont généralement moins fiables et plus difficile à corriger que les petits ligands, les logiciels de *docking* étant très sensibles à ces petites variations. Un dernier point devant être mentionné concerne le coût en terme de temps de calcul, rendant une approche par *docking* trop longue ou chère sur une base de données de cibles de l'ampleur de la PDB. En effet, l'ordre de grandeur pour un *docking aveugle* est d'une heure sur un processeur pour une macromolécule, soit plusieurs dizaines d'années pour plus de 200 000 macromolécules que contient la PDB.

L'approche du docking inverse a cependant été utilisée avec succès dans quelques expériences spécifiques. On cite en particulier la découverte d'une nouvelle cible pour le N6-isopentenyladenosine, réalisée grâce à un docking inverse du ligand sur 296 cibles à l'aide d'AutoDock [Scrima 2014]. La contrainte du temps de calcul a été adressée par un espace de cibles à explorer réduit, et la contrainte de comparabilité des scores a été contournée par une normalisation de chaque score par la moyenne d'un ensemble de 50 ligands sur chaque cible.

# 4 Approches par recherche de similarités

Les approches par recherche de similarités utilisent un principe d'inférence d'interaction afin de répondre à la problématique de prédiction de cibles par un autre point de vue. On présente ici le fonctionnement général de telles approches, les techniques algorithmiques classiques seront présentées dans le second chapitre.

## 4.1 Similarité et prédiction d'interaction chimique

### 4.1.1 Principe d'inférence de l'interaction chimique

Les approches par similarités reposent sur un principe suivant lequel *des cibles similaires interagissent avec des ligands similaires* [Klabunde 2007]. Ce principe peut se décliner suivant deux points de vue selon qu'on considère la similarité entre les ligands ou bien la similarité entre les cibles. Plus précisément, si le ligand requête est connu pour interagir avec une cible dite requête et que cette cible requête est similaire à une seconde cible, alors on peut inférer l'interaction entre le ligand requête et cette dernière cible. De manière analogue, si le ligand requête est similaire à

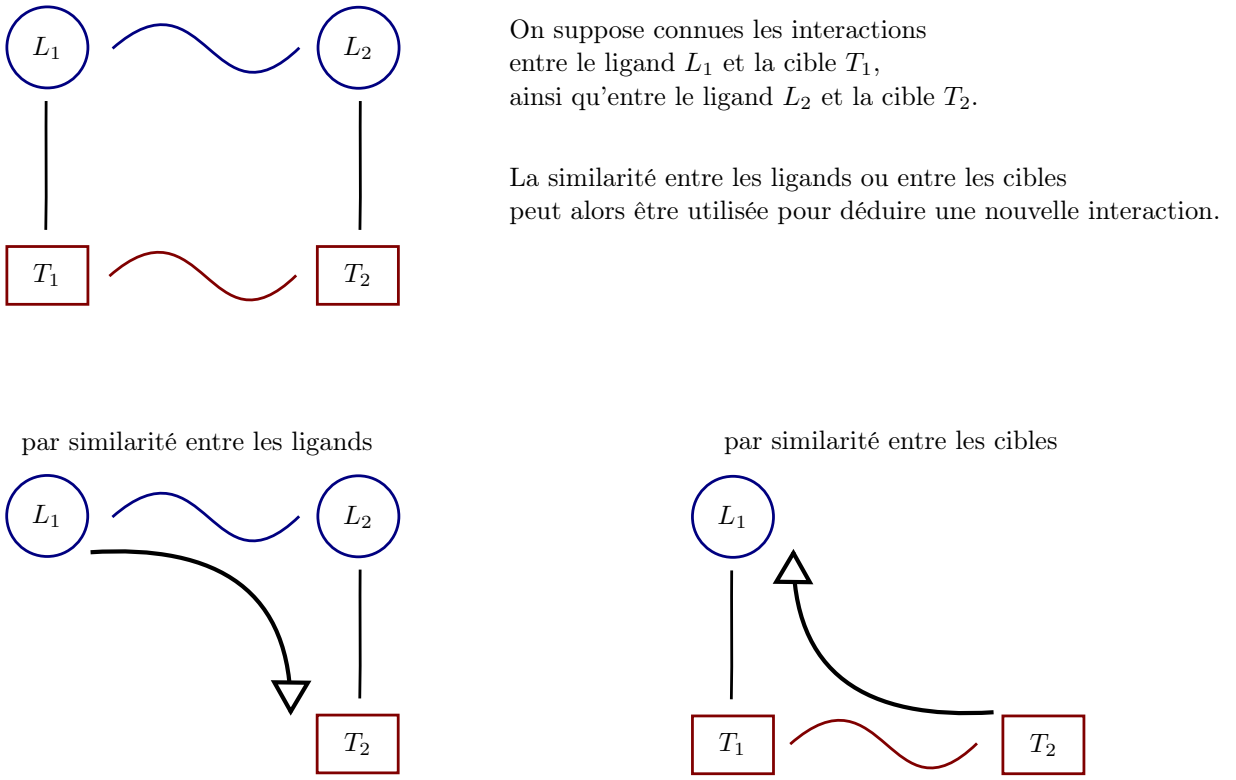



FIGURE 6 – Inférence de l'interaction entre le ligand $L_1$ et la cible $T_2$ par similarité entre ligands ou entre cibles.

un second ligand lui même connu pour interagir avec une cible, alors on peut inférer l'interaction entre le ligand requête et cette cible (figure 6).

L'avantage immédiat de cette approche est qu'elle permet de transformer le problème de prédiction d'interaction en un problème de recherche de similarité. Cela permet d'utiliser par exemple certaines heuristiques dans la recherche d'une géométrie similaire, alors que la géométrie « complémentaire » est plus difficile à définir. En revanche deux limitations intrinsèques à l'approche sont à noter, en particulier par rapport à l'approche de l'évaluation directe d'affinité ou *docking*. Le premier inconvénient concerne la non réciprocité du principe d'inférence de l'interaction, et le second est la dépendance à des données d'interactions connues avant de pouvoir démarrer ces approches.

### 4.1.2 Application à la prédiction de cibles

On considère toujours un ligand requête $L$ dont on recherche les cibles. Le processus consiste à trier un ensemble de cibles dites cibles candidates et notées $C_1 \dots C_n$ suivant un score de similarité. Contrairement à l'approche par *docking* qui évalue directement la *complémentarité* entre le ligand requête $L$ et chaque cible $C_i$, les approches par similarités nécessitent une information supplémentaire soit pour le ligand requête soit pour les cibles candidates. En effet l'approche par similarité des ligands nécessite de connaître pour chaque cible candidate un ligand connu pour évaluer la *similarité* avec le ligand requête. La seconde approche par similarité des cibles nécessite elle de connaître une première cible $C$ du ligand requête pour évaluer la *similarité* avec chaque cible candidate.

| | Docking inverse | Similarité ligands | Similarité cibles |
|---|---|---|---|
| Entrée | ligand $L$ | ligand $L$ | complexe $L-C$ |
| Base de recherche | cibles $C_i$ | complexes $L_i - C_i$ | cibles $C_i$ |
| Mesure évaluée | complémentarité $L/C_i$ | similarité $L \sim L_i$ | similarité $C \sim C_i$ |
| Sortie | classement des cibles | | |

TABLE 1 – Tableau récapitulatif des approches *in silico* de prédiction de cibles.

## 4.2 Évaluation de la similarité

### 4.2.1 Similarité entre deux ligand

On distingue deux grandes classes de représentations des molécules utilisées dans les approches par similarité des ligands. Un premier type de représentation, qu'on appellera *empreinte*, consiste à énumérer un certain nombre de descripteurs sur les molécules, c'est-à-dire à représenter une molécule comme un vecteur de taille fixe ayant des valeurs booléennes. Un second type de représentation qu'on appellera *géométrique* consiste à définir des points de l'espace associés à une ou plusieurs valeurs scalaires ou vectorielles dépendant des atomes environnants.

La génération d'une empreinte pour une molécule consiste à tester la présence ou l'absence de petits groupement d'atomes prédéfinis, appelés motifs, dans une molécule afin de créer un liste de valeurs binaires : pour un ensemble de $n$ motifs, on affecte le vecteur $(a_1, a_2, \ldots, a_n)$ à une molécule, où $a_i = 1$ si le $i^{\text{e}}$ motif est présent dans la molécule et $a_i = 0$ sinon. Le procédé de comparaisons entre deux empreintes est généralement dérivé du score de *Tanimoto* [Bajusz 2015] qui mesure la distance entre deux vecteurs binaires comme le ratio des cardinaux de l'intersection (« et » logique) par l'union (« ou » logique), correspondant au coefficient de similarité de Jaccard sur les ensembles.

$$\begin{aligned}\text{Tanimoto}\left((0,0,1,1,0,0,1,0),(0,1,0,1,1,0,0,0)\right) &= \frac{\#((0,0,1,1,0,0,1) \wedge (0,1,0,1,1,0,0))}{\#((0,0,1,1,0,0,1) \vee (0,1,0,1,1,0,0))} \\ &= \frac{\#(0,0,0,1,0,0,0)}{\#(0,1,1,1,1,0,1)} = \frac{1}{5}\end{aligned}$$

Les représentations géométriques reposent sur un principe analogue de recherche d'un certain nombre de motifs appelés *pharmacophores* annotés par une position et éventuellement une orientation. La méthode de comparaison consiste alors à effectuer un appariement des pharmacophores entre les deux représentations des molécules. La flexibilité du ligand peut être prise en compte, soit en générant un ensemble de conformations pour chaque ligand, soit dans le procédé d'appariement.

### 4.2.2 Similarité entre deux cibles

L'étude de la similarité des cibles se distingue de la similarité des ligands par la taille importante des macromolécules, c'est-à-dire que la similarité globale entre deux macromolécules n'est pas toujours pertinente et c'est une similarité locale qu'il faut déterminer [Vulpetti 2012]. Plus précisément, on souhaite généralement comparer deux *sites de liaison* et non deux macromolécules. Ce facteur d'échelle introduit une nouvelle distinction dans les approches par similarité selon que l'espace de recherche des cibles candidates est fourni comme un ensemble de sites candidats ou bien si la recherche du site potentiel sur la macromolécule candidate fait partie de l'approche.

Le premier point de vue permet d'utiliser des méthodes de représentations des sites originales et directement optimisées pour la comparaison, analogues aux empreintes. Cela suppose en revanche de disposer d'une méthode permettant pour chaque cible candidate de déterminer l'ensemble des sites potentiels qui

seront comparés au site requête, ce qui impose un premier biais dans l'espace de recherche des sites pouvant être prédits. De plus la forme des sites pouvant être modélisés est généralement contrainte afin d'être adaptée à la méthode de comparaison.

Le second point de vue part également du site requête défini par l'interaction entre le ligand requête et sa cible requête, mais permet de faire une recherche sur l'ensemble des macromolécules candidates, permettant notamment de prédire des sites sur une région qui n'aurait pas pu être déterminée *a priori*. Ce point de vue impose de pouvoir représenter une macromolécule entière, et non pas seulement un site de liaison, ce dernier étant alors simplement une partie de cette représentation restreinte à une région donnée. Les représentations géométriques des macromolécules sont variées. Elles peuvent notamment se faire au niveau de granularité de l'atome, du résidu, ou encore d'un descripteur analogue aux pharmacophores décrits pour les ligands. Les représentations peuvent également se distinguer selon qu'elles modélisent l'ensemble d'une macromolécule ou bien seulement la surface, constituée des atomes en contact avec le milieu extérieur qui sont susceptibles de participer à une interaction.

# 5 Algorithme BioBind de recherche de similarités

L'algorithme, conçu et implémenté dans le cadre de cette thèse, est nommé BioBind. Il s'agit d'une approche par similarité des cibles, dont le développement s'inscrit dans le contexte et les contraintes industrielles de l'entreprise finançant la thèse.

## 5.1 Méthode pré-existante insatisfaisante pour l'application industrielle

### 5.1.1 Contexte industriel

Le développement de l'algorithme BioBind s'inscrit dans le cadre de l'offre commerciale de l'entreprise Bionext SA qui propose une plate-forme BioSight regroupant différents logiciels et services pour faciliter la recherche en biologie moléculaire et pharmacologie, à destination d'un public académique et industriel. L'entreprise propose un service, appelé Target-Analysis, de prédiction de cibles. L'utilisateur fournit un complexe entre un ligand et une cible, et récupère une liste de cibles prédites issues de la PDB ou d'une autre base de données publique ou privée annotées et classées suivant différentes bases de données structurales et fonctionnelles. Il s'agit donc d'une approche par similarité des cibles, proposée en SaaS [2] au travers du visualiseur de molécule BioViz. L'objet de la thèse consistant à développer un nouvel algorithme de comparaison moléculaire pouvant être intégré, dans un premier temps, au service Target-Analysis.

### 5.1.2 Algorithme existant non satisfaisant

L'algorithme existant, LightSM, est une implémentation d'une approche par similarité des cibles basée sur une représentation de la surface des molécules à l'échelle de l'atome, et permettant une recherche de site de liaison sur la totalité de la surface mais avec une contrainte de forme sur le site de liaison.

La représentation des macromolécules est basée sur les formes alpha [Edelsbrunner 1994], pour produire un graphe dont les sommets sont les atomes de surface et les arêtes sont induites par le polytope de la forme alpha. Chaque atome est modélisé par sa position, une normale, et des propriétés scalaires.

Un site de liaison, que ce soit le site requête ou un des sites candidats, est modélisé comme un sous-graphe constitué par sélection des atomes à une distance inférieure à un seuil fixé d'un atome central, en considérant la distance le long des arêtes du graphe de la macromolécule. Cela permet de définir pour une molécule généralement autant de tels sous-graphe qu'il y a d'atomes en surface. Le site de liaison requête est construit comme le sous-graphe maximisant le nombre d'atomes en interaction parmi l'ensemble des sous-graphes d'une molécule, la propriété d'être en interaction étant définie par un seuil de distance entre

2. Software as a Service : Logiciel comme un service.

l'atome et un des atomes du ligand. L'espace de recherche des sites candidats dont la similarité doit être évaluée est construit de manière exhaustive comme l'ensemble des sous-graphes de l'ensemble des macromolécules candidates.

Le procédé de comparaison se fait en deux étapes. Une première étape de filtrage est basée sur une empreinte, une valeur scalaire définie sur chaque sous-graphe. Une seconde étape consiste en un problème d'optimisation : il s'agit de rechercher la superposition des deux sites maximisant un score défini par couplage des atomes.

Différents paramètres peuvent être adaptés concernant la taille des sous-graphes, le filtrage, et la comparaison. Cependant malgré plusieurs années de travail d'optimisation de l'algorithme, le pouvoir prédictif de l'algorithme dans l'application à la prédiction de cibles s'est avéré insuffisant.

## 5.2 Nouvel algorithme BioBind

### 5.2.1 Cahier des charges

La direction de la recherche dans l'entreprise, encadrant la thèse, était initialement orientée sur une analyse et une amélioration de plusieurs aspects précis de l'algorithme, avec un champ d'action possible restreint : la méthode de filtrage, la méthode d'alignement, les paramètres de la fonction de score. Après que certaines optimisations aient été développées et validées, le résultat global en terme de prédiction de cibles s'est avéré toujours insuffisant. Un changement dans la direction de la recherche a été entrepris, au début de la seconde année de thèse, et m'a permis d'élargir le cadre des travaux de recherches dans l'objectif de proposer une approche entièrement nouvelle, permettant d'espérer une amélioration sensible des résultats.

Les contraintes industrielles étaient essentiellement les suivantes : pour des raisons commerciales, l'algorithme qui allait être développé devait être implémenté et intégré dans la plate-forme de l'entreprise dans un délai relativement bref ; pour des raisons de propriété intellectuelle, aucune librairie informatique présente dans la base de code de l'entreprise autre que celle concernant la génération des formes alpha ne pouvait être réutilisée.

### 5.2.2 Développements réalisés

L'algorithme de recherche d'un site de liaison requête sur une macromolécule candidate a été entièrement repensé, tout en réutilisant certaines heuristiques qui étaient déjà présentes et optimisations qui avaient été réalisées.

Le modèle des macromolécules a été revu. Restant basé sur les formes alpha, le modèle du graphe simple a été remplacé par une variété orientée union de triangles modélisant la surface des macromolécules.

La notion de sous-graphe circulaire à la surface a été conservé, mais un niveau d'information supplémentaire caractérise un site de liaison (requête ou prédit). Cette notion de sous-graphe circulaire a permis notamment de réutiliser des optimisations réalisées dans l'heuristique d'alignement des sous-graphes, et les nouvelles méthodes de filtrage.

L'intégralité de l'algorithme a été implémenté dans le cadre de la thèse, ainsi que diverses interfaces permettant de tester l'approche sur la PDB ou d'autres jeux de données de test.

Chapitre II

# Similarité entre macromolécules pour la prédiction de cibles

## Introduction

Le développement d'une méthode de recherche de similarités entre macromolécules est motivé par la prédiction des interactions qu'il est possible d'en déduire. La notion de similarité, très intuitive, reste difficile à définir de manière précise. En effet la mesure de la similarité est dépendante du modèle de représentation choisi des molécules, des techniques algorithmiques employées, et des motivations pharmacologiques ou biologiques. Il existe de très nombreuses approches de recherche de similarités entre macromolécules appliquées à la prédiction d'interactions et cette grande variété dans les modèles et techniques employées rend l'évaluation et la comparaison entre les méthodes complexe. Il n'existe en effet aucun processus de validation faisant consensus, et permettant d'évaluer une approche donnée par rapport aux approches concurrentes de manière systématique.

Une première section de ce chapitre spécifie le problème informatique qui fait l'objet de notre étude, à partir des problématiques biologiques. Une seconde section propose une classification des différentes méthodologies permettant de réaliser la recherche d'une similarité locale entre un site de liaison et une macromolécule, en précisant les problèmes algorithmiques qui en découlent. Ensuite une troisième section présente certains algorithmes, afin notamment de décrire les spécificités liées à leur application pour la recherche de similarités entre macromolécules. Enfin une dernière section décrit les différents choix effectués pour le développement de BioBind, éclairés par une discussion sur les avantages et inconvénients des précédentes méthodes.

## Sommaire

# 1 Mesure de la similarité et prédiction d'interaction

La recherche de similarités entre macromolécules possède de nombreuses applications basées sur l'inférence d'une interaction chimique. Notre étude est centrée sur le problème de la prédiction de cibles secondaires qui constitue le contexte pour définir le problème de la recherche de similarité. Le choix du modèle des structures moléculaires considérées est d'une importance particulière, tout d'abord parce qu'il permet de formaliser la notion de similarité mais également car il guide ou semble parfois guidé par les algorithmes de recherche utilisés. Par ailleurs, de par la variété des problématiques biologiques et des méthodes algorithmiques, la validation et la comparaison des approches constitue un sujet d'étude important *per se*.

## 1.1 Inférence de l'interaction

Le problème de la recherche de similarité entre molécules est motivé par la prédiction d'une interaction pouvant être déduite à partir du principe d'inférence de l'interaction : « *similar receptors bind similar ligands* » [Klabunde 2007].

L'objet de notre étude est le problème de la recherche de similarité locale entre un site de liaison connu et un site de liaison prédit. Cette similarité est l'outil pour répondre à la problématique pharmacologique de la prédiction de cibles secondaires d'une petite molécule ligand. Il existe d'autres applications biologiques ou pharmacologiques qui peuvent se traduire par un problème semblable de recherche de similarité entre macromolécules, et certaines méthodes peuvent être utilisées pour répondre à différentes problématiques biologiques. Cependant chaque problématique induit des choix souvent spécifiques dans les modèles et algorithmes, et chaque approche est généralement spécialisée pour répondre à une problématique biologique précise. Ces problématiques, et leurs liens avec la similarité locale sont schématisées dans la figure 1.

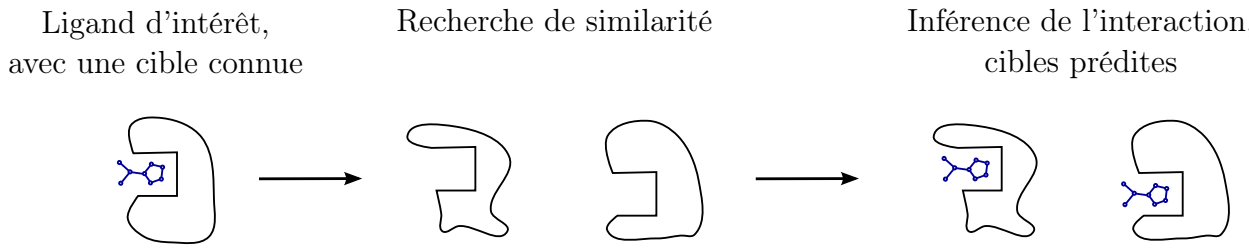
### 1.1.1 Prédiction de cibles secondaires

La prédiction de cibles secondaires [Lavecchia 2015] consiste à prédire une interaction entre une petite molécule ligand et une macromolécule cible, à partir de la donnée d'un complexe impliquant ce ligand avec une autre macromolécule similaire à la cible prédite. Cette problématique constitue la motivation initiale de l'algorithme BioBind développé dans le cadre de cette thèse, c'est ainsi l'objectif à partir duquel les approches seront évaluées et comparées. L'interaction, connue ou prédite, a lieu sur un site de liaison localisé sur une macromolécule, c'est donc plus précisément une similarité locale qui est recherchée entre le site de liaison de la cible connue et une région de la cible prédite. Les différentes méthodes traitant cette problématique par la recherche de similarité sont présentées dans la section 2.

### 1.1.2 Annotation fonctionnelle des macromolécules

L'annotation fonctionnelle [Watson 2005, Adams 2007] peut être vue comme un problème de prédiction d'interaction où l'on cherche notamment à déterminer l'ensemble des ligands pouvant se lier à une macromolécule donnée. Le même principe d'inférence de l'interaction peut être utilisé, afin de prédire un ligand via une similarité locale avec un site de liaison connu de ce ligand sur une autre macromolécule. Si le procédé de recherche d'un motif similaire est semblable, il ne s'agit cependant pas d'un problème équivalent à la prédiction de cibles secondaires. Une différence essentielle se situe dans la mesure de la similarité, qui doit prendre en compte des sites de tailles, formes, et compositions chimiques différentes pour proposer un classement dans cet ensemble hétérogène.

**Prédiction de cibles secondaires.** Il s'agit de prédire de nouvelles cibles pour un ligand d'intérêt, à partir d'une première cible connue.




**Annotation fonctionnelle.** Il s'agit de prédire de nouveaux ligands pour une macomolécule d'intérêt, à partir des autres cibles connues de ces ligands.

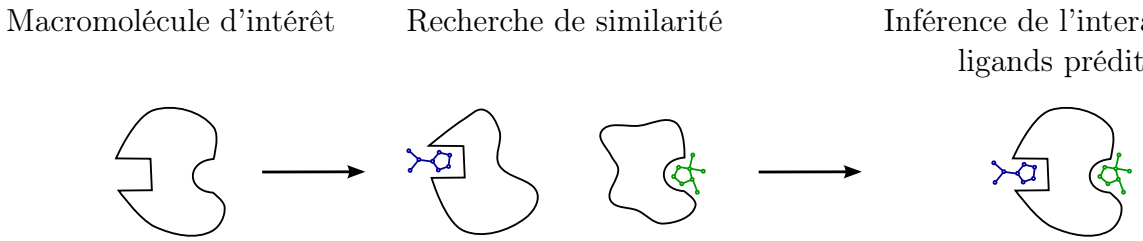



FIGURE 1 – Similarité et inférence de l'interaction appliquée à différentes problématiques biologiques.

### 1.1.3 Interactions chimique entre macromolécules

Les interactions entre macromolécules, par exemple entre protéines, définissent des assemblages qui peuvent avoir une fonction biologique différente des macromolécules considérées séparément. La prédiction des interactions entre macromolécules revient alors à prédire l'existence de certains assemblages. La principale différence avec le problème de la prédiction de complexes ligand-cible concerne la géométrie des sites de liaison, qui sont souvent plus grands et plats et parfois séparés en plusieurs composantes déconnectées [Janin 1990, Moreira 2007]. Ainsi, même s'il est parfois possible d'adapter les méthodes conçues pour la prédiction de cibles de petites molécules, le problème s'avère en général différent en pratique.

## 1.2 Définir la similarité

Le principe d'inférence de l'interaction est un concept naturel, et communément accepté. Il repose sur la notion de similarité entre molécules, une notion qui semble tout aussi naturelle, mais pose en pratique la problématique importante de la définition de la similarité.

### 1.2.1 Similarité et représentation

La similarité est définie sur une représentation des molécules. Le choix de la représentation est central dans le développement d'une approche de recherche de similarité pour plusieurs raisons. D'une part la représentation choisie doit être efficacement manipulable par les algorithmes de recherche, et d'autre part cette représentation contraint la définition qu'il devient possible de donner à la similarité.

La représentation d'une molécule peut se caractériser par son niveau de granularité, par exemple en représentant chaque atome ou bien en ne représentant que les résidus. Il est également possible de discriminer les représentations par leurs degrés d'abstraction par rapport à la structure chimique atomique, en considérant par exemple des groupes fonctionnels d'atomes ou bien une structure totalement abstraite sur laquelle les propriétés chimiques de l'ensemble du site doivent être projetées.

### 1.2.2 Flexibilité des macromolécules

La flexibilité des macromolécules, c'est-à-dire leur capacité à posséder plusieurs structures tertiaires selon leur environnement, peut être prise en compte dans la méthode d'évaluation de la similarité, typiquement en autorisant certaines variations dans les positions des atomes qui pénalisent peu l'évaluation totale de la similarité. En effet, si le modèle des macromolécules est basé sur la structure tertiaire, la similarité sera *a priori* évaluée de manière différente entre un site de liaison requête et deux structures tertiaires distinctes d'une même macromolécule. Ce phénomène apparaît par exemple souvent lorsque deux structures expérimentales sont disponibles pour une macromolécule : l'une avec le ligand présent dans le site, et l'autre sans le ligand ou avec un autre ligand. L'ajustement induit de la macromolécule modifie alors significativement la position des atomes au voisinage du ligand [Koshland 1995].

### 1.2.3 Similarité et distance

La mesure d'une similarité entre un site de liaison requête et une région d'une macromolécule est évaluée par une fonction généralement appelée « fonction de score ». Son rôle est essentiellement de permettre un classement entre les macromolécules, afin de proposer une liste de cibles prédites. Cependant dans le cas général cette fonction n'induit pas une *distance* entre les différents sites de liaison. En effet il est facile de transformer toute mesure de similarité en une fonction positive, et éventuellement symétrique, mais en général l'inégalité triangulaire n'est pas vérifiée. Cette dernière propriété assure essentiellement que si deux objets sont suffisamment proches d'un troisième, alors ils sont également proches entre eux. Autrement dit, si un site requête est très similaire à un premier site candidat, et que ce second site candidat est très similaire à un troisième site candidat, alors le site requête doit être similaire à ce dernier site.

Cette structure sur la mesure de similarité est difficile à garantir. Une mesure de similarité qui posséderait cette propriété aurait pourtant de nombreux avantages, comme la cohérence du score entre des conformations différentes mais suffisamment proches notamment de part la flexibilité des macromolécules, ainsi que pour l'analyse des résultats en permettant par exemple d'effectuer un *clustering* (ou partitionnement) sur l'ensemble des macromolécules et de proposer un représentant pour chaque classe de sites. L'étude proposée dans [Murakami 2013] construit par exemple un groupement hiérarchique des sites à partir d'une comparaison exhaustive réalisée avec le logiciel eF-seek [Kinoshita 2002].

### 1.3 Validations indirectes

La validation des approches de prédiction par similarité des cibles est complexe. En effet le problème de la recherche de similarité est motivé par le problème de la prédiction d'une interaction, le principe d'inférence de l'interaction permettant de considérer la possibilité d'une interaction comme une conséquence de la similarité. La similarité elle même n'a pas de définition canonique permettant une vérification de la correction des approches, et c'est donc uniquement sa conséquence supposée qui peut être vérifiée expérimentalement.

Comparer les différentes approches de recherche de similarité n'est pas plus évident, comme en témoigne l'absence de jeux de données et de métriques faisant consensus, et le faible nombre de publications de méthodes qui contiennent une comparaison aux méthodes antérieures [Nisius 2012]. Il s'agit d'une différence essentielle avec le problème de *docking*, pour lequel des jeux de données fiables comme l'*Astex diverse set* [Hartshorn 2007] ou le *Directory of Useful Decoys* [Mysinger 2012] existent et sont très utilisés, et pour lesquels des métriques canoniques sont définies même s'il est toujours possible d'en améliorer la pertinence [Kirchmair 2008].

## 2 Présentation générale des approches de recherche de similarité

Il existe de très nombreuses approches de recherche de similarité entre macromolécules reposant sur des modélisations des molécules et des définitions de la similarité très variées. Cette variété s'explique par de nombreux choix qui sont réalisés en fonction de la motivation biologique et des données disponibles.

La table 1 liste une sélection de logiciels de recherche de similarités locales. Les paragraphes qui suivent proposent une description des principales méthodes utilisées pour adresser les problématiques de la représentation des molécules, de l'évaluation de la similarité, de l'exploration d'une macromolécule, et de l'alignement entre deux sites.

Les deux premières problématiques concernant la représentation et l'évaluation de la similarité correspondent à la modélisation de la chimie. Les choix réalisés dans ces étapes doivent permettre de transposer la problématique bio-chimique de la similarité moléculaire, en un modèle informatique. Les deux problématiques suivantes d'exploration de la macromolécule et d'alignement correspondent aux méthodes algorithmiques employées pour effectuer la recherche dans le cadre du modèle précédant.

### 2.1 Représentation des molécules

La similarité entre deux molécules, ou plus précisément entre un site de liaison et une partie d'une macromolécule ne possède pas de définition naturelle. En effet la notion de similarité dépend du choix du modèle considéré ; c'est sur ce modèle qu'une mesure peut être définie pour évaluer la similarité.

Il existe essentiellement trois grandes catégories de représentation des molécules. Un premier type de représentation consiste en un ensemble de points dans l'espace tridimensionnel, chaque point représentant un élément réel ou abstrait de la molécule est annoté par un ensemble de caractérisations de ses propriétés

| Logiciel | Référence | Description |
|---|---|---|
| IsoMIF Finder | [Chartier 2015] | Sous motif commun entre points d'une grille modélisant les positions possibles pour les atomes d'un ligand potentiel. |
| FIM | [Wang 2015] | Discrétisation des fragments cotés ligand et cibles, avec matrice de correspondance. |
| APoc | [Gao 2013] | Sous motif commun entre les carbones alpha (orientés vers le beta). |
| BUMBLE | [Kasahara 2010] | Apprentissage des fragments de ligands observés pour chaque fragment de protéine, analyse de la distribution de ces fragments pour une protéine candidate. |
| Patch-Surfer2.0 | [Zhu 2015] | Descripteurs 3D de Zernike définis sur des régions recouvrant un site. |
| ProBiS | [Konc 2010] | Sous motif commun entre des groupes fonctionnels de trois résidus. |
| Hoffmann | [Hoffmann 2010] | Alignement de nuages d'atomes, via une descente de gradient sur un score continu. |
| FuzCav | [Weill 2010] | Empreinte par comptage de motifs prédéfinis, qui sont des triplets de carbones alpha. |
| Yin 2009 | [Yin 2009] | Distributions des courbures sur des régions recouvrant un site. |
| MolLoc | [Angaran 2009] | Méthode de Horn pour l'alignement des surfaces de Connolly. |
| PESD | [Das 2009] | Distribution d'une propriétés sur toutes les paires de sommets. |
| IsoCleft Finder | [Najmanovich 2008] | Sous motif commun entre les carbones alpha, puis entre les atomes. |
| SiteAlign | [Schalon 2008] | Alignement d'une sphère discrétisée où les propriétés des résidus sont projetées. |
| Sommer 2007 | [Sommer 2007] | Comparaison des moments invariants définis sur la distribution du volume des atomes. |
| Morris 2005 | [Morris 2005] | Comparaison des harmoniques sphériques capturant la forme de la surface. |
| SiteEngine | [Shulman-Peleg 2004] | Alignement par hachage géométrique. |
| SuMo | [Jambon 2003] | Sous motif commun entre triplets de groupes fonctionnels. |
| CSC | [Milik 2003] | Sous motif commun entre quadruplets d'atomes. |
| eF-seek | [Kinoshita 2002] | Sous motif commun entre ensembles d'atomes de surface. |
| CavBase | [Schmitt 2002] | Sous motif commun entre groupes fonctionnels. |

TABLE 1 – Logiciels de recherche de similarités entre macromolécules, dans l'ordre chronologique inversé des publications.

chimiques. Un second type de représentation consiste à modéliser la topologie de la surface de la macromolécule sur laquelle les propriétés chimiques sont projetées. Enfin un dernier type de représentation correspond à un descripteur entièrement abstrait modélisant la géométrie et les propriétés chimiques.

### 2.1.1 Ensemble de points

Un premier point de vue naturel pour représenter une molécule est de modéliser chaque atome par une sphère, en associant un certain nombre de propriétés physico-chimiques scalaires. Il s'agit ainsi d'une représentation consistant en un ensemble de points, les centres des atomes, associés à différentes propriétés qui peuvent être géométriques, comme le rayon, ou physico-chimiques comme la charge. L'approche présentée dans [Hoffmann 2010] propose un encodage des sites comme l'ensemble des atomes déterminés par une distance maximale avec un atome d'un ligand, chaque atome étant représenté par sa position et éventuellement un descripteur de charge. IsoCleft Finder [Najmanovich 2008] propose de même une modélisation des sites par un ensemble des atomes, les sites étant préalablement déterminés par un logiciel de détection de cavités. CSC [Milik 2003] et eF-seek [Kinoshita 2002] proposent de modéliser une macromolécule comme l'ensemble des atomes suffisamment proches de la surface pour être potentiellement impliqués dans une interaction.

Ce modèle se généralise naturellement en considérant d'autres « pseudo-centres », par exemple en déterminant un centre pour chaque résidu de la macromolécule. Ce dernier est généralement choisi comme le centre de l'atome du carbone alpha du résidu, c'est à dire le carbone appartement à la chaîne carbonée de la macromolécule. Ce modèle est utilisé par APoc [Gao 2013] et ProBiS [Konc 2010].

Une autre généralisation consiste à définir des groupes fonctionnels, parfois appelés fragments. Ces groupes modélisent les propriétés chimiques d'un ensemble d'atomes, chaque groupe ayant une fonction chimique propre. A partir d'un dictionnaire de ces groupes fonctionnels, une macromolécule est transformée en un ensemble de tels fragments représentés par leur type, leur position, éventuellement une orientation, alors que la composition atomique est généralement ignorée par la suite. Cette méthode se retrouve avec différentes variantes dans le choix du dictionnaire de fragments parmi FIM [Wang 2015], BUMBLE [Kasahara 2010], SuMo [Jambon 2003] ou CavBase [Schmitt 2002]. SiteEngine [Shulman-Peleg 2004] a une approche intermédiaire en conservant l'association entre les atomes et le groupe fonctionnel qui les contient.

### 2.1.2 Modèle de la surface interagissante

Alors que les représentations par des ensembles de points correspondent à la modélisation d'un ensemble d'atomes d'une macromolécule, éventuellement filtrés par leur position vis-à-vis du milieu extérieur, d'autres représentations cherchent à modéliser directement cette interface entre la macromolécule et le milieu extérieur.

IsoMIF Finder [Chartier 2015] propose de modéliser l'ensemble des points du milieu extérieur situés dans un voisinage de la macromolécule, par sélection de sommets sur une grille régulière pavant l'espace. PatchSurfer (resp. PatchSurfer2.0) [Sael 2010] (resp. [Zhu 2015]) utilise une triangulation de la surface par APBS [Baker 2000], et Yin [Yin 2009] une triangulation obtenue via MSMS [Sanner 1996].

### 2.1.3 Descripteur abstrait

Une empreinte est un descripteur représentant un site comme un vecteur de booléens. A partir d'un dictionnaire de motifs, la présence ou l'absence de chacun d'eux étant simplement notée. On appelle également empreinte un vecteur d'entiers, comptant le nombre occurrences de chaque motif, et FuzCav [Weill 2010] propose ainsi une description d'un site définie par 4833 entiers comptant le nombre de chaque motif présent.

Il existe d'autres techniques pour décrire un objet tridimensionnel, qui comme les empreintes sont indépendantes du repère de coordonnées. Sommer [Sommer 2007] utilise la théorie des moments invariants

permettant de décrire une distribution des volumes des atomes. PatchSurfer(2.0) [Zhu 2015] utilise des descripteurs dits de Zernike, et PESD [Das 2009] un encodage des distributions de propriétés dans l'espace, dans le même objectif. Les auteurs de [Yin 2009] proposent un encodage discrétisant la distribution des courbures, alors que la méthode proposée dans [Morris 2005] utilise un encodage par les harmoniques sphériques.

Une représentation originale proposée par SiteAlign [Schalon 2008] modélise un site par une sphère, discrétisée en 80 triangles, et placée en son centre sur laquelle les propriétés des résidus sont projetés. Contrairement aux techniques précédentes, cette méthode d'encodage est sensible à l'orientation de la molécule dans l'espace, mais propose un objet géométrique plus simple à manipuler.

## 2.2 Évaluation de la similarité

Les méthodes d'évaluation de la similarité sont évidemment très dépendantes du modèle choisi pour représenter les molécules. Il existe de nombreuses méthodes, présentées par exemple dans [Kellenberger 2008]. Au cours du processus de recherche, la similarité est évaluée pour une région candidate déterminée par exploration de la macromolécule candidate, et généralement pour une notion d'alignement ou de couplage définie sur le modèle. Ces deux aspects étant traités dans les deux sections suivantes, on considère ici uniquement la méthode consistant à affecter un score en supposant toutes les informations nécessaires à son calcul connues.

### 2.2.1 Superposition des ensembles de points

À partir de la donnée de deux ensembles de points plongés dans l'espace tridimensionnel, une mesure naturelle de la similarité est le RMSD. Il s'agit de mesurer le minimum de la moyenne des distances au carré parmi les couplages possibles entre les ensembles de points. Cette mesure du RMSD est par exemple utilisée dans ProBiS [Konc 2010].

Les auteurs de [Hoffmann 2010] proposent une méthode de score définie pour une superposition et qui ne nécessite la construction d'aucun appariement spécifique. Un noyau de convolution, qui est une forme de produit scalaire, est défini à partir de l'ensemble exhaustif des couples de points, qui a l'avantage de proposer une fonction de score continue en la transformation géométrique qui induit la superposition.

### 2.2.2 Appariement des pseudo-centres

La plupart des méthodes de scores sont basées sur un appariement entre les pseudo-centres. Un premier point de vue consiste à considérer uniquement la taille de l'appariement, modélisant le nombre d'éléments communs. Le score est alors une variante du score de Tanimoto prenant en compte le cardinal de l'appariement relativement aux cardinaux de chacune des deux représentations comparées. Il s'agit de la méthode choisie par eF-seek [Kinoshita 2002] et CavBase [Schmitt 2002].

Une généralisation de cette méthode consiste à pondérer chacune des paires par un score défini entre chaque couple de pseudo-centres. Ce score de paire prend en général en compte la distance si une superposition est également déterminée, et une mesure de similarité entre les propriétés physico-chimiques des pseudo-centres associés.

### 2.2.3 Modèle abstrait et empreinte

Les représentations moléculaires abstraites sont justement motivées par l'existence d'une mesure naturelle de similarité ou de distance définie sur le modèle choisi. Certains modèles, comme les moments invariants, ne nécessitent en particulier aucune notion d'alignement entre les représentations, alors que d'autres comme le modèle de la sphère discrétisée de SiteAlign [Schalon 2008] sont définis pour une orientation donnée entre deux sphères.

La méthode usuelle pour comparer deux empreintes, c'est-à-dire deux vecteurs de booléens ou d'entiers, est le score de Tanimoto, qui consiste à considérer le nombre de motifs présents simultanément dans les deux empreintes par rapport au nombre total de motifs présent dans au moins l'une des empreintes. Ce score peut également être adapté pour donner plus d'importance à certains motifs qu'à d'autres.

## 2.3 Exploration de la macromolécule candidate

À partir de la représentation du site de liaison requête, on souhaite déterminer des régions de la macromolécule candidate qui constituent des sites candidats similaires au site requête. La première difficulté consiste donc à déterminer sur une grande molécule, un ensemble de régions plus restreintes où la similarité avec le site de liaison requête doit être évaluée.

### 2.3.1 Détection *a priori* des sites de liaison candidats

Une première approche pour répondre au problème de l'exploration de la macromolécule candidate consiste à déterminer *a priori*, indépendamment du site requête particulier qui est recherché, l'ensemble des régions de la macromolécule candidate susceptibles de former un site de liaison.

Il est connu que les sites de liaison des ligands sont généralement des cavités à la surface des macromolécules et de nombreux logiciels de détection de cavités sont performants sur les *benchmarks* réalisés [Liang 1998, Yuan 2013]. Différentes méthodes et algorithmes sont présentés dans la section 3.3.

Cette approche possède l'inconvénient que les sites de liaisons prédits doivent correspondre au modèle qui peut être détecté, ne pouvant pas correspondre à toutes les géométries possibles. Il existe par exemple des sites de liaison situés dans des régions plates qui ne sont pas détectables par cette approche, ou bien des cavités permettant de lier des ligands de tailles variés pour lesquelles il est impossible de prédire *a priori* les limites exactes de la zone d'interaction.

### 2.3.2 Recouvrement par fragments réguliers

Une seconde approche pour explorer la macromolécule candidate consiste à déterminer un motif qu'il est possible de générer de manière exhaustive sur les macromolécules, indépendamment du motif recherché. Un tel motif représentatif et alors sélectionné sur le site requête et l'ensemble des motifs sont considérés sur la macromolécule candidate. Ces motifs sont par exemple des disques géodésiques [Yin 2009].

Cette approche repose sur trois éléments essentiels. Tout d'abord il est nécessaire de définir un fragment qui peut être généré sans ambiguïté de manière exhaustive sur l'ensemble de la macromolécule candidate. Ensuite il doit être possible de sélectionner un fragment recouvrant correctement le site de liaison requête. Et enfin une fonction d'évaluation de la similarité doit être correctement définie, et d'une complexité raisonnable puisqu'elle sera évaluée un grand nombre de fois.

De même que pour la détection *a priori* des sites candidats, cette approche limite la variabilité des sites de liaison requêtes qu'il est possible de rechercher efficacement à ceux pour lesquels il est possible de construire un fragment suffisamment représentatif. En revanche l'espace des sites candidats n'est par contraint par une sélection préalable. Cela permet en particulier de déterminer des sites de liaisons prédits sur des régions de macromolécules non connues pour lier un ligand, et pour lesquelles cette capacité à lier n'est pas prévisible avec les outils de détection des cavités.

### 2.3.3 Sous-motif commun maximal

Le problème de la recherche du sous-motif commun maximal consiste à déterminer simultanément une partie du site requête et une partie de la macromolécule candidate, ces deux parties étant similaires. Ce point de vue est le plus général concernant la variabilité des formes acceptables dans le site de liaison requête, ainsi que dans le modèle de représentation des molécules, et ne pose aucune condition sur la localisation des sites candidats.

Le problème de la recherche de sous-motif commun maximal se traduit par un problème algorithmique difficile (NP-complet [Hartmanis 1982]), et certaines heuristiques et simplifications dans les modèles sont utilisées. L'approche générale consiste à représenter les sites et macromolécules par des graphes dont les sommets représentent les atomes ou résidus des molécules et les arêtes représentent des liens chimiques ou topologiques, il s'agit alors de déterminer deux sous-graphes isomorphes traduisant la similarité des représentations. Le problème algorithmique de la recherche de sous-graphe commun maximal et les heuristiques propres au domaine d'application de la similarité moléculaire sont présentés dans la section 3.2.

## 2.4 Alignement des sites de liaison

La construction d'un score évaluant la similarité entre un site requête et une macromolécule candidate dépend généralement d'une notion d'alignement. Il peut s'agir d'un alignement au sens propre, c'est-à-dire la donnée d'une transformation géométrique permettant la superposition tridimensionnelle, et il peut également s'agir d'un appariement entre descripteurs. Enfin certaines représentations ne nécessitent pas d'être alignées pour évaluer leurs similarités.

### 2.4.1 Recherche de la superposition globale optimale

En considérant une mesure de la similarité définie pour une superposition, le problème de l'alignement entre un site requête et une région donnée de la macromolécule candidate consiste à déterminer la transformation géométrique maximisant ce score. Dans ce cas la région de la macromolécule candidate qui est évaluée correspond exactement au site éventuellement prédit, cette méthode dépend donc de la correction d'une méthode d'exploration fournissant des sites candidats bien définis. La méthode dans [Hoffmann 2010] propose par exemple une recherche d'alignement optimal utilisant la propriété de continuité de la fonction de score pour pouvoir appliquer une descente de gradient.

Une autre méthode consiste à déterminer des motifs d'ancrage, définis sur le site requête et la région candidate, permettant de définir à partir d'un couple de tels motifs une transformation géométrique. Des triplets de points de la représentation moléculaires sont ainsi utilisés par SiteEngine [Shulman-Peleg 2004] afin d'extraire un consensus à partir des transformations géométriques induites par chaque couple de triplets.

### 2.4.2 Simplification du modèle

L'alignement peut être défini directement à partir de la représentation des molécules permettant d'évaluer la similarité, mais il peut également être construit sur une représentation simplifiée qui est ensuite traduite à nouveau dans le modèle initial. IsoCleft Finder [Najmanovich 2008] fonctionne en trois étapes : dans un premier temps un appariement définissant un motif commun entre les résidus est déterminé, ensuite à partir de cet appariement un premier alignement géométrique est appliqué, enfin un nouvel appariement est déterminé entre les atomes. Ce dernier appariement peut ainsi être soumis à une contrainte sur la distance obtenue lors du premier alignement afin de simplifier le problème.

### 2.4.3 Absence d'alignement

Certaines représentations abstraites ne nécessitent aucune forme d'alignement pour évaluer la similarité. C'est par exemple le cas des empreintes définies par FuzCav [Weill 2010], des moments invariant de Sommer [Sommer 2007], ou des harmoniques sphériques utilisées dans les travaux de Morris [Morris 2005]. La qualité d'une telle méthode dépend alors de la correction de la détection préalable des sites candidats et la pertinence de la représentation vis-à-vis des propriétés géométriques et chimiques.

# 3 Algorithmes généraux et heuristiques spécifiques

Les sections précédentes montrent que les choix dans les méthodes de représentation des molécules et dans la définition de la similarité sont très variés. En revanche certains algorithmes sont particulièrement souvent utilisés, parfois dans des contextes différents. On présente dans cette section plusieurs problèmes algorithmiques généraux, afin de présenter les techniques spécifiques qui peuvent être utilisées dans le contexte de l'évaluation de la similarité des molécules.

## 3.1 Représentation invariante de la géométrie

Différents modèles mathématiques sont applicables pour représenter la géométrie d'un ensemble de sphères ou d'une surface, indépendamment du choix du repère. Cela signifie que la représentation est identique pour deux objets égaux à rotation près. Ces modèles permettent d'encoder la représentation d'une molécule ou d'une partie d'une molécule, et permettent de mesurer une distance entre deux représentations. L'objectif est alors de choisir une représentation qui encode suffisamment les propriétés géométriques et chimiques pour que la proximité des modèles corresponde effectivement à la similarité moléculaire. Inversement deux molécules suffisamment similaires doivent engendrer deux modèles suffisamment proches.

### 3.1.1 Moments invariants

Les moments invariants [Mamistvalov 1998] sont un modèle mathématique permettant d'encoder certaines propriétés d'une distribution de masse indépendamment de l'orientation du repère de coordonnées, et permettant de définir une distance sur ce modèle.

On considère une distribution de masse $\rho(x_1, x_2, x_3)$ centrée en l'origine. Le moment $\mu_{p_1p_2p_3}$ d'ordre $(p_1, p_2, p_3)$ de la distribution est alors défini comme :

$$\mu_{p_1p_2p_3} = \int\int\int x_1^{p_1} x_2^{p_2} x_3^{p_3} dx_1 dx_2 dx_3$$

Dans le cas général ces moments sont dépendants de l'orientation de l'objet que représente la distribution de masse. Il existe cependant des invariants par rotation comme $O_3 = (\mu_{200} + \mu_{020} + \mu_{002})/\mu_{000}$. On peut de même définir plusieurs invariants en fonction des différents moments $\mu_{ijk}$. Ainsi toute distribution de masse peut être encodée par un vecteur réel, dont chaque coordonnée correspond à l'un des invariants. Afin de pouvoir utiliser la distance euclidienne classique sur ces vecteurs, et pour éviter un biais lorsque les coordonnées ont des amplitudes trop différentes, une normalisation de chaque coordonnée est réalisée par rapport à un ensemble de vecteurs.

Cette méthode est appliquée dans [Sommer 2007] pour représenter une région d'une molécule en représentant un ensemble d'atomes comme une distribution de masse, où la masse de chaque atome est modélisée par une gaussienne, ce qui permet un calcul facile des moments. Cette approche considère donc uniquement la géométrie des sites de liaison dans l'évaluation de la similarité, et plus précisément une approximation du volume de l'union des atomes.

### 3.1.2 Distributions de forme

Les distributions de forme (*shape distributions*) [Osada 2002] permettent de définir pour un polygone tridimensionnel, une représentation invariante par rotation de l'objet, sur laquelle une distance naturelle peut être définie. Le principe consiste à considérer une fonction définie sur un polygone, et à déterminer la distribution de ses valeurs. La comparaison entre deux objets peut alors être ramenée à la comparaison entre deux distributions, par exemple en utilisant la norme $L^N$ de la différence ($\|g\|_{L^N} = (\int g^N)^{\frac{1}{N}}$), il suffit de vérifier que la fonction est bien intégrable.

Un exemple simple, noté $D2$ dans [Osada 2002], d'une telle fonction est la distance entre deux points de la représentation. On considère alors la distribution des distances : pour chaque distance possible on détermine la probabilité que deux points aléatoires soit séparés par cette distance.

Cette méthode est adaptée pour définir des « distributions de forme encodant des propriétés » (*Property-encoded shape distributions*) [Das 2009]. Les propriétés chimiques des sommets sont prises en compte pour déterminer des distributions différentes pour chaque propriété. Pour cela une couleur est affectée à chaque sommet, et une distribution est déterminée pour chaque couple de couleurs possible.

## 3.2 Sous-graphe commun maximal

Le problème du plus grand sous-graphe commun, ou du sous-graphe commun maximal (ou *MCS, Maximal Common Subgraph*) consiste à partir de deux graphes à déterminer deux sous-graphes isomorphes de taille maximale. L'isomorphisme permet en particulier de caractériser la correspondance entre les sommets ainsi associés.

Ce problème apparaît naturellement dans la recherche de motifs communs entre représentations moléculaires, chacun des deux graphes de départ correspondant à la représentation respectivement du site requête et de la macromolécule candidate. Il peut également modéliser la recherche d'un couplage entre les sites requête et candidat afin de déterminer une mesure de similarité.

### 3.2.1 Sous-graphe commun maximal et clique maximale

Étant donnés deux graphes $G = (V_G, E_G)$ et $H = (V_H, E_H)$ le problème du sous-graphe commun maximal consiste à déterminer un sous-graphe $G'$ de $G$ isomorphe à un sous-graphe $H'$ de $H$, de taille maximale. Il s'agit d'un problème NP-complet [Hartmanis 1982]. Ce problème peut être traduit en la recherche d'une clique maximale dans un graphe de compatibilité [Raymond 2002] (ou *modular product*[1]) $K = (V_K, E_K)$ dont les sommets $V_K = V_G \times V_H$ sont les couples de sommets des graphes $G$ et $H$ et dont les arêtes $E_K$ sont les $((g_1, h_1), (g_2, h_2))$ tels que :

— Soit $(g_1, g_2) \in E_G$ et $(h_1, h_2) \in E_H$
— Soit $(g_1, g_2) \notin E_G$ et $(h_1, h_2) \notin E_H$

### 3.2.2 Compatibilité des sommets et arêtes

La donnée d'un isomorphisme entre deux sous-graphes ou de manière équivalente d'une clique du graphe de compatibilité revient à proposer un couplage entre une partie des sommets du premier graphe et une partie des sommets du second. Deux critères de compatibilité, sur les sommets et sur les couples de sommets, peuvent être définis pour contraindre les couplages possibles.

On considère un critère de compatibilité sur les sommets, $v : (V_G, V_H) \to \{0, 1\}$ qui évalue la similarité des propriétés du sommet, par exemple la charge des atomes représentés par un sommet. On considère également un critère défini sur les arêtes $e : E_G \times E_H \to \{0, 1\}$ qui évalue typiquement une correspondance approximative entre les longueurs des arêtes.

Le graphe de compatibilité $K$ est alors défini comme suit :

— $V_K = \{(g, h) \in V_G \times V_H : v(g, h) = 1\}$
— $E_K = \{((g_1, h_1), (g_2, h_2)) \in V_K^2 : \ (g_1, g_2) \in E_G, \ (h_1, h_2) \in E_H, \ e((g_1, h_1), (g_2, h_2)) = 1\}$

Cette méthode a été appliquée aux représentations des molécules [Raymond 2002] où une molécule est modélisée par un graphe dont les sommets sont les atomes et les arêtes sont les liaisons covalentes. La même méthode est appliquée à la structure secondaire des protéines [Grindley 1993]. L'algorithme performant dans le cas général de Bron-Kerbosch [Bron 1973] est fréquemment utilisé, et certaines heuristiques sont développées dans le contexte plus restreint des graphes moléculaires [Depolli 2013].

1. Vocabulaire en Français « graphe de compatibilité » repris de [Minot 2015].

Plusieurs techniques permettent de réduire le temps de calcul moyen du problème. Un premier point de vue consiste simplement à réduire la taille des graphes considérés :

— En utilisant une représentation des molécules à une granularité moins fine, en considérant les résidus plutôt que les atomes par exemple.
— En limitant la taille du motif pouvant être recherché, typiquement les arêtes des graphes modélisant les molécules sont limitées par une longueur maximale.
— En contraignant suffisamment les critères de compatibilité sur les sommets et les arêtes du graphe de compatibilité.

Ces techniques ne réduisent en revanche pas la complexité théorique de l'algorithme. Il est également possible d'utiliser des heuristiques directement dans la résolution du problème de la clique maximale, mais la correction du résultat n'est plus garantie. IsoMIF Finder [Chartier 2015] utilise par exemple l'heuristique énoncée par les auteurs de l'algorithme Bron-Kerbosch [Bron 1973] selon laquelle les plus grandes cliques sont énumérées en premier dans leur algorithme pour se restreindre à un nombre fixé de cliques, sans attendre l'énumération exhaustive de toutes les cliques.

## 3.3 Détection de cavités

Le problème de la détection de cavités consiste à déterminer, à la surface d'un objet tridimensionnel, un ensemble de régions « enfouies » dans le volume de l'objet. Cette détection de cavités est utilisée pour répondre à la problématique de la détection *a priori* de régions de macromolécules pouvant être un site de liaison.

Il n'existe pas de définition précise de « cavité », la validité d'une méthode est donc vérifiée à partir d'un jeu de données de macromolécules et de sites de liaisons connus de petites molécules ligands sur ces macromolécules. L'applicabilité de la méthode repose en particulier sur le principe suivant lequel les sites de liaison sont généralement des « poches » qui forment des creux à la surface de la macromolécule [Laskowski 1996]. Plus précisément, il s'agit de déterminer des régions du vide entourant une macromolécule, susceptibles de contenir un ligand dans une position permettant une interaction non-covalente.

Plusieurs méthodes reposent sur le schéma général suivant :

1. Détecter des points enfouis (*buried*), suivant différentes définition de l'enfouissement.
2. Regrouper (*cluster*) ces points pour former des régions connexes.
3. Évaluer la capacité potentielle de liaison (*druggability*) du site ainsi déterminé.

Une liste de logiciels populaires est présentée dans la table 2. Les principales techniques algorithmiques employées pour décrire la notion de cavité sont présentées dans les paragraphes qui suivent.

### 3.3.1 Placement de sphères vides sur des rayons traversant la macromolécule

Afin de déterminer des sphères vides SURFNET [Laskowski 1995] fonctionne en plaçant des sphères entre chaque couple d'atomes de la macromolécule, et en ajustant le rayon pour qu'elle n'intersectent aucun autre atome. L'ensemble des sphères ainsi obtenues est reporté sur une grille, chaque sphère contribuant aux points voisins de la grille selon leur distance en suivant une gaussienne. Les espaces vides sont alors délimités par une ligne de niveau sur cette grille.

LIGSITE [Hendlich 1997] et son successeur LIGSITEcsc [Huang 2006] reposent sur un principe similaire, en considérant l'ensemble des lignes « protéine - solvant - protéine » dans LIGSITE ou « surface - solvant - surface » dans LIGSITEcsc où la surface de Connolly [Connolly 1983] est préalablement déterminée. Une grille est construite pour paver l'espace, et chaque rayon ainsi défini apporte une contribution à toutes les cellules qu'il traverse. Les cellules sont alors regroupées selon cette valeur, afin de former des régions connexes, qui sont alors classées selon leur taille, et selon le degré de conservation des résidus pour LIGSITEcsc. Ce dernier degré de conservation ConSurf-HSSP [Glaser 2005] permet de favoriser les résidus les plus stables dans les évolutions génétiques. Les rayons « protéine - solvant - protéine » sont

| Logiciel | Référence | Description |
|---|---|---|
| CAVITY | [Yuan 2013] | « Effacement » du milieu extérieur afin de délimiter les sites. Évaluation des volumes et surfaces hydrophobe et donneur H. |
| DoGSiteScorer | [Volkamer 2012] | Utilisation de méthodes d'apprentissage machine pour évaluer les sites de liaison (SVM). |
| SiteMap | [Halgren 2009] | Évaluation de la capacité à lier un médicament en fonction des propriétés physico-chimiques. |
| Fpocket | [Le Guilloux 2009] | Sphères vides du complexe alpha. |
| PocketPicker | [Weisel 2007] | Enfouissement de chaque point d'une grille déterminée suivant 30 directions, et regroupement. |
| Binding-Response | [Zhong 2007] | Sphères vides sur les normales des sommets de la surface de Connolly, et regroupement. |
| Travel Depth | [Coleman 2006] | Plus court chemin entre chaque point d'une triangulation de la surface et l'enveloppe convexe. |
| CAVER | [Petrek 2006] | Tunnel entre l'enveloppe convexe et la surface, avec la profondeur et largeur du tunnel. |
| SURFNET-ConSurf | [Glaser 2006] | Placement de sphères vides centrées entre deux atomes de la macromolécule. Évaluation du degré de conservation des résidus avec ConSurf. |
| LIGSITEcsc | [Huang 2006] | Détection de rayons « surface - solvant - surface » via la surface de Connolly et évaluation de la conservation des résidus par ConSurf. |
| Q-SiteFinder | [Laurie 2005] | Placement d'un groupe méthyl sur chaque sommet de la grille pour évaluer l'énergie, regroupement des sommets favorables. |
| PocketFinder | [An 2005] | Placement d'un groupe aliphatique sur chaque sommet de la grille pour évaluer l'énergie, lissage itératif des valeurs. |
| ConSurf | [Armon 2001, Glaser 2003] | Mesure de degré de conservation de chaque résidu à partir d'un alignement de séquences au sein d'une famille de protéines. |
| PASS | [Brady 2000] | Placement de couches successives de sphères vides sur la surface. |
| LIGSITE | [Hendlich 1997] | Détection des rayons « protéine - solvant - protéine ». |
| CAST | [Liang 1998] | Utilisation du concept du flux discret sur la forme alpha pour la caractérisation des cavités et des ouvertures vers l'extérieur. |
| APROPOS | [Peters 1996] | Variation du complexe alpha en fonction de la valeur de alpha. |
| SURFNET | [Laskowski 1995] | Placement de sphères vides centrées entre deux atomes de la macromolécule. |
| POCKET | [Levitt 1992] | Parcours d'un rayon suivant les 3 axes, détection des événements atome - solvant - atome. |

TABLE 2 – Logiciels de détection de cavités. Présentés dans l'ordre chronologique inversé des publications.

construits suivant les 3 axes canoniques, et les diagonales. PocketPicker [Weisel 2007] propose un procédé essentiellement similaire, mais en calculant un indice d'enfouissement sur chaque point de la grille selon 30 directions. Il s'agit du logiciel utilisé dans l'étude ultérieure [Weisel 2009] qui montre que le volume et l'enfouissement d'une cavité sont fortement corrélés à sa capacité à lier un ligand.

### 3.3.2 Autre placement de sphères vides

Il existe d'autres méthodes qui ont été utilisées avec succès afin de placer des sphères vides dans les espaces susceptibles de constituer des sites de liaison. Le logiciel Binding-Response [Zhong 2007] définit des sphères qui sont placées sur la normale des atomes surface, suivant la triangulation de surface de Connolly.

Le logiciel PASS (*Putative Active Site with Spheres*) [Brady 2000] propose une méthode d'accrétion de couches successives de sphères sur l'ensemble de la surface d'une macromolécule. Une première phase consiste à placer des sphères vides de rayon fixé, simultanément tangentes à des triplets d'atomes. Ces sphères sont filtrées afin de ne conserver que celles qui se trouvent dans une région suffisamment concave, qui n'intersectent pas la molécule, et qui sont suffisamment éloignées d'une autre telle sphère. À partir de cet ensemble de sphères une seconde phase consiste à répéter une étape d'ajout de nouvelles sphères tangentes aux précédentes, jusqu'à ce qu'aucune nouvelle sphère ne puissent être ajoutée. Enfin la dernière phase consiste à regrouper les ensembles de sphères en régions définissant les sites prédits.

Enfin le logiciel Fpocket [Le Guilloux 2009] utilise la théorie des formes alpha, qui permet de déterminer un ensemble de sphères simultanément tangentes à trois atomes mais n'intersectant aucun autre atome. Ces dernières sphères sont filtrées selon un rayon maximal et minimal pour éliminer les régions non accessibles, ou bien au contraire trop exposées. Les sphères sont également annotées par les propriétés physico-chimiques des atomes en contact, puis regroupées pour constituer les sites prédits.

### 3.3.3 Plus court chemin vers l'extérieur

Travel Depth [Coleman 2006] et CAVER [Petrek 2006] modélisent la notion de plus court chemin entre un point de la surface et l'enveloppe convexe de la macromolécule.

Travel Depth considère une triangulation d'une surface délimitant le volume de la macromolécule. À partir d'une grille régulière, chaque cellule de la grille est assignée à l'une des quatre catégories suivantes : à l'extérieur de l'enveloppe convexe (O), à l'extérieur de la surface mais à l'intérieur de l'enveloppe convexe (B), contenant un point de la surface (S), entièrement à l'intérieur de la surface (I). Un graphe modélise l'adjacence des cellules (B), (S) et (I), les arêtes du graphe étant pondérées par la distance euclidienne entre les centres des cellules. La profondeur des cellules (B) et (S) est alors calculée par l'algorithme de Dijkstra de calcul des plus courts chemins, les cellules (I) étant initialisées à une profondeur nulle, et les cellules (B) et (S) à une profondeur infinie.

CAVER considère également une grille dont les sommets peuvent être à l'extérieur de l'enveloppe convexe, à l'intérieur de la molécule, ou bien entre les deux. Un graphe d'adjacence entre les points de la grille est également construit, mais contrairement à Travel Depth ce ne sont pas les arêtes mais les sommets qui sont valués par une mesure d'un coût pour les traverser. Ce coût est construit pour modéliser la « difficulté » de passage d'un ligand par ce sommet et correspond à l'inverse du rayon maximal d'une sphère qui n'intersecte pas la molécule. La recherche du plus court chemin est alors également réalisée par l'algorithme de Dijkstra. La prise en compte de la largeur du tunnel pour atteindre un point de la surface permet de favoriser les poches à la fois les plus enfouies mais également les plus facilement accessibles par un ligand depuis l'extérieur.

### 3.3.4 Évaluation de l'énergie pour un ligand virtuel

Une alternative au placement de sphères vides pour lesquelles un indice d'enfouissement est calculé consiste à modéliser un fragment de ligand virtuel pour déterminer un score énergétique entre ce fragment

virtuel et la macromolécule. Q-SiteFinder [Laurie 2005] et PocketFinder [An 2005] fonctionnent tous deux de manière similaire, en plaçant respectivement un groupe méthyl et un groupe aliphatique sur l'ensemble des sommets d'un grille, afin d'en estimer l'énergie et les résidus susceptibles d'interagir.

### 3.3.5 Évaluer la capacité à être la cible d'un médicament

D'autres approches permettent d'évaluer la capacité d'un site à être effectivement la cible d'un médicament, après que la géométrie du site ait été définie [Hussein 2016]. La méthode proposée dans PockDrug-Server [Hussein 2015] consiste par exemple à annoter des sites préalablement déterminés par proximité d'un ligand ou par une méthode géométrique telle que Fpocket [Le Guilloux 2009], pour proposer un score à chaque site dépendant de descripteurs chimiques par une technique d'apprentissage machine.

# 4 Motivations des choix pour le développement de BioBind

Nous avons présenté le problème de la recherche de similarité entre macromolécules, en décrivant d'une part les approches existantes et d'autre part certains algorithmes clés. Le développement de notre algorithme BioBind est basé sur l'analyse des avantages et inconvénients des différentes approches, afin de déterminer les choix pertinents pour l'application souhaitée.

## 4.1 Avantages et inconvénients des différentes approches

### 4.1.1 Nécessité de détecter les sites *a priori*

Nous avons vu que le choix de la représentation des molécules est central dans toute approche de recherche de similarité. Une première distinction fondamentale entre les différentes méthodes concerne la possibilité de représenter n'importe quelle partie d'une macromolécule ou bien seulement des sites déterminés *a priori* de manière indépendante du site requête. L'inconvénient immédiat de ce dernier choix est qu'il est nécessaire d'échantillonner au préalable l'ensemble des cibles candidates pour construire les sites candidats, et quelle que soit la méthode de détection des cavités il n'est pas possible de déterminer tous les sites potentiels qui ne sont par exemple pas dans des cavités ou bien lorsqu'une cavité correspond à plusieurs sites de tailles différentes selon le ligand effectif. Cela permet en revanche d'utiliser des représentations spécifiques comme les empreintes, ou d'autres modèles mathématiques abstraits, afin de définir une mesure de la similarité.

### 4.1.2 Complexité de la recherche de sous-motif commun

De très nombreuses approches reposent sur une méthode de recherche de motif commun entre deux ensemble de pseudo-centres, en transformant le problème en une recherche de cliques maximales sur un graphe produit. L'attrait de cette méthode est que le problème est bien défini, avec des algorithmes exacts permettant d'évaluer les heuristiques éventuelles choisies en tant que telles et indépendamment de la problématique biologique. Cependant le temps de calcul en pratique reste trop important dès que le graphe de compatibilité est trop grand et dense, ainsi les critères de compatibilité entre les sommets sont souvent trop stricts par rapport au degré de variabilité existant entre des sites que l'on souhaite pourtant considérer similaires.

D'autres approches fonctionnent essentiellement en considérant des « super-motifs », typiquement des triplets de pseudo-centres de la représentation choisie, pour déterminer des appariements locaux qui peuvent guider un appariement plus global du site, à partir de la transformation géométrique induite par le couple de motifs. De la même manière que pour la recherche de clique maximale, le nombre de triplets est généralement très grand si le critère de construction n'est pas suffisant contraint, ce qui en limite l'application pratique.

### 4.1.3 Représentation réduite

Enfin, le choix d'une représentation réduite, en considérant par exemple des groupes fonctionnels au lieu de l'ensemble des atomes, permet de réduire le temps de calcul pratique des différents algorithmes en diminuant la taille des entrées mais sans changer la complexité. Ce choix permet également de réduire la sensibilité aux coordonnées atomiques, c'est à dire que des petites variations dans les positions d'atomes individuels modifient moins la position globale d'un regroupement d'atomes. En revanche le niveau de granularité de l'atome est généralement ignoré lors de la mesure de la similarité et ainsi la qualité du modèle de réduction de la représentation est primordiale alors qu'elle est difficile à valider en tant que telle.

De manière analogue, le choix d'un modèle mathématique abstrait (comme les moments invariants ou descripteurs de Zernike) simplifie la méthode de mesure de similarité, en revanche la pertinence du modèle lui même pour représenter les propriétés physico-chimiques et géométriques est d'autant plus importante.

## 4.2 Choix pour notre algorithme BioBind

Les choix réalisés dans la conception de notre algorithme BioBind concernent d'une part le modèle et la définition de la similarité et d'autre part la méthode de recherche. Cette distinction permet une séparation claire entre la définition d'un problème d'optimisation et les techniques de résolution. L'ensemble de la solution BioBind est orientée sur une problématique pharmacologique précise, la prédiction de cibles secondaires potentielles pour des ligands, qui motive les choix dans la conception de l'algorithme présenté dans le prochain chapitre, et permet de définir une méthode pertinente de validation de l'approche.

### 4.2.1 Représentation des molécules et mesure de la similarité

L'objectif suivi pour définir une représentation des molécules consiste à représenter la notion d'accessibilité au milieu extérieur par une surface, avec un niveau de granularité à l'échelle de l'atome. La théorie des formes alpha est utilisée pour proposer un modèle de surface triangulée, qui correspond de manière exacte à la notion d'accessibilité définie comme la capacité d'une molécule d'eau à entrer en contact avec un atome de la macromolécule. Ce modèle permet par ailleurs de définir un site de liaison, connu ou prédit, comme un sous-ensemble de la représentation sans autre contrainte spécifique sur la forme des sites.

On définit la similarité à partir d'une superposition qui détermine à son tour un schéma de correspondance entre les sommets de deux surfaces triangulées. Chaque sommet est associé à un unique atome, cependant la distinction entre sommet et atome est nécessaire car un même atome peut être associé à plusieurs sommets de notre modèle. Il s'agit d'une conséquence de notre méthode de régularisation de la topologie de la surface qui est également détaillée dans le chapitre suivant.

### 4.2.2 Recherche et alignement de motifs

À partir de ces définitions, le problème de la recherche de similarité locale peut être défini comme un problème d'optimisation de la superposition d'un site sur la surface d'une macromolécule. On rappelle qu'aucune contrainte de forme n'est imposée *a priori* sur le site qui est recherché, en particulier l'espace de recherche sur la macromolécule n'est pas limité à un ensemble de régions prédéterminées. En effet nous souhaitons que notre approche soit la plus exhaustive possible dans l'ensemble des sites pouvant être prédits, qui ne sont pas nécessairement des sites connus, ni même obligatoirement des cavités, la seule limite étant le degré de couverture de la base de données de structures tridimensionnelles utilisée.

La méthode de résolution pour déterminer la meilleure superposition entre un site et une surface est basée sur une fragmentation exhaustive de la surface en petits disques géodésiques, qui sont superposés pour proposer un ensemble d'alignements ensuite étendus sur l'intégralité du site. Cette heuristique,

conceptuellement similaire à la méthode consistant à considérer les triplets d'une représentation constituée d'un ensemble de points, met à profit la structure topologique de la surface.

Chapitre III

# BioBind - Bind Is Not Docking

## Introduction

Notre algorithme, BioBind, est une approche de prédiction de cibles par évaluation de similarités entre macromolécules. Il repose sur une modélisation des surfaces moléculaires qui utilise la théorie des formes alpha, cette représentation permettant de définir le problème de la recherche de similarités locales entre un motif donné et l'ensemble d'une surface moléculaire. Le problème de la prédiction de cibles consiste alors à classer un ensemble de molécules en fonction du degré de similarité avec un motif recherché.

L'approche est formalisée par un problème d'optimisation où il s'agit de rechercher la meilleure superposition entre deux régions qui maximise une mesure de similarité. Étant donnés un site de liaison requête et une cible candidate, l'espace de recherche est ainsi constitué de l'ensemble des sites candidats pouvant être définis sur la surface de la cible candidates et de l'ensemble des superpositions réalisables.

L'espace de recherche étant trop vaste pour être exploré de manière exhaustive, une heuristique a été développée consistant essentiellement à définir des régions circulaires, approximant la notion de disque géodésique, qui sont générées de manière exhaustive à la surface des molécules. La régularité géométrique de ces régions permet d'utiliser des heuristiques efficaces pour superposer des régions circulaires, ces dernières superpositions étant ensuite traduites à l'échelle du site requête qui est le motif initial recherché.

## Sommaire

# 1 Présentation générale de l'approche

Du point de vue du problème de la prédiction de cibles, l'objectif de notre approche BioBind est de proposer une liste ordonnée de cibles à partir d'un site requête, triées suivant la capacité prédite à lier le même ligand. L'approche consiste à évaluer une mesure de similarité locale entre le site requête et une région de chaque cible candidate, le principe d'inférence de l'interaction permettant de traduire cette similarité comme une capacité à lier le même ligand.

Le problème consiste donc en la recherche d'un motif, le site de liaison requête, sur une unique cible candidate ainsi que l'évaluation de la similarité locale. C'est cette valeur attribuée pour chacune des cibles candidates qui est en effet utilisée pour définir le classement résultant ; elle doit donc être comparable entre des cibles différentes. Notre approche de résolution est basée sur une représentation de la surface des macromolécules et une notion de fragmentation de cette surface afin d'adresser la problématique de la complexité de la recherche.

## 1.1 Problème de la recherche de similarités locales pour la prédiction de cibles

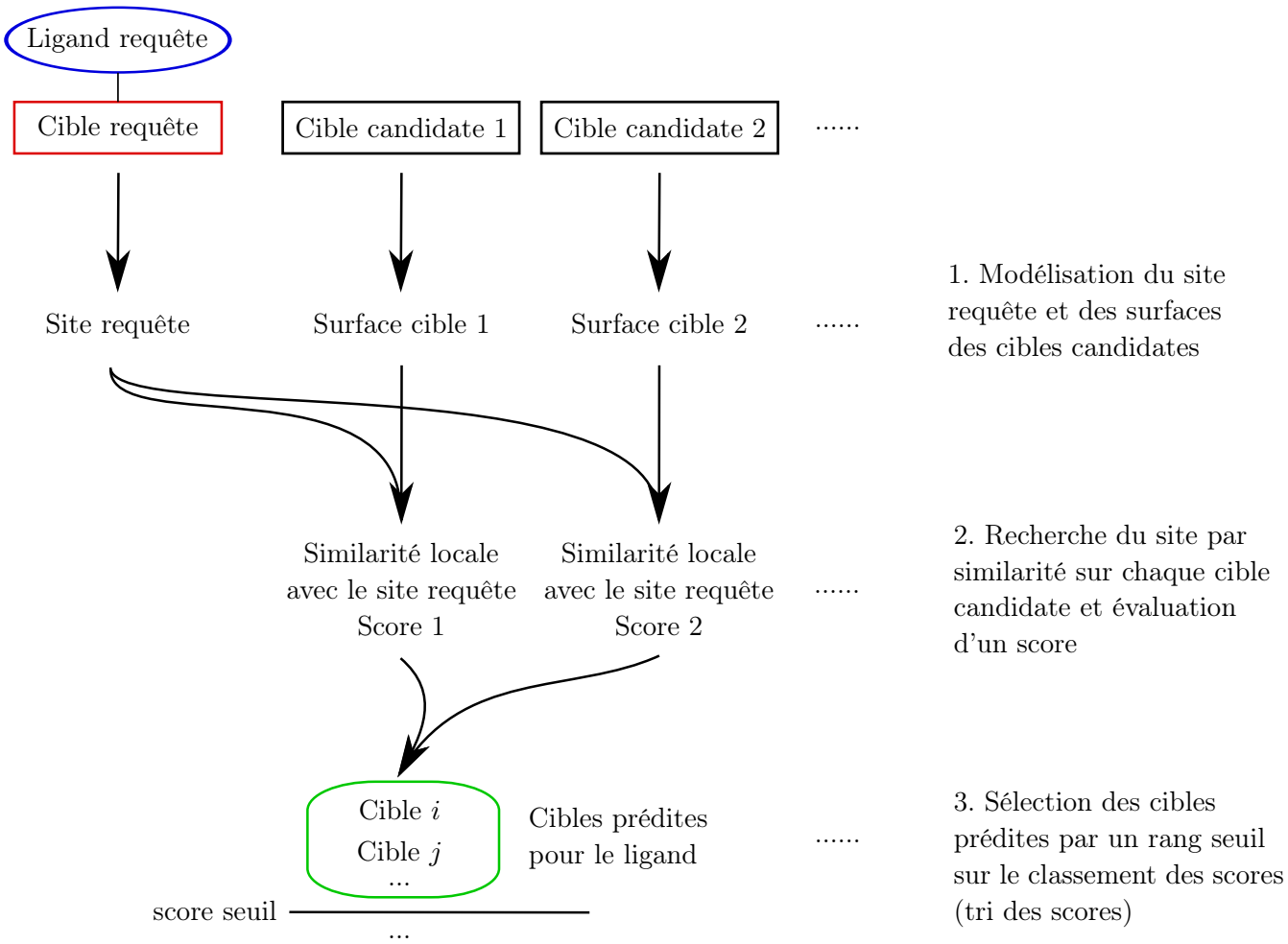



Figure 1 – Processus global de la prédiction de cibles par similarité de BioBind.

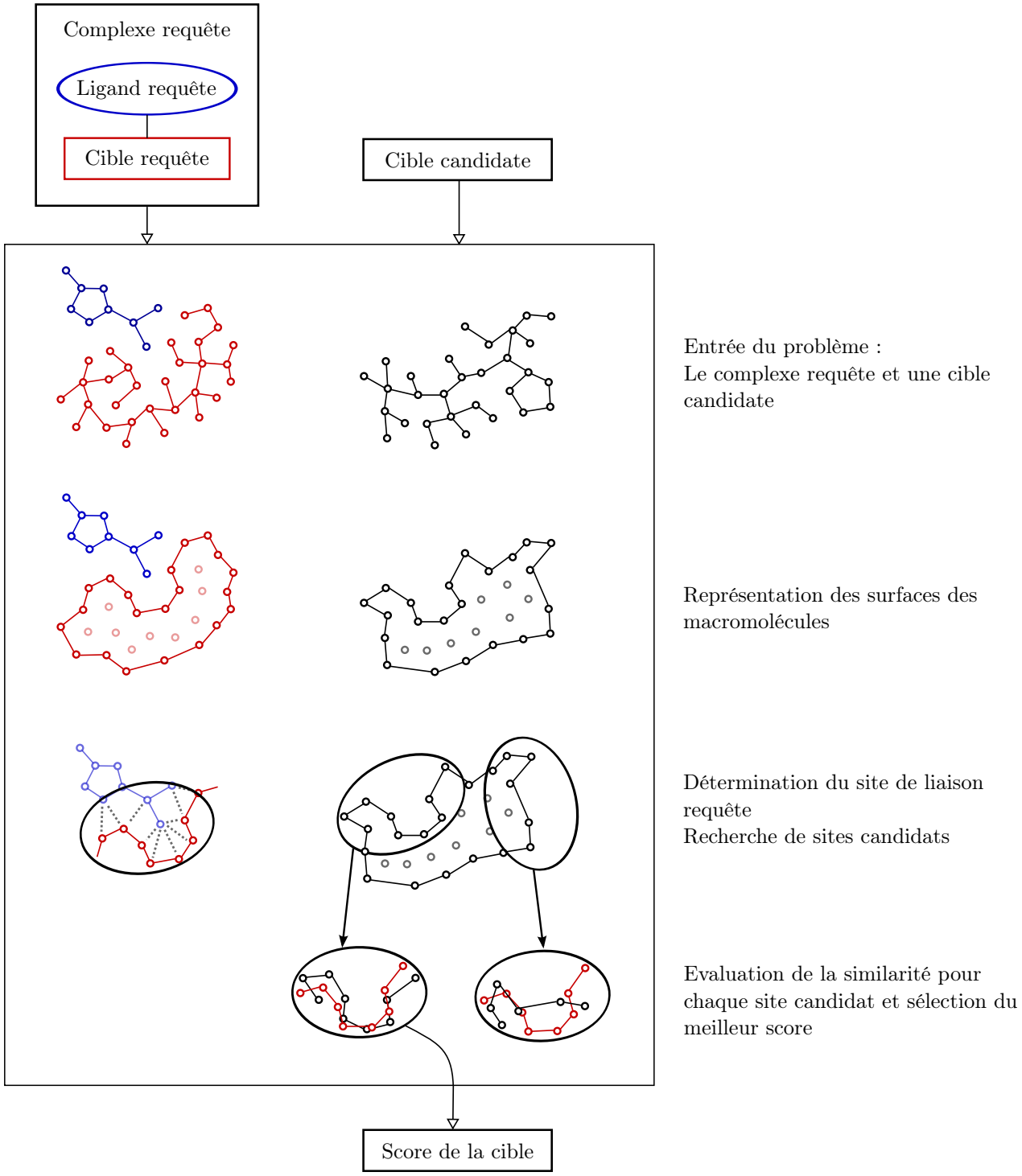



FIGURE 2 – Problème d'évaluation de la similarité locale entre le complexe requête et une cible candidate.

Le problème de la prédiction de cibles est un problème de classification binaire : étant données une molécule *ligand requête* et un ensemble de macromolécules *cibles candidates*, il s'agit de partitionner en deux l'ensemble des cibles candidates en sélectionnant un sous-ensemble de *cibles prédites*. L'approche choisie, par similarité des cibles, consiste à déterminer cet ensemble de cibles prédites par similarité avec une première *cible requête* également fournie en entrée et dont l'interaction avec le ligand requête est connue.

Le processus général est schématisé dans la figure 1. Dans un premier temps, le site de liaison requête et les surfaces des cibles candidates sont modélisés à partir d'un complexe requête fourni par l'utilisateur ainsi qu'une base de données de macromolécules cibles à explorer. Ensuite pour chacune des surfaces candidates, le site requête est recherché, et une évaluation de la similarité est attribuée. Enfin la dernière étape du processus consiste à trier les cibles suivant cette valeur de similarité afin de proposer un sous-ensemble des cibles qui sont prédites pour interagir avec le ligand requête par le choix d'un seuil dans la liste résultat. Le processus de recherche du motif pour une cible candidate donnée est plus précisément présenté dans la figure 2, correspondant au problème suivant :

**Entrée :** Le complexe requête constitué des conformères de la molécule ligand requête et de la macromolécule cible requête, ainsi que le conformère d'une macromolécule cible candidate.

**Sortie :** Un score de similarité affecté à la cible candidate, ainsi que la donnée d'un site prédit correspondant.

## 1.2 Approche pour la détermination de la meilleure superposition

À partir d'une définition d'un concept de région qui modélise une partie de la surface d'une macromolécule, le problème consiste à déterminer pour une région requête donnée la meilleure région candidate et la meilleure superposition. La notion de « meilleur » faisant référence à une mesure de similarité définie pour une superposition donnée d'un couple de régions données. Une première idée naturelle pour explorer l'espace de recherche consiste à lister toutes les régions de la cible candidate qui ont la même forme que le site requête. C'est à dire une méthode exhaustive sur l'ensemble des sites candidats. Cependant dans le cas général, la notion de « même forme » n'est pas définie. Une seconde idée également naturelle consiste à « parcourir » l'ensemble de la surface candidate en considérant toutes les régions candidates pouvant être obtenues par projection de la région requête. De la même manière la notion de parcours de cet espace de superposition n'est pas évidente.

Ces difficultés motivent la considération d'un type de région particulier, ayant des propriétés géométriques permettant d'utiliser différentes heuristiques. Une *région circulaire de surface*, ou région circulaire, consiste en une approximation d'un disque géodésique autour d'un sommet central. D'une part il est possible de définir une génération exhaustive des régions circulaires de surface sur la surface de toute macromolécule. D'autre part cette structure permet d'utiliser des outils simples pour établir une superposition et une évaluation de la similarité rapide entre deux régions circulaires. La forme régulière de ces régions permet de définir une méthode d'alignement plus rapide, utilisant un axe de rotation privilégié sur un disque géodésique. Elle permet également d'utiliser une méthode de filtrage, afin d'éliminer une partie importante des alignements qu'il est nécessaire de réaliser.

L'heuristique développée en utilisant les régions circulaires repose sur le principe suivant lequel si deux sites requête et candidat sont suffisamment similaires, alors ils partagent en particulier deux régions circulaires qui sont également similaires. Ainsi, en partant de tous les couples de régions circulaires similaires il est possible de compléter l'information au niveau des sites dans une étape de recomposition du site prédit.

# 2 Modèle de la surface des molécules

La représentation de la surface des macromolécules est construite à l'aide de la théorie des formes alpha [Edelsbrunner 1995]. Cette surface représente l'ensemble des atomes accessibles au milieu extérieur, susceptibles d'interagir avec des atomes d'une autre molécule. Une régularisation du polytope résultant est effectuée afin de définir d'une part des caractéristiques géométriques pour chaque sommet et d'autre part une notion précise de régions de surface utilisées dans nos algorithmes.

## 2.1 Modèle des molécules

Une molécule ou plus précisément un conformère est modélisé comme un ensemble d'atomes $M = \{a_i\}_{i=1}^{n}$. Chaque atome $a \in M$ est décrit par sa position $\text{pos}(a)$, l'élément chimique $\text{elt}(a)$, le type de résidu auquel il appartient $\text{res}(a)$. Ces informations sont directement extraites des fichiers fournis en entrée (par exemple au format mmCIF [Bourne 1995]), que ce soit pour le ligand et la cible requête, comme pour chaque cible candidate. Ce modèle est étendu par BioBind en associant des valeurs supplémentaires à chaque atome, afin de mieux caractériser les molécules.

Tout d'abord, un typage des atomes est effectué en associant à chaque atome $a$ une valeur énumérative $\text{type}(a)$ caractérisant les propriétés physico-chimiques de l'atome. Ce typage est effectué à partir de la nature de l'atome, du résidu, et de sa position au sein du résidu, par une fonction qui sera référencée comme pm-type [1]. Cette valeur modélise les types d'interaction chimique dans lesquelles l'atome est susceptible d'être impliqué avec un ligand. Elle est utilisée dans la méthode d'évaluation de la similarité entre les molécules.

$$\text{type}(a) = \text{pm-type}(\text{elt}(a), \text{res}(a), \text{position au sein du résidu})$$

Ensuite une seconde valeur est associée à chaque atome $a$, dépendant uniquement de l'élément chimique. Il s'agit du rayon de Van der Waals qui correspond à la distance au delà de laquelle les interactions non-covalentes peuvent se produire [Rowland 1996]. Nous utilisons les valeurs proposées par le CCDC (*Cambridge Crystallographic Data Centre*) [2]. Cette valeur notée $\text{rad}(a)$ est utilisée pour définir les propriétés géométriques d'une molécule, et en particulier pour la construction du modèle de la surface défini dans la suite. Il s'agit du rayon utilisé pour modéliser une molécule comme un ensemble de sphères, chacune correspondant à un atome.

## 2.2 Surface des molécules

### 2.2.1 Forme alpha d'une macromolécule

Les formes alpha ont été initialement introduites par H. Edelsbrunner pour formaliser une notion intuitive de « forme d'un ensemble de points » [Edelsbrunner 1983]. Les formes alpha, alpha pondérées, et beta, sont présentées dans l'annexe A. Une revue détaillée des formes alpha pondérées et leurs applications est proposée dans [Edelsbrunner 2010], on présente ici seulement l'application qui en est faite pour la modélisation des macromolécules dans le cadre de BioBind.

Notre modèle d'une molécule détermine pour chaque atome, parmi d'autres propriétés, sa position et son rayon. Cela permet de voir une molécule comme un ensemble de sphères, sur lequel on peut ainsi appliquer la théorie des formes alpha pondérées. La forme alpha d'une molécule est un polytope sur les centres des atomes, dont la frontière correspond aux atomes accessibles au milieu extérieur. Plus précisément, étant donné un rayon de test, choisi comme le rayon de Van der Waals d'une molécule d'eau

1. Types atomiques et affectation des types internes à Bionext SA (développés par Pascal Muller pascal.muller@bionext.com).
2. http://www.ccdc.cam.ac.uk/support-and-resources/ccdcresources/Elemental_Radii.xlsx

de 1,5 Angström, pour chaque facette il est possible de placer une molécule d'eau simultanément en contact avec tous les atomes de la facette et n'intersectant aucun autre atome de la molécule.

Ce polytope est construit comme une union de simplexes (tétraèdres, triangles, segments, et points) qui sont dits *exposés* parmi les simplexes d'une *triangulation régulière* [Edelsbrunner 1995] qui vérifient la propriété de Delaunay [Loisy 1951], et l'implémentation utilisée dans le cadre de BioBind pour le calcul de la forme alpha pondérée est une librairie propriétaire de Bionext SA. Ce polytope ne constitue pas une surface au sens d'une *variété*, c'est à dire que ce polytope n'est pas topologiquement localement similaire au plan euclidien $\mathbb{R}^2$. La figure 3 présente un exemple de point singulier de la forme alpha. Cependant cette similarité locale avec le plan est utile pour notre modèle, permettant de définir différentes constructions géométriques, et motive la régularisation qui en est faite.

### 2.2.2 Régularisation de la topologie

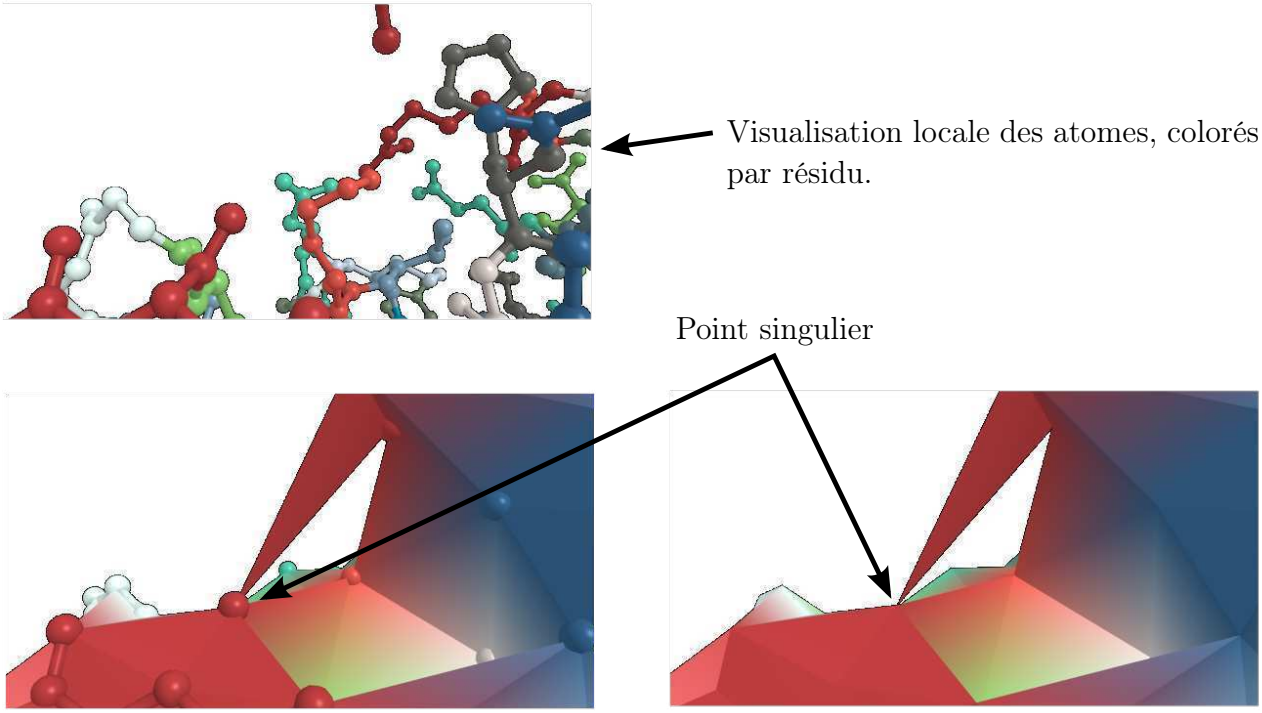



Figure 3 – Point singulier de la forme alpha.

Afin de pouvoir manipuler la surface d'une molécule, on utilise une régularisation qui consiste à extraire une variété orientée sans bord à partir de la forme alpha. Une variété de dimension 2 est topologiquement similaire au plan euclidien dans le sens où pour tout point de la variété il existe un voisinage homéomorphe à un ouvert du plan (déformation continue, sans « couper »). Cela permet notamment de définir une distance en surface ou *distance géodésique* pertinente, ainsi que des propriétés géométriques locales comme la normale et une courbure en chaque point de la surface.

Cette construction considère uniquement les facettes triangles de la forme alpha, en ignorant les arêtes dites pendantes (qui ne sont inclues dans aucun triangle), et les points dits isolés (qui ne sont inclus dans aucune arête). En pratique cette perte d'information est négligeable dans le contexte du modèle d'une macromolécule. En effet les points isolés n'existent pas car cela signifierait qu'un atome d'une molécule est situé à une distance de tout autre atome supérieure à la somme de leurs rayons de Van der Waals augmentée du rayon d'une molécule d'eau, ce qui contredit la possibilité d'une liaison chimique définissant une molécule. Les arêtes pendantes sont elles *très* rares, car elles correspondent à des chaînes d'atomes par ailleurs déconnectées du reste de la molécule, énergétiquement peu favorables.

L'algorithme schématisé dans la figure 4 considère l'ensemble des triangles qui sont les facettes du polytope, dont l'orientation est donnée dans la forme alpha par l'information du coté *intérieur* et *extérieur* de la molécule. Le processus consiste alors à *recoller* les triangles compatibles en fusionnant les sommets partagés par deux triangles ainsi associés. En particulier, certains atomes restent représentés par plusieurs sommets distincts. La structure de données résultante est une variante du modèle des demi-arêtes (HDS, *Half-edge Data Structure* [Kettner 1997]) permettant de représenter une variété orientée sans bord triangulée, et l'algorithme garantit que chaque arête est partagée par exactement deux triangles.

### 2.2.3 Caractérisations géométriques

Une surface $S = (V, E, T)$ d'une molécule $M$ est modélisée par un ensemble de sommets $V$, d'arêtes $E \subset V \times V$, et de triangles $T \subset V \times V \times V$, munis des opérations suivantes schématisées dans la figure 5 :

$\forall s, t \in V,$ tels que $(s, t) \in E$,

| | |
|---|---|
| $\text{atom}(s) = a \in M$ | atome associé |
| $\text{neighb}(s) = \{w \in V : (s, w) \in E\}$ | sommets voisins |
| $\text{left-trg}(s, t) = u \in V,$ tel que $(t, s, u) \in T$ | triangle gauche |
| $\text{right-trg}(s, t) = v \in V,$ tel que $(s, t, v) \in T$ | triangle droit |

On note que la fonction $\text{atom} : V \to M$ n'est pas injective en général, c'est à dire qu'un même atome de la molécule peut être partagé par plusieurs sommets de la surface. Cette situation se présente lorsqu'un atome correspond à un point singulier de la forme alpha, et qui a été régularisé par notre algorithme. Les fonctions $\text{left-trg}, \text{right-trg} : E \to V$ sont bien définies et surjectives par construction de la surface comme une variété union de triangles. On note enfin qu'il est possible de considérer le graphe $G = (V, E)$ induit par la structure de la surface, qui sera appelé le *graphe de surface* de la molécule $M$.

L'orientation des triangles permet de définir une normale pour chacun d'eux, comme le produit vectoriel de deux arêtes consécutives du triangle, dirigée vers l'extérieur de la molécule. Cette information est reportée sur chaque sommet $s$ en associant pour chacun la moyenne $\text{nor}(s)$ des normales des triangles adjacents. Pour un sommet $s$ donné, une courbure $\text{curv}(s)$ est définie comme la moyenne pour chaque voisin des produits scalaires entre le vecteur unitaire dans la direction du voisin par la normale du sommet $s$. En particulier la courbure prend ses valeurs dans l'intervalle $[-1; 1]$, les valeurs positives correspondant à des « creux » et les valeurs négatives à des « bosses ».

## 2.3 Régions de surface

On définit la notion de *région de surface* de la manière suivante. Étant donnée une surface $S = (V, E, T)$, un ensemble de sommets $W \subseteq V$ définit une région $R = S|_W = (W, E|_W, T|_W)$ où $E|_W$ est la restriction des arêtes de $E$ aux sommets de $W$, et de même $T|_W$ est la restriction des triangles de $T$ aux sommets de $W$. On définit le bord $\text{border}(R)$ comme l'ensemble des sommets de $P$ qui ont des voisins dans $S$ qui ne sont pas dans $R$. On note que les fonctions left-trg et right-trg ne sont pas toujours définies sur les arêtes du bord d'une région, comme on peut le voir en figure 6, si le sommet de la surface complétant l'arête pour former le triangle ne fait pas partie de la région.

Ces régions permettent de modéliser les parties de la surface correspondant aux sites de liaisons, ainsi que d'autres motifs qui sont définis sur la surface des molécules dans notre algorithme de recherche présenté dans la suite. Deux régions peuvent être comparées afin de mesurer la similarité pour une superposition donnée, selon la méthode décrite dans la section suivante.

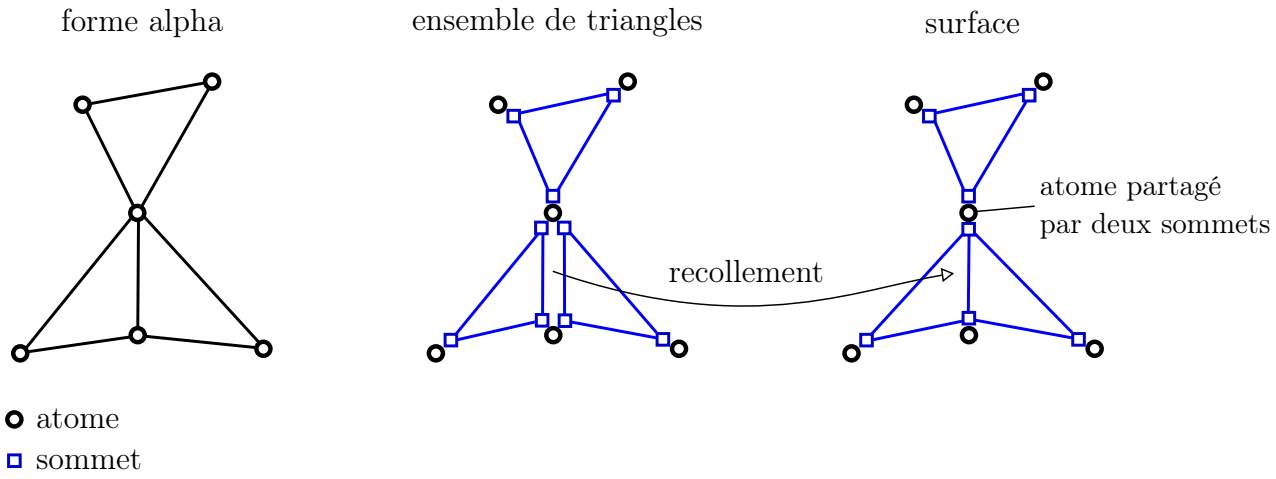
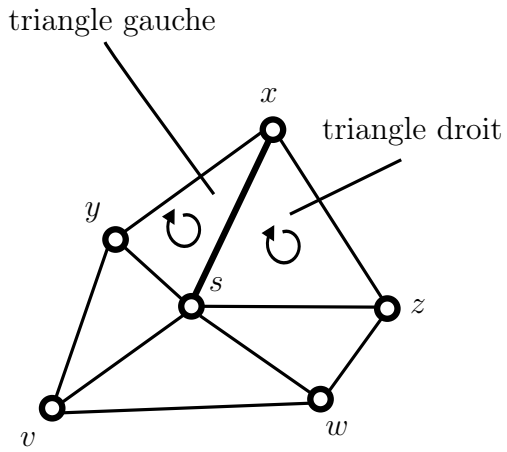



FIGURE 4 – Algorithme de régularisation de la surface. L'atome au centre correspond à un point singulier de la forme alpha, qui donne lieu à deux sommets par notre algorithme de régularisation de la surface.




$$\begin{aligned} \text{neighb}(s) &= \{x, y, z, v, w\} \\ \text{left-trg}(s, x) &= y \qquad (x, s, y) \in T \\ \text{right-trg}(s, x) &= z \qquad (s, x, z) \in T \end{aligned}$$

FIGURE 5 – Représentation de la surface au voisinage d'un sommet $s$.

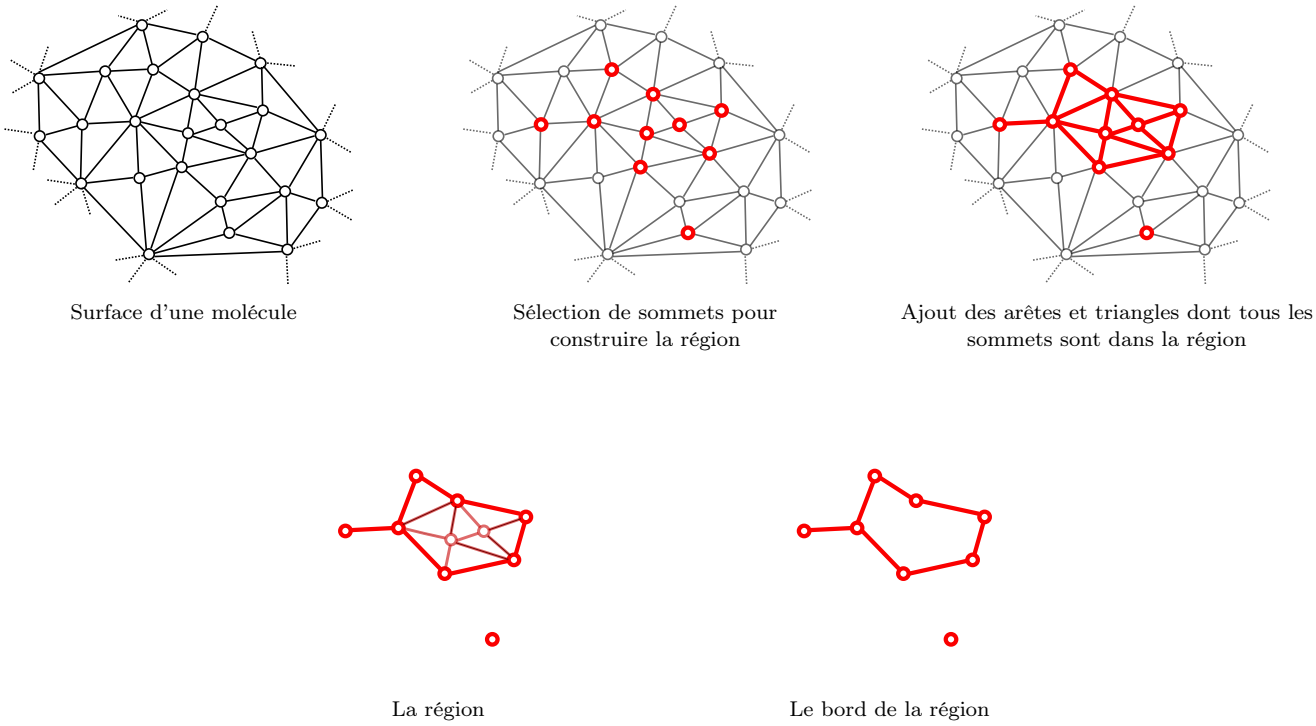



FIGURE 6 – Construction d'une région de surface. À partir d'une sélection de sommets sur une surface, les arêtes et triangles dont tous les sommets sont inclus dans cette sélection de sommets sont ajoutés. Le bord correspond alors à l'ensemble des sommets qui ont un voisin dans la surface qui n'est pas dans la région, ainsi que les arêtes dont les deux extrémités sont sur le bord.

# 3 Évaluation de la similarité locale

Nous rappelons que l'objectif de notre algorithme est de déterminer à partir d'une région requête, la meilleure région candidate et la meilleure superposition des régions. C'est-à-dire la région candidate et sa superposition sur le site requête qui maximise l'évaluation de la similarité.

## 3.1 Superposition entre deux régions

À partir d'une région site requête, dont la construction est détaillée dans la section 4.1, on considère une région candidate de la surface de la cible candidate ainsi qu'une superposition des deux régions afin de définir une mesure de similarité. Plus généralement notre mesure de similarité est définie pour deux régions quelconques, qui sont référencées comme la région requête et la région candidate, car cette mesure est utilisée afin d'évaluer la similarité d'un site candidat mais également dans le déroulement de notre algorithme pour d'autres régions dites circulaires qui sont utilisées dans l'approche de résolution (section 5).

Pour une région donnée de la surface d'une cible candidate, notre mesure de similarité est définie en considérant une superposition avec la région requête. C'est à dire que les deux régions sont plongées dans l'espace afin de considérer une superposition. Plus précisément, dans notre implémentation, la région requête est conservée dans sa position initiale, et une transformation géométrique est appliquée à la région candidate afin de déterminer une superposition. Ce point de vue permet d'identifier une transformation géométrique et une superposition par application de la transformation à la région candidate. Ce sont uniquement les déplacements, c'est-à-dire les isométries directes, qui sont considérées. Cela signifie essentiellement qu'aucune flexibilité n'est traitée dans la phase de superposition.

## 3.2 Mesure de similarité

La mesure de similarité entre une région requête $Q$ et une région candidate $T$ superposée consiste à mettre en correspondance des sommets de chaque région. Une évaluation de similarité est déterminée pour tous les couples de sommets en fonction des caractéristiques de chaque sommet selon la méthode présentée dans la prochaine section (3.2.1). Ensuite un sous-ensemble des ces couples est déterminé, plusieurs méthodes utilisées étant présentées dans la section suivante (3.2.2).

### 3.2.1 Composantes et score entre deux sommets

**Caractéristiques des sommets** Quatre caractéristiques sont considérées sur chaque sommet $s$ d'une région permettant de définir un score de similarité entre deux sommets :

- **La position :** La position du sommet est la position de l'atome associé, $\text{pos}(s) \in \mathbb{R}^3$ représenté par ses trois coordonnées réelles (section 2.1).
- **Le typage :** Le type du sommet correspond au typage de l'atome associé, $\text{type}(s)$ prenant ses valeurs dans un ensemble fini de typages prédéfinis (section 2.1).
- **La normale :** La normale du sommet construite grâce au modèle de la surface comme une variété union de triangles, $\text{nor}(s) \in \mathcal{S}^3$ (vecteur unitaire de la sphère unité) représenté par ses trois coordonnées réelles (section 2.2.3).
- **La courbure :** La courbure du sommet construite à partir des normales et des sommets voisins, $\text{curv}(s) \in [-1; 1]$ (section 2.2.3).

Pour chacune de ces caractéristiques une mesure de similarité normalisée dans l'intervalle $[0, 1]$ est construite pour chaque couple de valeur. Concernant la position, la normale, et la courbure, cette mesure normalisée est définie à partir d'une différence $\Delta$ entre deux valeurs, un seuil de tolérance $\tau$, ainsi qu'une fonction de normalisation $\sigma$ représentée dans la figure 7 qui transforme le quotient $\frac{\Delta}{\tau}$ de la différence des valeurs par le seuil de tolérance en une valeur normalisée entre 0 et 1 :

$$\sigma_{\text{pos}}(x,y) = \sigma\left(\frac{\Delta_{\text{pos}}(x,y)}{\tau_{\text{pos}}}\right) \qquad \text{avec } \Delta_{\text{pos}}(x,y) = \|\,\text{pos}(x) - \text{pos}(y)\|$$

$$\sigma_{\text{nor}}(x,y) = \sigma\left(\frac{\Delta_{\text{nor}}(x,y)}{\tau_{\text{nor}}}\right) \qquad \text{avec } \Delta_{\text{nor}}(x,y) = \text{angle}(\text{nor}(x), \text{nor}(y))$$

$$\sigma_{\text{curv}}(x,y) = \sigma\left(\frac{\Delta_{\text{curv}}(x,y)}{\tau_{\text{curv}}}\right) \qquad \text{avec } \Delta_{\text{curv}}(x,y) = |\,\text{curv}(x) - \text{curv}(y)|$$

La mesure de similarité normalisée entre deux valeurs de typage des sommets et directement définie par une matrice considérant l'ensemble des couples possibles. Cette matrice est définie de manière interne à Bionext SA [3] pour l'ensemble des valeurs possibles.

**Score d'une composante** Pour chaque caractéristique $k$, on définit un score de composante comme une fonction qui associe à un couple de sommets $(q, t)$ un triplet $(\text{ws}_k, \text{w}_k, \text{ns}_k)$ constitué respectivement d'un score pondéré $\text{ws}_k(q,t)$ d'un poids $\text{w}_k(q,t)$ et d'un score normalisé $\text{ns}_k(q,t)$. Le score normalisé ns correspond à la mesure de similarité des caractéristiques précédemment décrites, le poids w est déterminé en paramètre et le score pondéré ws vérifie les propriétés suivantes :

$$\begin{aligned} \text{ws}_k(q,t) &= \text{w}_k(q,t) \times \text{ns}_k(q,t) \\ \text{w}_k(q,t) &\in [0, \infty[ \\ \text{ns}_k(q,t) &\in [0, 1] \end{aligned}$$

3. Scores entre types atomiques internes à Bionext SA (développés par Pascal Muller pascal.muller@bionext.com).

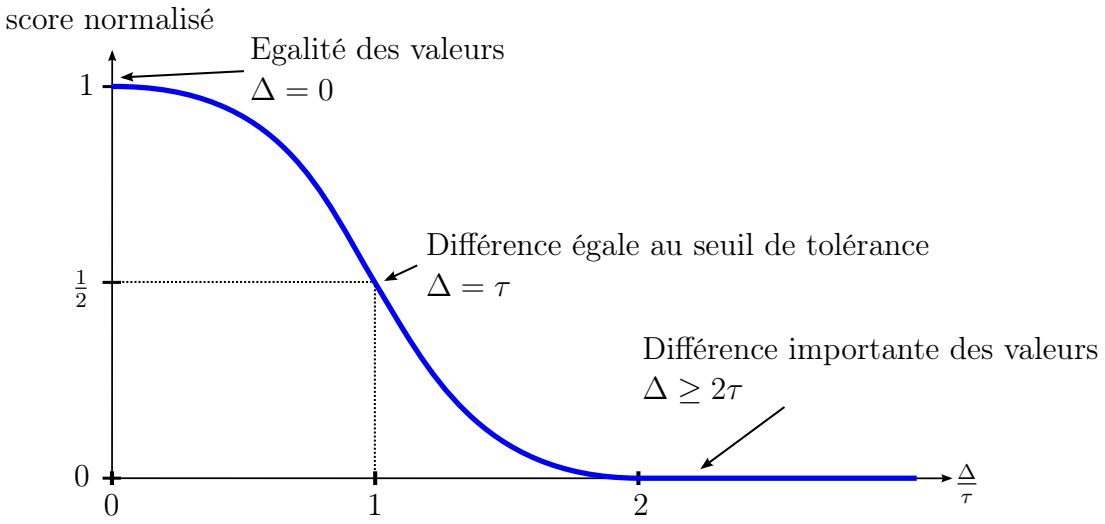



FIGURE 7 – Fonction de normalisation $\sigma$. La différence $\Delta$ entre deux valeurs d'une caractéristique est comparée au seuil de tolérance $\tau$, et la sigmoïde permet de définir une valeur de score entre 0 et 1.

Selon les choix effectués dans les paramètres des scores de composante, trois propriétés optionnelles peuvent être vérifiées. En particulier elles le sont dans l'implémentation par défaut de notre programme, mais elle ne sont pas indispensables pour les calculs qui suivent :

$$\mathrm{ns}_k(q,q) \geq \mathrm{ns}_k(q,t) \tag{III.1}$$
$$\mathrm{ns}_k(q,t) = \mathrm{ns}_k(t,q) \tag{III.2}$$
$$\mathrm{ns}_k(q,q) = 1 \tag{III.3}$$

**Influence** Une fonction d'influence associe à deux atomes $q$ et $t$ une valeur $\inf(q,t) \in [0,1]$. L'objectif est de modéliser l'importance d'une paire donnée de sommets dans un appariement global entre deux régions. Aucune propriété spécifique n'est requise pour la suite des calculs, cependant les trois propriétés suivantes optionnelles modélisent de manière naturelle une évaluation de l'influence comme l'inverse d'une distance, et sont également vérifiées dans notre implémentation par défaut :

$$\inf(q,q) = 1 \tag{III.4}$$
$$\inf(q,q) \geq \inf(q,t) \tag{III.5}$$
$$\|q-t\| \leq \|q-t'\| \Rightarrow \inf(q,t) \geq \inf(q,t') \tag{III.6}$$

Exemples : La fonction constante $\forall x,y,\ \inf(x,y) = 1$ qui ne privilégie aucun couple de sommets. Ou la fonction seuil $\inf_x(q,t) = 1$ si $\|q-t\| \leq x$ et $\inf_x = 0$ sinon, qui revient à ignorer les couples de sommets trop éloignés.

**Score entre deux sommets** On considère un ensemble de composantes $K$ et une fonction influence inf, on définit alors le score entre deux sommets comme le triplet $(\mathrm{ws}, \mathrm{w}, \mathrm{ns})$ :

$$\mathrm{ws}(q,t) = \inf(q,t) \sum_{k \in K} (\mathrm{ws}_k(q,t))$$
$$\mathrm{w}(q,t) = \sum_{k \in K} (\mathrm{w}_k(q,t))$$
$$\mathrm{ns}(q,t) = \frac{\mathrm{ws}(q,t)}{\mathrm{w}(q,t)}$$

On vérifie que les propriétés suivantes sont vérifiées :

$$\mathrm{ws}(q,t) = \mathrm{w}(q,t) \times \mathrm{ns}(q,t) \tag{III.7}$$
$$\mathrm{w}(q,t) \in [0,\infty[ \tag{III.8}$$
$$\mathrm{ns}(q,t) \in [0,1] \tag{III.9}$$

De plus si les équations III.1, III.2, et III.3 sont vérifiées pour toutes les composantes, et que les équations III.4, et III.5 sont vérifiées pour l'influence, alors on a aussi :

$$\mathrm{ns}(q,q) \geq \mathrm{ns}(q,t) \tag{III.10}$$
$$\mathrm{ns}(q,t) = \mathrm{ns}(t,q) \tag{III.11}$$
$$\mathrm{ns}(q,q) = 1 \tag{III.12}$$

### 3.2.2 Appariement et score entre deux régions

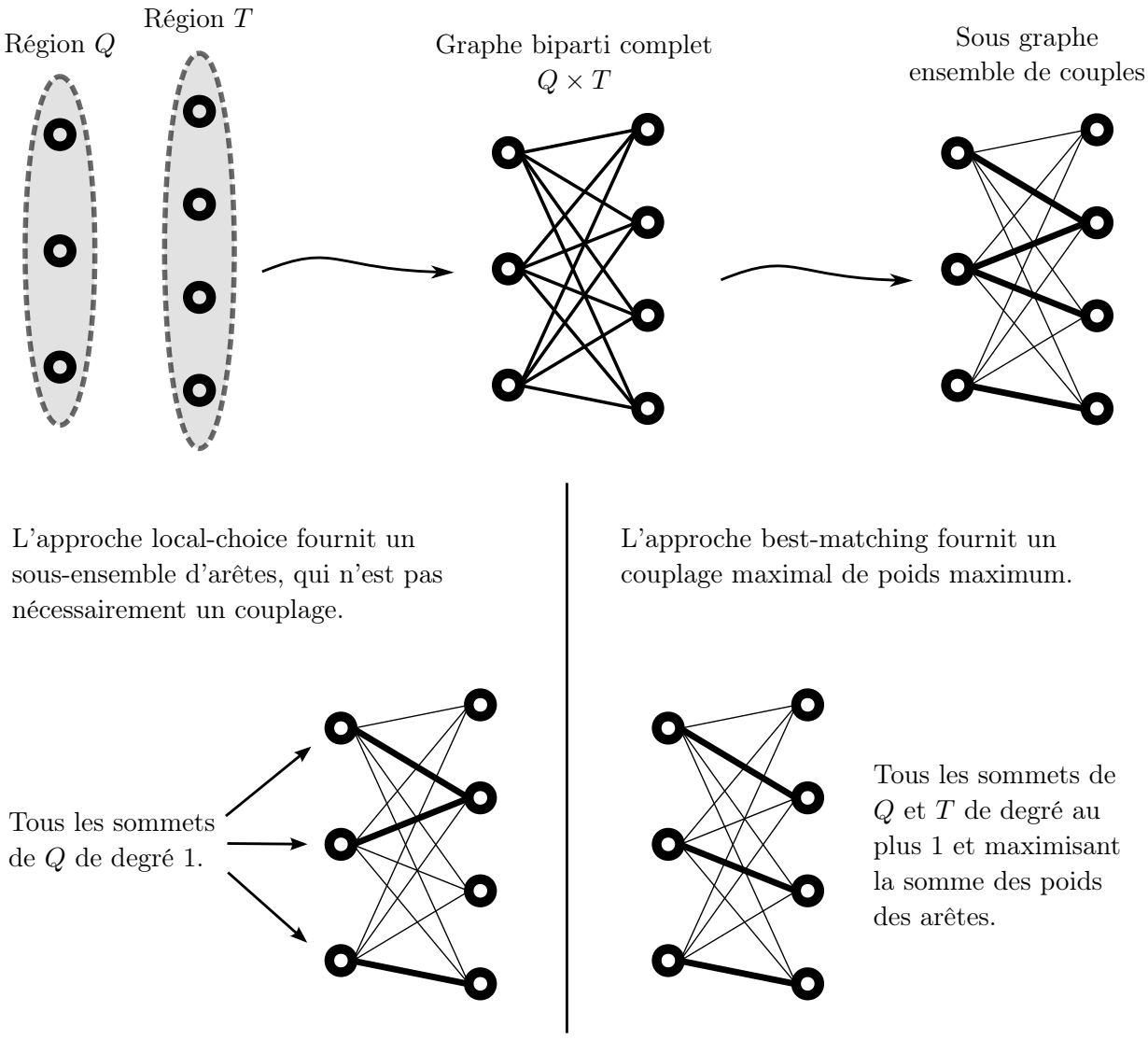



Figure 8 – Mise en correspondance de sommets dans la mesure de similarité. Un graphe biparti complet est établi où chaque arête est valuée par une mesure de similarité entre deux sommets. Un sous graphe est alors déterminé en sélectionnant une partie de ces arêtes.

On considère toujours une région requête $Q$ et une région candidate $T$, dont les sommets sont respectivement $\{q_i\}_{i=1}^n$ et $\{t_j\}_{j=1}^m$. La méthode de score est basée sur une mise en correspondance $A \subset Q \times T$ entre des sommets de $Q$ et des sommets de $T$. Comme illustré dans la figure 8, on considère le graphe biparti complet entre les sommets de $Q$ et les sommets de $T$ pondérés par la fonction de score entre sommets comme définie précédemment, plus précisément le poids $p$ des arêtes correspond au score pondéré $p(q,t) = \mathrm{ws}(q,t)$ de l'équation III.7, ensuite un sous-ensemble des arêtes de ce graphe biparti est sélectionné.

Pour déterminer cette sélection deux approches ont été étudiées. Une première méthode, local-choice, consiste à choisir le meilleur sommet de $T$ pour chaque sommet de $Q$. Pour chaque sommet $q$ de $T$, il s'agit de choisir le sommet $t$ de $T$ qui maximise le poids $p(q,t)$. En particulier la complexité de l'algorithme est quadratique $\mathcal{O}(\#Q \times \#T)$ correspondant à la complexité de la construction du graphe biparti des scores entre sommets en considérant l'évaluation d'un score entre deux sommets comme une opération élémentaire. Une seconde méthode best-matching consiste à déterminer le couplage de poids maximal, c'est-à-dire un couplage qui maximise la somme des poids. On utilise pour cela l'algorithme dit hongrois ou de Kuhn-Munkres [Kuhn 2010] de complexité cubique $\mathcal{O}(\max(\#Q, \#T)^3)$. Les deux approches se formalisent de la façon suivante :

$$\begin{aligned}
\text{local-choice}(Q,T,p) &= \{(q,t) : q \in Q,\ t \in T \text{ qui maximise } p(q,t)\} \\
\text{best-matching}(Q,T,p) &= A \subset Q \times T \text{ le couplage qui maximise } p(A) = \sum_{(q,t)\in A} p(q,t)
\end{aligned}$$

On note que la première approche local-choice ne fournit en général pas un couplage car un sommet de $T$ peut être associé à plusieurs sommets de $Q$. En revanche chaque sommet de $Q$ est associé à exactement un sommet de $T$, en particulier il y a donc exactement $\#Q$ paires. La seconde méthode propose effectivement un couplage maximal, c'est à dire exactement $\min(\#Q, \#T)$ couples où chaque sommet apparaît au plus une fois. Dans le cas général si les régions n'ont pas le même nombre de sommets, certains sommets de la région la plus grande resteront non appariés.

Une fois l'ensemble $A \subset Q \times T$ déterminé, il s'agit de sommer les scores de chaque paire de sommets. Une première possibilité consiste à utiliser le score qui a permis de définir le poids dans l'approche d'appariement, mais il est également possible d'utiliser une différente fonction de score entre sommets. Cela signifie que le score entre deux sommets n'est pas nécessairement le même pour guider l'appariement (fonction poids) et pour évaluer le score final de la région. Il est par exemple possible de ne considérer que la position pour définir l'appariement, en introduisant les autres propriétés uniquement lors de l'évaluation du score entre les régions. Pour un appariement $A$ des sommets des régions $Q$ et $T$ on définit ainsi la fonction de score :

$$\mathrm{ws}(Q,T) = \sum_{(q,t)\in A} \mathrm{ws}(q,t) \tag{III.13}$$

$$\mathrm{w}(Q,T) = \sum_{(q,t)\in A} \mathrm{w}(q,t) \tag{III.14}$$

$$\mathrm{ns}(Q,T) = \frac{\mathrm{ws}(Q,T)}{\mathrm{w}(Q,T)} \tag{III.15}$$

C'est le score normalisé (III.15) qui est alors utilisé comme mesure de similarité pour le couple de régions superposées. On note une extension naturelle de la méthode qui consiste à réaliser une moyenne arithmétique classique entre plusieurs fonctions de scores, chacune basée sur un poids et une méthode d'appariement différents. De cette manière il est possible de considérer un appariement des sommets pour évaluer les caractéristiques géométriques différent de l'appariement permettant d'évaluer les caractéristiques chimiques.

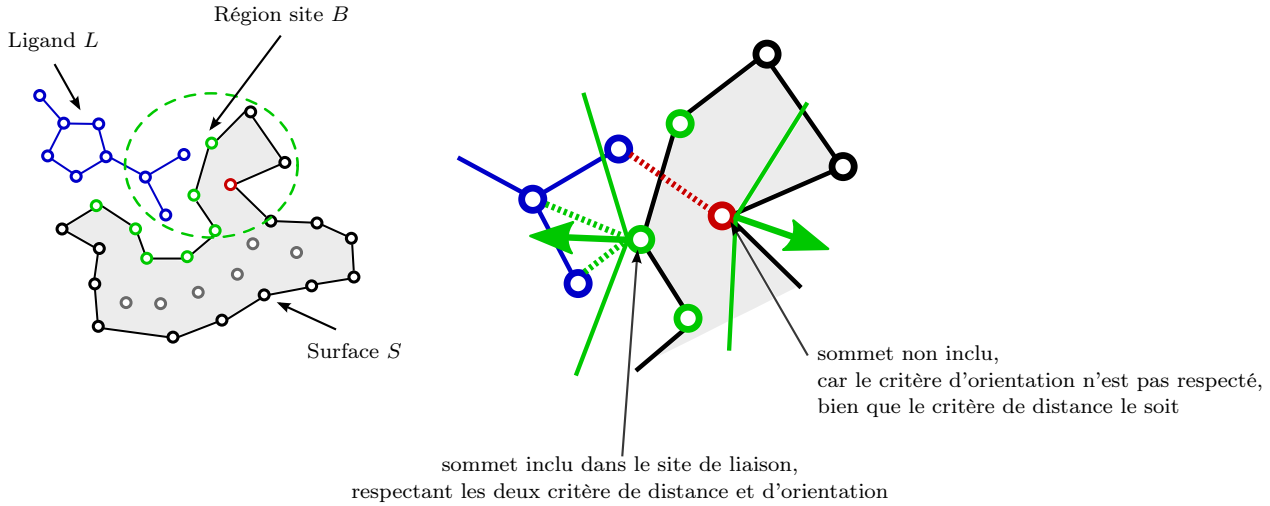



FIGURE 9 – Construction de la région site requête en fonction des paramètres de distance et d'orientation. Les cônes représentés sur deux des sommets schématisent la contrainte d'orientation de la normale vers les atomes du ligand.

# 4 Problème d'optimisation

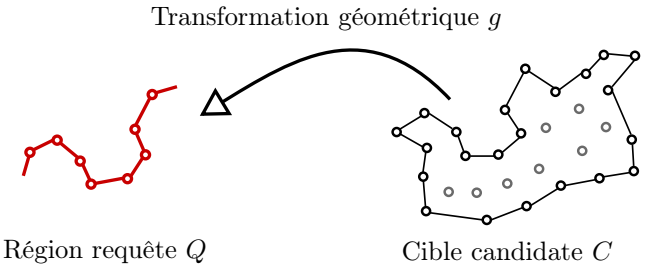
À partir du modèle des surfaces molécules, et de la donnée d'une mesure de similarité, il s'agit de déterminer la superposition du site requête sur une région de la cible candidate maximisant cette mesure de similarité. La méthode d'évaluation de similarité détaillée dans la section précédente est définie pour la donnée simultanée d'une région candidate sur la cible candidate et de la transformation géométrique définissant la superposition sur le site requête.

La recherche de motif se traduit alors comme un problème d'optimisation dont l'espace de recherche est constitué de l'ensemble des couples formés par la donnée d'une région de la cible candidate et d'une transformation géométrique, chaque transformation géométrique étant associée à la superposition induite. Une première technique pour réduire l'espace de recherche consiste à déterminer successivement la transformation géométrique, puis la région candidate par projection uniquement à partir de la superposition, selon une méthode de projection schématisée dans la figure 10. Cela permet d'éviter l'exploration simultanée de l'ensemble des régions candidates et des superpositions réalisables.

## 4.1 Construction de la région site requête

Le site de liaison requête est la région qui est le motif recherché par similarité par notre algorithme. Il peut être fourni manuellement par l'utilisateur ou bien à partir d'un complexe requête selon la construction automatique présentée dans la figure 9. Il est défini comme une région de surface de la cible requête. Afin de déterminer une région, il suffit de sélectionner un sous-ensemble des sommets de la surface de la cible requête (définition d'une région dans la section 2.3). Cette sélection est alors déterminée en fonction de la molécule ligand requête, suivant deux critères : la distance entre le sommet et un atome du ligand, et l'orientation de la normale du sommet par rapport à la direction vers l'atome du ligand.

Plus précisément, on considère un facteur de distance $d$ et un seuil d'orientation $c$. Si $S$ est la surface de la cible requête, et $L$ la molécule ligand requête, les sommets de la région site de liaison requête $B$ sont les sommets $s \in S$ pour lesquels il existe un atome $a \in L$ vérifiant à la fois :



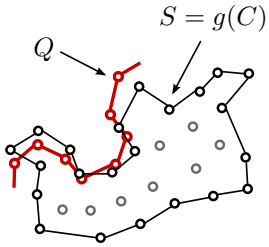
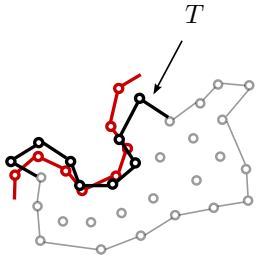
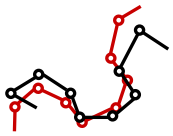




**Image**
Détermination de l'image $S$ de la cible candidate par la transformation géométrique.




**Projection**
Détermination de la région candidate $T$ par projection de la région requête $Q$.



**Score**
Evaluation de la superposition des régions $Q$ et $T$.

FIGURE 10 – Construction de la région candidate par projection de la région requête. Cette technique permet de définir une fonction objectif sur l'espace des transformations géométriques, la région candidate étant automatiquement déterminée.

$$\begin{aligned} \|\operatorname{pos}(a) - \operatorname{pos}(s)\| &\leq d \cdot (\operatorname{rad}(s) + \operatorname{rad}(a)) \\ c \cdot \|\operatorname{pos}(a) - \operatorname{pos}(s)\| &\leq \langle \operatorname{pos}(a) - \operatorname{pos}(s), \operatorname{nor}(s) \rangle \\ &\text{où } \langle \cdot, \cdot \rangle : \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R} \text{ est le produit scalaire} \end{aligned}$$

Par exemple en choisissant $d = 2$ et $c = 0$, on considère qu'un sommet de la surface de la cible fait partie du site de liaison s'il existe un atome du ligand à une distance inférieure au double de la somme des rayons Van der Waals, et situé dans le demi-espace délimité par le sommet dans la direction de sa normale.

## 4.2 Projection de la région requête en un site candidat

Pour un point de l'espace de recherche, c'est-à-dire pour une transformation géométrique $g$ donnée, on définit une fonction objectif. Cette fonction est construite comme un score de similarité entre deux régions

superposées : le site de liaison requête et le *site de liaison candidat*. Ce dernier site de liaison candidat ou plus simplement site candidat, correspond à une *projection* du site requête sur la cible candidate à partir de la transformation géométrique $g$ (figure 10).

Si $Q$ est la région site requête et $C$ la surface candidate, on considère $S = g(C)$ l'image de la cible candidate. Le site candidat $T$ est alors construit comme la région de l'image de la cible candidate $S$ contenant le même nombre de sommets que $Q$, et minimisant la distance entre chaque sommet de $T$ et l'un des sommets de $Q$. De manière plus précise, on associe à chaque sommet $t$ de la cible candidate une distance $d(t)$ qui est le minimum des distances euclidiennes entre $t$ et chaque sommet $q$ du site requête aligné $Q$. On définit alors $T$ par sélection des $\#Q$ sommets ayant les plus petites valeurs, où $\#Q$ désigne le nombre de sommets de $Q$.

## 4.3 Espace de recherche des transformations

L'espace de recherche du problème d'optimisation correspond à l'ensemble des superpositions du site requête sur la cible candidate, c'est-à-dire exactement l'espace des isométries directes de $\mathbb{R}^3$ qui peuvent agir sur le site requête. Cet espace est modélisé à l'aide des quaternions pour représenter les rotations et des vecteurs réels pour représenter les translations. Ce choix est guidé par la stabilité des opérations de composition entre les transformations, c'est-à-dire n'amplifiant pas les erreurs d'arrondis à chaque opération, qui sont nombreuses dans les algorithmes d'alignement qui seront utilisés.

L'espace de recherche est défini comme l'espace des transformations géométriques. C'est en particulier un espace en bijection avec $\mathcal{S}^3 \times \mathbb{R}^3$ où $\mathcal{S}^3$ est la sphère unité de dimension 3 plongée dans l'espace $\mathbb{R}^4$ qui représente l'ensemble des rotations et $\mathbb{R}^3$ représente l'ensemble des translations. Cet espace ne peut être exploré de manière exhaustive suivant une discrétisation raisonnable, y compris en bornant cet espace en se limitant aux transformations qui envoient le site requête dans un espace restreint contenant la cible candidate.

La structure de la surface permet l'utilisation de l'heuristique des régions circulaires, qui fera l'objet de la section suivante. Nous présentons d'abord le fonctionnement de la fonction objectif définie pour un alignement donné

### 4.3.1 Rotations et quaternions

On considère l'ensemble $\mathcal{R}$ des rotations vectorielles, la rotation d'axe porté par un vecteur non nul $v = (v_x, v_y, v_z) \in \mathbb{R}^3$ et d'angle $\theta \in \mathbb{R}/2\pi\mathbb{Z}$ étant notée $r_v^\theta$. Les rotations sont représentées par le groupe $\mathcal{H}$ des quaternions unitaires.

$$\mathcal{H} = \{(u, a, b, c) \in \mathbb{H} : u^2 + a^2 + b^2 + c^2 = 1\}$$

Il existe une bijection entre l'ensemble des rotations et l'ensemble des quaternions unitaires quotienté par la relation d'équivalence $\sim$ définie par $q_1 \sim q_2 \Leftrightarrow q_1 = \pm q_2$. Cette correspondance étant compatible avec les lois multiplicatives, c'est à dire que la composition des rotations correspond à la multiplication des quaternions, ou plus formellement la fonction $\pi$ suivante est un morphisme de groupe :

$$\begin{aligned}\pi : \mathcal{R} &\longrightarrow \mathcal{H}_{/\sim} \\ r_v^\theta &\longmapsto \left\{ \pm \left( \cos\left(\frac{\theta}{2}\right), v_x \sin\left(\frac{\theta}{2}\right), v_y \sin\left(\frac{\theta}{2}\right), v_z \sin\left(\frac{\theta}{2}\right) \right) \right\}\end{aligned}$$

$$\begin{aligned}\pi^{-1} : \mathcal{H} &\longrightarrow \mathcal{R} \\ (u,a,b,c) &\longmapsto r_v^\theta \quad \text{avec} \quad v = \frac{(a,b,c)}{\sqrt{1-u^2}}, \theta = 2\arccos(u) \\ -(u,a,b,c) &\longmapsto r_{-v}^{-\theta} = r_v^\theta \quad \text{donc } \pi^{-1} \text{ est bien défini sur le quotient } \mathcal{H}_{/\sim}\end{aligned}$$

$$\pi\left(r_{v_1}^{\theta_1} \circ r_{v_2}^{\theta_2}\right) = \pi\left(r_{v_1}^{\theta_1}\right) \cdot \pi\left(r_{v_2}^{\theta_2}\right)$$

### 4.3.2 Transformations

En notant $\mathcal{T}$ l'espace des translations, identifié avec l'ensemble $\mathbb{R}^3$ des vecteurs, on considère l'ensemble des transformations géométriques qui sont des isométries directes $\mathcal{G}$. Ces déplacements sont exactement l'ensemble des composées de rotations vectorielles et de translations, qui forment un groupe agissant sur $\mathbb{R}^3$.

$$\begin{aligned}\mathcal{H} \times \mathcal{T} &\simeq \mathcal{G} \\ (r_v^\theta, t) &\mapsto t \circ r_v^\theta\end{aligned} \qquad \begin{aligned}\circ : \mathcal{G} \times \mathcal{G} &\to \mathcal{G} \\ (g_1, g_2) &\mapsto t_1 \circ r_1 \circ t_2 \circ r_2 \\ &= (r_1 \circ r_2, r_1(t_2) + t_1)\end{aligned}$$

## 4.4 Méthode d'exploration des transformations

Afin de résoudre le problème de la recherche de la meilleure superposition, on propose une méthode générique d'exploration des transformations géométriques. Il s'agit essentiellement de décrire un arbre de transformations à explorer ainsi qu'une méthode de parcours.

### 4.4.1 Arbre des transformations

On considère une liste d'ensembles de transformations :

$$L = (L^1, L^2, \cdots, L^n)$$

$$\begin{aligned}L^1 &= \{g_1^1, g_2^1, \cdots, g_{m_1}^1\} \\ L^2 &= \{g_1^2, g_2^2, \cdots, g_{m_2}^2\} \\ &\vdots \\ L^n &= \{g_1^n, g_2^n, \cdots, g_{m_n}^n\}\end{aligned}$$

On construit un arbre enraciné, labellisé par des transformations géométriques (figure 11) :
— La racine est labellisée par l'identité.
— Pour $1 \leq p \leq n$, chaque nœud de profondeur $p-1$ est labellisé par une transformation $g$ a $m_p$ fils, chacun labellisé par $g_i^p \circ g$ pour $g_i^p$ dans $L^p$.

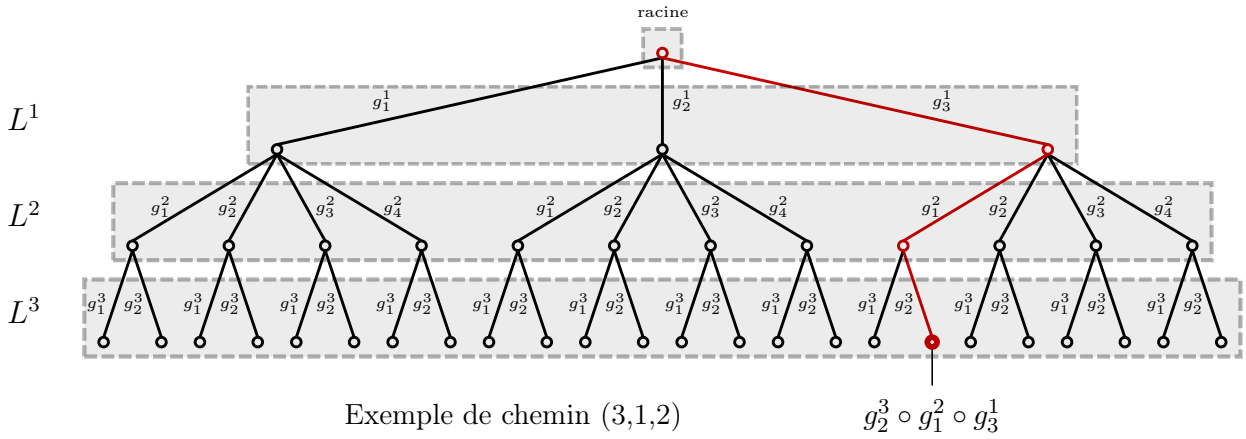



FIGURE 11 – Arbre de transformations géométriques. Les valeurs des sommets sont reportées sur les arêtes parentes pour améliorer la lisibilité.

### 4.4.2 Exploration de l'arbre

On suppose maintenant qu'on dispose d'une fonction $f : \mathcal{G} \to \mathbb{R}$ permettant d'évaluer chaque transformation selon la mesure de similarité définie dans la section précédente (3.2) pour la superposition induite par application de la transformation à la région candidate. Cette fonction permet ainsi d'attribuer un score à chaque nœud de l'arbre.

On définit une distance $d : \mathcal{G} \times \mathcal{G} \to \mathbb{R}_+$ qui mesure un écart entre différentes superpositions. Nous avons considéré deux approches pour cela. Une première approche consiste à mesurer la moyenne des distances entre les images des sommets par chacune des transformations, c'est-à-dire que la distance entre deux transformations $g_1$ et $g_2$ appliquées à une région $Q$ est définie comme $d(g_1, g_2) = \frac{1}{n} \sum_{q \in Q} \|g_1(q) - g_2(q)\|$. Une seconde approche consiste à définir un ensemble de points autour de $Q$, plus précisément un tétraèdre placé sur le centre de gravité de $Q$, avant d'utiliser la même mesure d'écart entre les images des sommets de ce tétraèdre. L'avantage de cette seconde méthode est de proposer un algorithme plus rapide car un nombre fixe d'images est à considérer, en revanche la dimension du tétraèdre doit être adapté à la dimension de la région pour apporter une mesure pertinente. Ainsi la seconde méthode est utilisée lorsque les tailles des régions sont fixées et que la dimension optimale du tétraèdre a pu être déterminé expérimentalement.

On propose alors un algorithme de parcours de cet arbre à partir des paramètres suivants définis pour chaque niveau de profondeur $1 \leq i \leq p$ :

- Un nombre maximal d'alignements à conserver $c_i \in \mathbb{N}$.
- Un seuil minimal de score $s_i \in \mathbb{R}$.
- Une distance minimale entre les alignements à conserver $d_i \in \mathbb{R}_+$.

L'algorithme d'alignement progressif multiple (algorithme 1) propose une méthode d'exploration de l'arbre des transformations. Le nombre de nœuds évalués à chaque niveau de profondeur $i$ est égal au cardinal de $T$ (ligne (1)) qui est exactement le cardinal de $R^{i-1}$ multiplié par $m_i$, or le cardinal de $R^{i-1}$ (ligne (2)) est majoré par $c_{i-1}$. Ainsi le nombre total d'évaluations peut être majoré par :

$$\mathcal{O}(\#\text{ évaluation de } f) = \sum_{i=1}^{p} c_{i-1} \times m_i \quad \text{en définissant } c_0 = 1.$$

### 4.4.3 Phases d'échantillonnage et d'ajustement

Deux phases essentielles se distinguent dans la méthode de superposition décrite. La première phase, l'échantillonnage consiste à déterminer $L^1$, c'est à dire à proposer des superpositions initiales. Toutes les

**Entrées :**
- $L$ arbre des transformations
- $f$ fonction d'évaluation du score de similarité
- $d$ fonction d'évaluation de distance entre superpositions
- $c_i$ nombre maximal de superpositions à conserver (profondeur $1 \leq i \leq p$)
- $s_i$ score minimal d'une superposition (profondeur $1 \leq i \leq p$)
- $d_i$ distance minimale entre deux superpositions (profondeur $1 \leq i \leq p$)

**Résultat :** ensemble de transformations feuilles de l'arbre $L$

$R^0 \leftarrow \{\text{id}\}$ ▷ la racine de l'arbre

**pour** $i \leftarrow 1, \cdots, p$ **faire**

(1) $T \leftarrow \emptyset$

**pour tous** $g \in R^{p-1}$ *et* $h \in L^p$ **faire**

$T \leftarrow T \cup \{h \circ g\}$ ▷ tous les fils du niveau précédant

**fin**

(2) $R^i \leftarrow \emptyset$

**pour tous** $t \in T$ *trié par* $f$ **faire**

**si** $|T^p| < c_i$ *et* $f(t) \geq s_i$ *et* $d(t, r) \geq d_i, \forall r \in R^i$ **alors**

$R^i \leftarrow R^i \cup \{t\}$ ▷ sélection selon les paramètres

**fin**

**fin**

**fin**

**retourner** $R^p$

**Algorithme 1 :** Alignement progressif multiple.

étapes suivantes $L^i, i > 1$ peuvent alors se voir comme un ajustement de la superposition proposée.

Les transformations choisies dans la phase d'ajustement dans notre implémentation consistent à explorer les rotations et translations dont les axes correspondent aux diagonales d'un cube centrée sur la région requête. Les amplitudes sont décroissantes à chaque niveau de l'arbre.

La première phase d'échantillonnage est en revanche plus complexe. Alors que la phase d'ajustement correspond à une discrétisation d'un ensemble de transformations de faible norme (en considérant par exemple les distance entre les images), la phase d'échantillonnage correspond *a priori* à une exploration de l'espace total de toutes les transformations géométriques. Une première idée consiste à déterminer dans un premier temps la translation qui superpose les centres de gravités, cependant l'ensemble des rotations centrées en ce points restent à explorer. Cela motive une heuristique pour un tel échantillonnage qui tire parti d'une régularité dans la forme de régions, dites circulaires, construites dans ce but.

# 5 Approche de résolution par les régions circulaires

## 5.1 Autre point de vue sur la projection

Le problème de la recherche de régions similaires à une région requête consiste à déterminer à la fois la meilleure région candidate et la meilleure superposition maximisant une mesure de similarité. Nous avons vu dans la section précédente qu'il est possible de déterminer la région candidate à partir de la superposition, ce qui permet de réduire l'espace de recherche à l'ensemble des transformations géométriques. Cet espace de recherche reste impossible à explorer de manière exhaustive.

Il est cependant possible de voir le problème suivant un second point de vue, schématisé dans la figure 12 et consistant cette fois à déduire la superposition à partir de la donnée de la région candidate. Dans ce cas de la même manière que la projection permettait de proposer une région candidate à partir de la superposition il est nécessaire de définir une méthode pour proposer une superposition à partir de la données d'un couple de régions requête et candidate. L'espace de recherche est ainsi ramené à l'exploration des régions candidates sur la surface de la cible candidate, cependant cela reste dans le cas général impraticable. En effet le problème consistant à échantillonner l'ensemble des régions candidates « de la même forme » que le site requête est un problème difficile à formaliser. Cela motive l'introduction d'un nouveau type de régions, les régions circulaires, qui ne modélisent pas directement un site, mais une partie de la surface approximant un disque géodésique pour lesquelles un échantillonnage exhaustif se définit naturellement.

## 5.2 Construction des régions circulaires

### 5.2.1 Définition des régions circulaires et propriétés géométriques

On considère une surface $S$ d'une molécule, un sommet $c \in S$, et un rayon $\lambda \in \mathbb{R}_+$. On définit la région $\mathrm{CR}_\lambda(c)$ comme la région contenant tous les sommets de $S$ qui sont à une distance de $c$ inférieure à $\lambda$. La distance est évaluée par somme des longueurs des arêtes du graphe de surface induit par $S$. La méthode de construction consiste à explorer l'arbre des plus courts chemins enracinée en $c$. On note en particulier qu'une telle région est toujours connexe, et le nombre de sommets dans la région peut varier pour un même rayon $\lambda$ fixé selon le sommet central $c$ choisi.

On considère un second type de région circulaire $\mathrm{CR}^n(c)$, définie par son centre $c$ et un nombre de sommets fixé $n$. Il s'agit alors de construire la région contenant les $n$ sommets les plus proches de $c$ en considérant toujours la distance définie par la somme des longueurs des arêtes. On note qu'une telle région n'est pas définie si la composante connexe de $S$ contenant $c$ est de cardinal strictement inférieur à $n$. Ce sont ces dernières régions de taille fixe (en nombre de sommets) qui sont utilisées dans BioBind, avec une valeur $n$ en paramètre du programme.

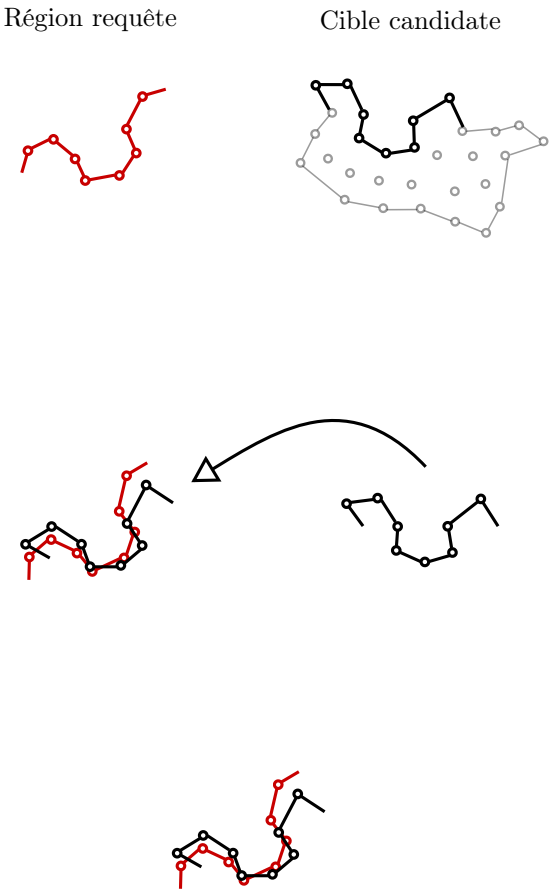



FIGURE 12 – Second point de vue : déterminer la superposition à partir de la région candidate.

### 5.2.2 Échantillonnage des régions circulaires

Ces régions sont échantillonnées en premier lieu sur le site de liaison requête, pour former l'ensemble des *régions circulaires requêtes*. Le choix des régions à sélectionner est guidé par une estimation de la qualité du recouvrement défini comme le pourcentage de sommets dans l'une au moins des régions par le nombre total de sommets dans le site de liaison.

Plus précisément en notant $B$ le site de liaison sur une surface $S$, en notant $c_i$ les centres sélectionnées, on définit la couverture :

$$\begin{aligned} \mathrm{cov} : \mathcal{P}(S) &\longrightarrow [0,1] \\ \{c_1, c_2, \cdots, c_q\} &\longmapsto \frac{|\bigcup_i \mathrm{CR}^n(c_i)|}{|B|} \end{aligned}$$

Il s'agit de déterminer un compromis entre le nombre de régions et la couverture. On note qu'il n'est pas toujours possible d'obtenir une couverture maximale car certains sommets du site de liaison peuvent se trouver dans des composantes connexes de $S$ où les régions circulaires de taille $n$ ne sont pas définies.

| | Couverture du site pour chaque stratégie | | | |
|---|---|---|---|---|
| Nombre de fragments | EXACT | CENTER | BORDER | AUTO |
| 1 | 39 | 39 | 36 | 39 |
| 2 | 68 | 66 | 66 | 68 |
| 3 | 83 | 80 | 77 | 81 |
| 4 | N/A | 87 | 86 | 88 |
| 5 | N/A | 91 | 88 | 92 |
| 6 | N/A | 93 | 90 | 94 |

Table 1 – Benchmark sur la stratégie de fragmentation du site requête. La couverture du site par l'union des fragments est présentée pour un nombre de fragments fixé entre 1 et 6. Il s'agit d'une moyenne des valeurs obtenues sur 6 sites de liaison de tailles et formes variées. La stratégie EXACT n'est testée que jusqu'à 3 fragments en raison du temps de calcul impraticable au delà. On observe que les stratégies CENTER et BORDER donnent des couvertures correctes au regard de la méthode exacte, et *a fortiori* notre méthode AUTO qui reste par construction systématiquement plus performante que les deux autres.

L'approche exacte permettant de déterminer les $q$ centres maximisant la couverture pour un nombre de région $q$ fixé, est très coûteuse en temps de calcul car il faut alors explorer les $q$ parmi $\#S$ choix possibles. Ainsi une heuristique a été développée, et validée sur un *benchmark* présenté dans la table 1 de sites de liaison de tailles et formes variées. Cet algorithme repose sur deux méthodes gloutonnes choisissant une région après l'autre pour maximiser soit la couverture partielle des régions sélectionnées, soit la couverture partielle favorisant les sommets du bord par une pondération dans le calcul d'une couverture modifiée. L'algorithme consiste à déterminer la meilleure couverture pouvant être obtenue en alternant ces deux méthodes gloutonnes, mais en limitant le nombre de changements de méthode afin de conserver une complexité globale polynomiale en la taille du site et le nombre de régions souhaitées.

L'échantillonnage sur la cible candidate consiste simplement à générer exhaustivement l'ensemble des régions circulaires centrées en chaque sommet de la surface, qui sont appelées *régions circulaires candidates*.

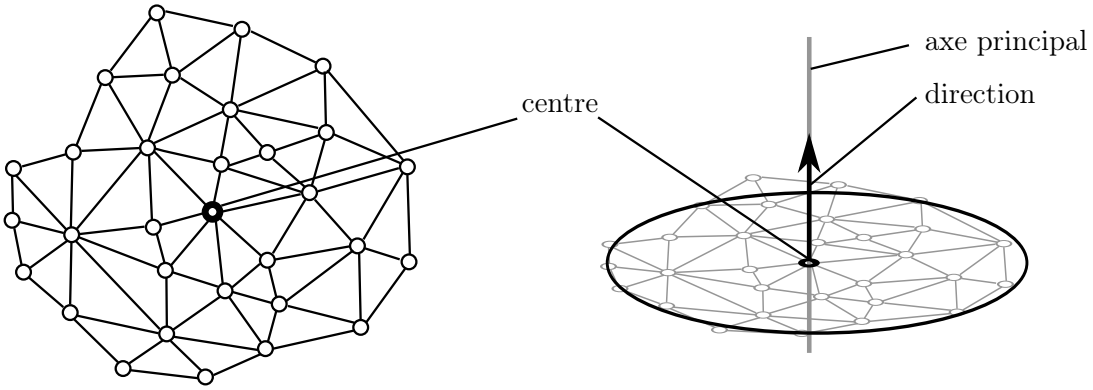



FIGURE 13 – Axe principal défini sur une région circulaire.

## 5.3 Méthode de superposition entre deux régions

### 5.3.1 Problème de comparaison de régions circulaires permettant de guider la résolution du problème initial

On définit un problème de comparaison de régions circulaires consistant à superposer chaque région circulaire requête avec chaque région circulaire candidate. Les contraintes géométriques placées sur les régions circulaires permettent de définir un problème de comparaison entre régions circulaires plus simple que la recherche directe du site requête ; la méthode de comparaison est précisée dans les paragraphes suivants. Nous montrerons ensuite comment l'information obtenue sur un couple de régions circulaires peut être étendue sur l'intégralité du site requête, afin de proposer des points pertinents de l'espace de recherche du problème initial.

### 5.3.2 Heuristique de l'axe principal

L'heuristique de l'*axe principal* consiste à associer à une région circulaire $P$ un *centre principal* main-center$(P) \in \mathbb{R}^3$ et une *direction principale* main-direction$(P) \in \mathcal{S}^3$ (vecteur unitaire) avec une propriété de stabilité vis-à-vis des superpositions pertinentes. C'est à dire que pour toute transformation $g$ qui induit une superposition d'une région $T$ sur une région $Q$ vérifiant une similarité géométrique, alors l'image de l'axe principal de $T$ est proche de l'axe principal de $Q$ (figure 13), ce qui signifie que les deux mesures suivantes doivent être minimisées :

— La distance entre les images des centres : $\|g(\text{main-center}(T)) - \text{main-center}(Q)\|$

— L'angle entre les images des directions : $(g(\text{main-direction}(T)), \text{main-direction}(Q))$

Deux choix naturels sont possibles pour définir le centre principal : le centroïde (centre de gravité) de la région, et le sommet ayant servi de centre pour la génération de la région circulaire. Après que ces deux possibilités aient été testées, la méthode du centroïde a été sélectionnée dans l'implémentation BIOBIND.

Plusieurs méthodes ont également été expérimentées pour définir la direction principale. Une première méthode consiste à effectuer une moyenne des normales des sommets de la régions, c'est la méthode implémentée dans BIOBIND. Une seconde méthode qui a été implémentée consiste à utiliser la normale du *meilleur plan contenant* la région, ou contenant la restriction au bord de la région. La détermination d'un tel plan est résolue en déterminant le plan minimisant la somme des distances au carré entre les sommets considérés et le plan, via une analyse des composantes principales [Wold 1987] réalisée par une décomposition en valeurs singulières [Klema 1980] d'une matrice déterminée à partir des coordonnées des points. Cette seconde méthode s'est cependant avérée moins performante sur nos jeux de données, pour son application à l'heuristique de superposition décrite dans la suite.

On considère une région circulaire $Q$ et une région circulaire $T$ que l'on souhaite superposer sur $Q$ afin d'évaluer la similarité avec une fonction $f$. L'heuristique de l'axe principal consiste à explorer uniquement

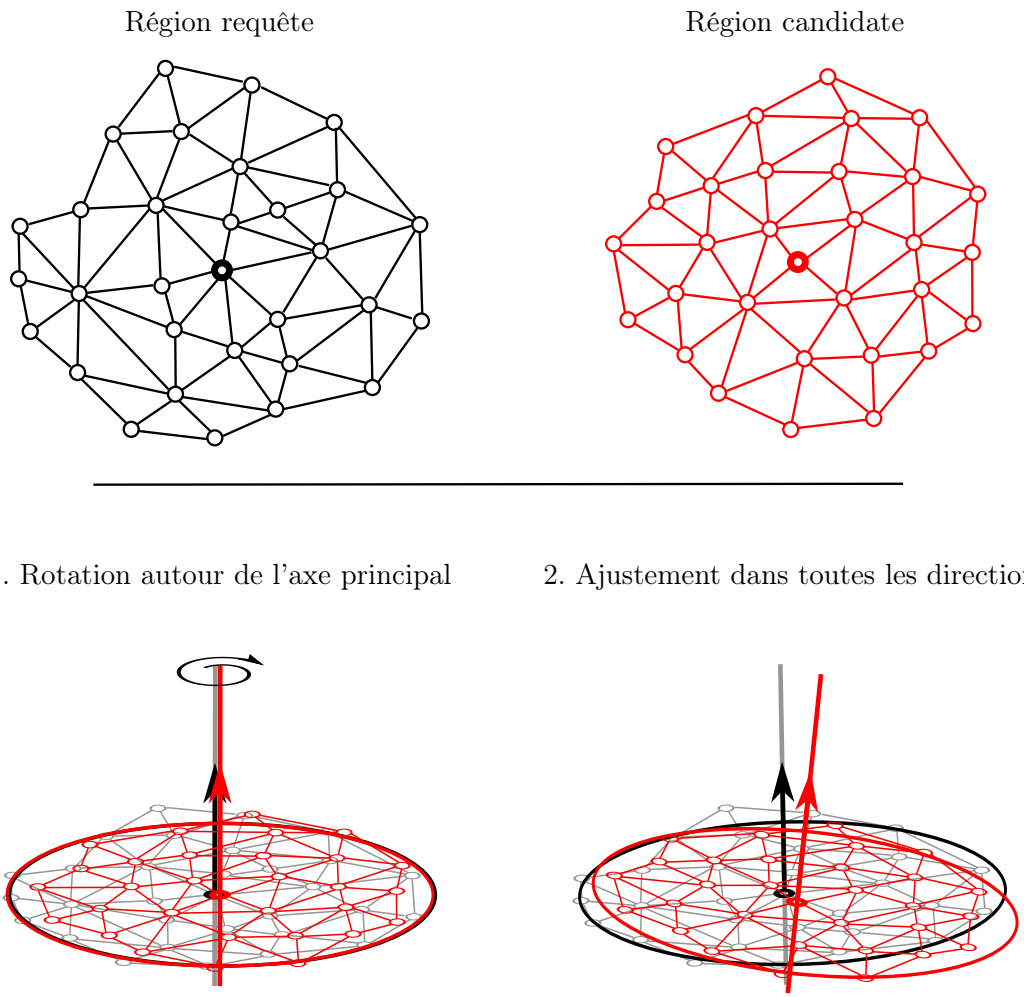



FIGURE 14 – Alignement en deux étapes : la rotation complète autour de l'axe principal puis un ajustement de plus faible amplitude dans toutes les directions.

les alignements qui envoient l'axe principal de $T$ à proximité de l'axe principal de $Q$ dans la première phase d'échantillonnage de notre algorithme d'exploration des transformations (section 4.4).

Pour cela on construit l'arbre d'exploration des alignements de la manière suivante (figure 14). Une première transformation géométrique $g$ qui envoie main-axis($T$) sur main-axis($Q$) est déterminée. Pour un pas angulaire $\alpha$ choisi, les rotations $g_k$ autour de cet axe sont échantillonnées, d'angles $0 \leq k\alpha < 2\pi$. Le premier niveau de profondeur de l'arbre, $L^1$, correspondant à ces rotations $g_k \circ g$.

Ensuite une ou plusieurs étapes d'ajustement sont construites en considérant pour un pas angulaire $\alpha$ et un pas scalaire $\lambda$ les rotations et translations autour des huit axes portés par les diagonales d'un cube centré en l'origine. A chaque profondeur successive, les normes $\alpha$ et $\lambda$ sont décroissantes.

### 5.3.3 Heuristique pour un filtrage rapide

L'objectif du filtrage avant la superposition de deux régions circulaires est de prédire quelles régions ne pourront pas être superposées de façon à donner un score suffisant. Pour ces couples de régions ainsi filtrées, l'alignement maximisant le score n'est pas calculé. On souhaite éliminer le plus de couples de régions possibles, tout en évitant d'éliminer celles qui sont effectivement similaires.

On propose une méthode de couplage entre les sommets selon les caractéristiques qui ne dépendent pas de la position (le type physico-chimique et la courbure). Ainsi un tel couplage maximal n'est pas toujours réalisable géométriquement, en revanche s'il existe un couplage réalisable géométriquement il le sera *a fortiori* à cette étape de filtrage. Cela garantit la sensibilité du filtrage.

Le principe du filtrage consiste donc à ne pas considérer les caractéristiques que sont la position et la normale des sommets afin de déterminer un appariement indépendant de la configuration spatiale des sommets du couple de régions. Cela signifie notamment qu'il n'y a aucune contrainte pouvant *a priori* limiter un sommet à être couplé avec tout autre sommet. Cependant l'heuristique de l'*axe principal* est à nouveau utilisée ici. En effet tout sommet d'une région peut se repérer par ses coordonnées suivant la hauteur de la projection sur cet axe, et suivant la distance à cet axe. L'intérêt de ces coordonnées est qu'elles restent invariantes à rotation près autour de cet axe. Cette heuristique est utilisée pour limiter l'espace des couplages réalisables, afin d'augmenter la spécificité du filtrage.

## 5.4 Recomposition d'un site candidat

### 5.4.1 Regroupement ou extension des paires de régions circulaires

À partir des couples de régions alignés, il s'agit d'utiliser la transformation géométrique qui est alors appliquée à l'ensemble du site de liaison requête, afin d'obtenir par projection un site candidat superposé. La mesure de similarité décrite en section 3.2 peut alors être appliquée. Cette première méthode, par *extension*, consiste donc simplement à utiliser la superposition trouvée sur un couple de régions circulaires.

Une seconde méthode, par *regroupement*, consiste à déterminer un ensemble de couples de régions circulaires $(Q_i, T_i)$ dont les transformations géométriques $g_i$ induisant les superpositions sont compatibles, c'est à dire que pour $i \neq j$ les images du site requête $g_i(Q)$ et $g_j(Q)$ sont suffisamment proches. La construction de ces ensembles compatibles est réalisée par une approche gloutonne en considérant les couples par ordre de scores décroissant, une transformation étant ajoutée à l'ensemble si le RMSD des images est suffisamment faible avec toutes les régions déjà prises dans l'ensemble. Une fois l'ensemble des transformations compatibles obtenu $g_1, \cdots, g_n$, on définit une *moyenne* des superpositions de la façon suivante. On considère d'une part le site de liaison requête copié $n$ fois, et d'autres part les $n$ images du site. À partir de ces deux listes de $n|B|$ points, on détermine la superposition qui minimise le RMSD entre les copies du site et les $n$ images du site par l'algorithme de Kabsch [Kabsch 1983].

### 5.4.2 Reconstitution du site candidat et score

Quelle que soit la méthode par regroupement ou extension à partir des couples résultats de régions circulaires, on obtient un superposition d'une ou plusieurs régions circulaires requêtes qui est transposée sur l'intégralité du site de liaison requête. Cette superposition constitue un point de notre espace de recherche du problème initial, permettant d'appliquer la fonction de score après projection du site candidat.

Une optimisation est réalisée sur cette superposition, en explorant un voisinage de la transformation géométrique proposée. On utilise pour cela la seconde phase de notre algorithme d'alignement progressif multiple décrit dans la section 4.4 en considérant uniquement un ajustement de la première transformation par exploration d'un nombre fini de superpositions. La valeur maximale du score ainsi obtenue est sélectionnée parmi toutes les évaluations de la fonction objectif réalisées pour définir le score attribué à la cible candidate. Cette valeur est suffisante pour trier les cibles et ainsi répondre au problème de classification binaire.

En plus de la donnée du score affecté à la cible candidate, la transformation géométrique réalisant le score maximum peut être appliquée sur le ligand requête pour former un *complexe résultat* constitué par l'image du ligand requête et la cible candidate. Ce complexe résultat est fourni à l'utilisateur afin de pouvoir être utilisé par une méthode complémentaire de prédiction d'interaction, par inspection visuelle ou à l'aide d'un logiciel.

# 6 Récapitulatif des différentes étapes successives

**Modèle des molécules.** Les molécules sont représentées comme des ensembles d'atomes, chacun étant annoté par des caractéristiques géométriques et physico-chimiques (section 2.1). Certaines caractéristiques sont présentes dans les fichiers fournis en entrée, d'autre sont récupérées selon des base de données de référence, comme le rayon de Van der Waals déterminé par le CCDC (ccdc.cam.ac.uk), et enfin un typage des atomes est défini selon un procédé interne à l'entreprise.

**Construction de la surface moléculaire.** À partir du modèle des molécules une triangulation est déterminée où chaque sommet est associé à un atome (section 2.2). Cette construction repose sur la théorie des formes alpha (section 2.2.1) ainsi qu'un algorithme développé afin d'obtenir une variété orienté (section 2.2.2), permettant de définir une notion de région de surface (section 2.3).

**Détermination du site requête.** Le site requête est une région de la surface requête (section 4.1). Les sommets de la région sont déterminés par un critère de distance et d'orientation vers les atomes du ligand requête au sein du complexe requête.

**Échantillonnage des régions circulaires.** Les régions circulaires requêtes sont sélectionnées afin de maximiser la couverture du site requête, à l'aide d'une heuristique polynomiale en le nombre de régions souhaitées (section 5.2.2), proche de la méthode exacte exponentielle impraticable. Les régions circulaires candidates sont toutes conservées (génération exhaustive, section 5.2.2).

**Filtrage des couples de régions circulaires.** Les couples de régions circulaires requête et candidate sont filtrées selon une évaluation d'une borne supérieure d'un score, en ne tenant pas compte des caractéristiques qui peuvent varier selon la superposition (section 5.3.3). L'implémentation consiste à utiliser un couplage de poids maximal des sommets selon leurs propriétés invariantes par transformation géométrique.

**Superposition des régions circulaires.** La superposition des régions circulaires est réalisée en utilisant l'algorithme d'alignement progressif multiple défini dans la section 4.4. Une première phase d'échantillonnage des superpositions utilise l'heuristique de l'axe principal (section 5.3.2), avant une seconde phase d'ajustement de la superposition maximisant le score de similarité.

**Score d'une région circulaire candidate.** La mesure de similarité définie dans la section 3.2 est utilisée.

**Recomposition d'un site candidat.** La recomposition d'un site (section 5.4) consiste à projeter le site requête sur l'image de la surface de la cible candidate. La transformation géométrique utilisée pour définir l'image de la surface candidate est déduite des superpositions des régions circulaires.

**Superposition des sites.** La superposition des sites utilise uniquement la seconde phase d'ajustement de notre algorithme d'alignement progressif multiple (section 4.4).

**Score d'un site candidat.** La mesure de similarité définie dans la section 3.2 est utilisée, de la même manière que pour les régions circulaires, cependant les paramètres définissant le rôle de chaque caractéristique et la méthode de choix des couples de sommets ont été optimisés indépendamment.

**Réponse au problème de classification binaire.** Les scores des sites candidats sont utilisés pour classer les cibles candidates correspondantes, et ainsi proposer une classification entre les cibles prédites par similarité avec le site de liaison requête et les autres par la donnée d'un rang seuil dans ce classement (section 1.1).

# 7 Conclusion

A partir de la donnée d'un complexe requête, l'algorithme BioBind permet d'une part de proposer des régions candidates et d'autre part d'en évaluer la similarité avec le site de liaison requête. Cette mesure de similarité a pour objectif d'évaluer la capacité de la cible candidate à lier le même ligand présent dans le complexe requête, reposant sur le principe d'inférence de l'interaction. La méthode de recherche à elle pour objectif de sélectionner les régions candidates ainsi que de proposer une superposition, afin de maximiser ce dernier score.

Le choix du modèle de la surface des molécules est central dans notre approche car il conditionne la possibilité de définir notre évaluation de la similarité par appariement de sommets entre deux régions de deux molécules. Les sommets appariés, qui contribuent à la valeur du score attribué sont associés à des atomes dits « accessibles » qui sont sélectionnés de manière exacte par l'usage de la théorie des formes alpha, en suivant le modèle du contact de Van der Waals avec une molécule d'eau provenant du milieu extérieur de la molécule. Ce choix du modèle de surface triangulée est également pré-requis par notre définition de région de surface utilisée à plusieurs niveaux de granularité : pour représenter un site complet mais également une fragmentation exhaustive en régions régulières.

L'espace de recherche du problème d'optimisation consistant à déterminer la meilleure région candidate pour la meilleure superposition maximisant la mesure de similarité s'avère impraticable à explorer de manière exhaustive. Une idée naturelle pour simplifier l'exploration d'un tel espace consiste à séparer l'espace des régions candidates de l'espace des superpositions afin de les déterminer successivement au lieu de maximiser directement notre fonction objectif de mesure de similarité sur le produit cartésien de ces espaces. Un premier point de vue revient alors à explorer l'espace des superpositions en déduisant la région candidate, alors que le second point de vue demande à explorer l'ensemble des régions candidates pour ensuite les superposer. Dans le cas général ces deux approches restent impraticables, que ce soit par la taille et l'absence de structure de l'espace des transformations géométriques ou de l'ensemble des régions candidates.

Afin de résoudre le problème de l'exploration de l'espace de recherche une heuristique est utilisée qui repose sur l'utilisation de régions dites circulaires. La structure régulière de ces régions permet de proposer des algorithmes efficaces pour d'une part échantillonner l'ensemble des régions sur une surface donnée et d'autre part superposer deux telles régions via un axe de rotation privilégié. Cette comparaison exhaustive entre des régions circulaires qui échantillonnent à la fois le site de liaison requête et la surface de la cible candidate permet de proposer des points d'entrée dans la recherche de la superposition globale du site de liaison. Cette étape de recomposition du site candidat permet d'appliquer notre mesure de similarité pour un sous-ensemble pertinent de régions candidates déterminées par notre heuristique.

On note par ailleurs que notre algorithme est fortement paramétré, avec par exemple de nombreux paramètres concernant la définition des régions, le calcul des scores de similarités entre les sommets, les approches d'appariement des sommets pour déterminer les scores, ou encore les différents algorithmes de superpositions. Ainsi pour le même algorithme, des paramètres différents peuvent avoir des conséquences très importantes en pratique sur d'une part la qualité des résultats et d'autre part le temps de calcul. En effet même si l'ordre de grandeur du temps de calcul de chaque partie de l'algorithme n'est pas modifié, notre algorithme est appliqué à l'échelle de la PDB, soit plusieurs centaines de milliers de macromolécules cibles. Ainsi le temps de calcul est un élément essentiel de l'approche et un facteur 100 peut par exemple rendre une approche non réaliste pour les problèmes réels. Certains compromis ont été effectués dans les paramètres en évaluant différents jeux de données afin notamment de proposer une version par défaut qui

est utilisée dans les validations présentées dans le chapitre suivant.

Chapitre IV

# Évaluation des performances de BioBind

## Introduction

La qualité de notre algorithme BioBind décrit dans le précédent chapitre est évaluée pour la problématique biologique de la prédiction de cibles secondaires. La formalisation de cette question comme un problème de classification binaire, c'est-à-dire la séparation des cibles candidates entre les positifs et négatifs selon un paramètre seuil, permet d'utiliser des métriques classiques qui sont adaptées.

Nous définissons trois jeux de données, les deux premiers sont issus de la littérature et le troisième a été conçu spécifiquement pour la validation de BioBind. Ce dernier jeu de données vise à corriger certains défauts, d'une part concernant le choix des ligands souvent redondants et peu pertinent pour la pharmacologie, et d'autre part dans la définition des cibles négatives plus diversifiées et représentatives. Notre algorithme se compare favorablement à deux autres approches, un première approche par similarité des cibles, ProBiS, et une seconde par docking, Vina.

## Sommaire

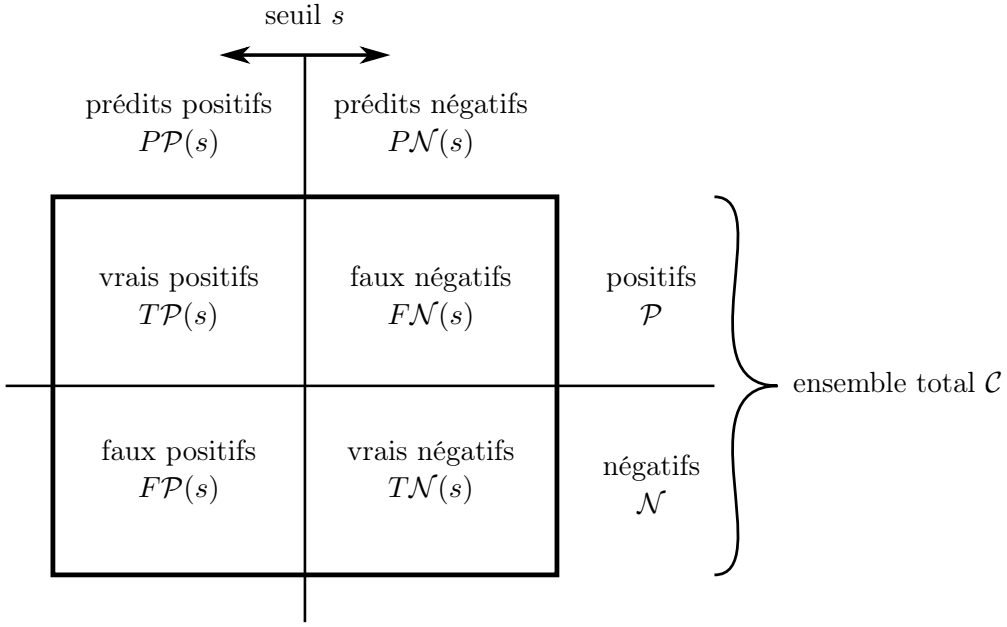



FIGURE 1 – Partition de l'ensemble des résultats, positifs et négatifs, suivant le rang choisi comme seuil entre les prédits positifs et les prédits négatifs.

# 1 Méthodes pour évaluer et comparer différentes approches

Une approche de prédiction de cibles peut être vue comme un classificateur binaire sur un ensemble de cibles candidates. Cela permet d'utiliser des métriques usuelles qui seront adaptées pour mesurer de manière plus pertinente la qualité d'un résultat tel qu'il est destiné à être lu par l'utilisateur.

## 1.1 Sensibilité et spécificité du classificateur binaire

### 1.1.1 La prédiction de cibles comme un classificateur

On considère une instance du problème de la recherche de cibles, constituée d'une molécule ligand requête $L$, et d'un ensemble de macromolécules cibles candidates $\mathcal{C}$. Les méthodes de prédiction de cibles fonctionnent généralement en fournissant une liste triée des cibles de $\mathcal{C}$, suivant l'affinité prédite avec le ligand requête. On peut alors utiliser un rang seuil $s$ qui partitionne $\mathcal{C} = P\mathcal{P}(s) \cup P\mathcal{N}(s)$ en deux ensembles : les résultats *prédits positifs* $P\mathcal{P}(s)$ dont le rang est inférieur à $s$, et les résultats *prédits négatifs* $P\mathcal{N}(s)$ dont le rang est supérieur à $s$. On obtient ainsi un nombre de cibles prédites positives $\#P\mathcal{P}(s)$ égal à la valeur du rang seuil $s$ choisi.

On suppose maintenant qu'on connaît le résultat souhaité, c'est à dire une partition $\mathcal{C} = \mathcal{P} \cup \mathcal{N}$ telle que les cibles de $\mathcal{P}$ sont exactement celles qui peuvent former un complexe avec le ligand. On dit que $\mathcal{P}$ sont les résultats *positifs* et $\mathcal{N}$ sont les *négatifs*. On définit alors pour un rang seuil $s$ les ensembles suivants (détaillés en figure 1) :

- $T\mathcal{P}(s) = P\mathcal{P}(s) \cap \mathcal{P}$, les *vrais positifs*
- $T\mathcal{N}(s) = P\mathcal{N}(s) \cap \mathcal{N}$, les *vrais négatifs*
- $F\mathcal{P}(s) = P\mathcal{P}(s) \cap \mathcal{N}$, les *faux positifs*
- $F\mathcal{N}(s) = P\mathcal{N}(s) \cap \mathcal{P}$, les *faux négatifs*

À partir de ces ensembles et pour un rang seuil $s$ donné, on définit la *sensibilité* (ou *True Positive Rate*) d'une approche comme le quotient des vrais positifs par les positifs. On définit de même la *spécificité*

(ou *True Negative Rate*) comme le quotient des vrais négatifs par les négatifs. La *précision* (ou *Positive Predictive Value*) est définie comme le quotient des vrais positifs par les prédits positifs.

$$\begin{aligned}
\text{TPR}(s) &= \frac{\#\mathcal{TP}(s)}{\#\mathcal{P}} = \frac{\#\mathcal{TP}(s)}{\#\mathcal{TP}(s) + \#\mathcal{FN}(s)} && \in [0,1] && \text{la sensibilité} \\
\text{TNR}(s) &= \frac{\#\mathcal{TN}(s)}{\#\mathcal{N}} = \frac{\#\mathcal{TN}(s)}{\#\mathcal{TN}(s) + \#\mathcal{FP}(s)} && \in [0,1] && \text{la spécificité} \\
\text{PPV}(s) &= \frac{\#\mathcal{TP}(s)}{\#\mathcal{PP}(s)} && \in [0,1] && \text{la précision}
\end{aligned}$$

### 1.1.2 Courbe ROC, évaluation du classificateur

Le fait de considérer les approches de prédiction de cibles comme des classificateurs binaires permet d'utiliser des méthodes classiques comme les courbes ROC et les aires AUC [Carvalho 2014]. Ces mesures permettent d'évaluer la qualité des résultats d'un classificateur binaire, fournissant ainsi un moyen de comparer plusieurs classificateurs entre eux sur une même instance du problème de prédiction de cible.

La courbe ROC (ou *Receiver Operating Characteristic*, figure 2) est définie comme la sensibilité en fonction de la spécificité. Elle est constituée de l'ensemble des points $(\text{TPR}(s), 1 - \text{TNR}(s))$ pour toutes les valeurs de seuils $s$ possibles $0 \leq s \leq \#\mathcal{C}$. En particulier le point $(0,0)$ est toujours présent pour le rang seuil $s = 0$ toutes les cibles sont prédites négatives, il n'y a donc aucun vrai positif et tous les vrais négatifs. Réciproquement le point $(1,1)$ est également toujours présent, car pour le rang seuil maximal $s = \#\mathcal{C}$ toutes les cibles sont prédites positives, il n'y a donc aucun vrai négatif et tous les vrais positifs.

L'AUC (ou *Area Under Curve*), définie comme l'aire sous la courbe ROC, constitue une métrique globale du classificateur sur l'ensemble des cibles candidates. Il s'agit par définition d'une partie du carré unité, donc toujours comprise entre 0 et 1. La valeur 1 correspondant au classificateur idéal pour lequel toutes les cibles positives sont classées avant toute autre cible négative. L'AUC peut aussi s'interpréter comme la probabilité qu'un positif choisi au hasard soit mieux classé qu'un négatif choisi au hasard, en particulier la valeur 0.5 correspond à l'espérance d'un classificateur aléatoire.

Un cas particulier est à considérer si une approche de prédiction de cible ne fournit pas un classement complet. C'est-à-dire qu'un rang n'est attribué que pour une partie de l'ensemble des cibles, la tête du classement résultat. Ce problème est contourné en affectant des rangs arbitrairement mauvais aux cibles positives qui ne sont pas classés, afin de pouvoir définir systématiquement les différentes métriques. Cela introduit un biais dans les mesures globales comme l'AUC mais nous verrons que les mesures pertinentes sont justement restreintes à la tête du classement, où ce biais n'a pas d'influence.

### 1.1.3 Adaptation de la métrique pour évaluer la tête du classement

La figure 3 présente une adaptation de la métrique. En effet la mesure de l'AUC permet d'évaluer globalement un classificateur, pour l'ensemble des choix de rang seuils possibles en couvrant toutes les valeurs de sensibilité et spécificité. Cependant une méthode de prédiction de cibles, telle que BioBind, est conçue dans l'objectif de pouvoir déterminer les cibles d'un ligand, parmi un *très grand* ensemble de cibles candidates. L'objectif est de déterminer un ensemble de cibles potentielles suffisamment petit, sur lequel de nouvelles expériences plus précises pourront être réalisées. Ainsi, l'intérêt se situe dans les premiers vrais positifs, et la sensibilité ou spécificité pour des valeurs de seuil trop grandes est une mesure peu pertinente de la qualité de l'approche. Une nouvelle mesure est construite pour prendre en compte uniquement les classifications produites par les rangs seuil suffisamment petits, afin de rendre mieux compte du résultat attendu pour un problème réel.

Pour une spécificité minimale de $X\%$, on définit le rang seuil $s_X$ pour lequel la spécificité reste au dessus de $X\%$, correspondant à une lecture plus réaliste où seuls les premiers résultats sont considérés.

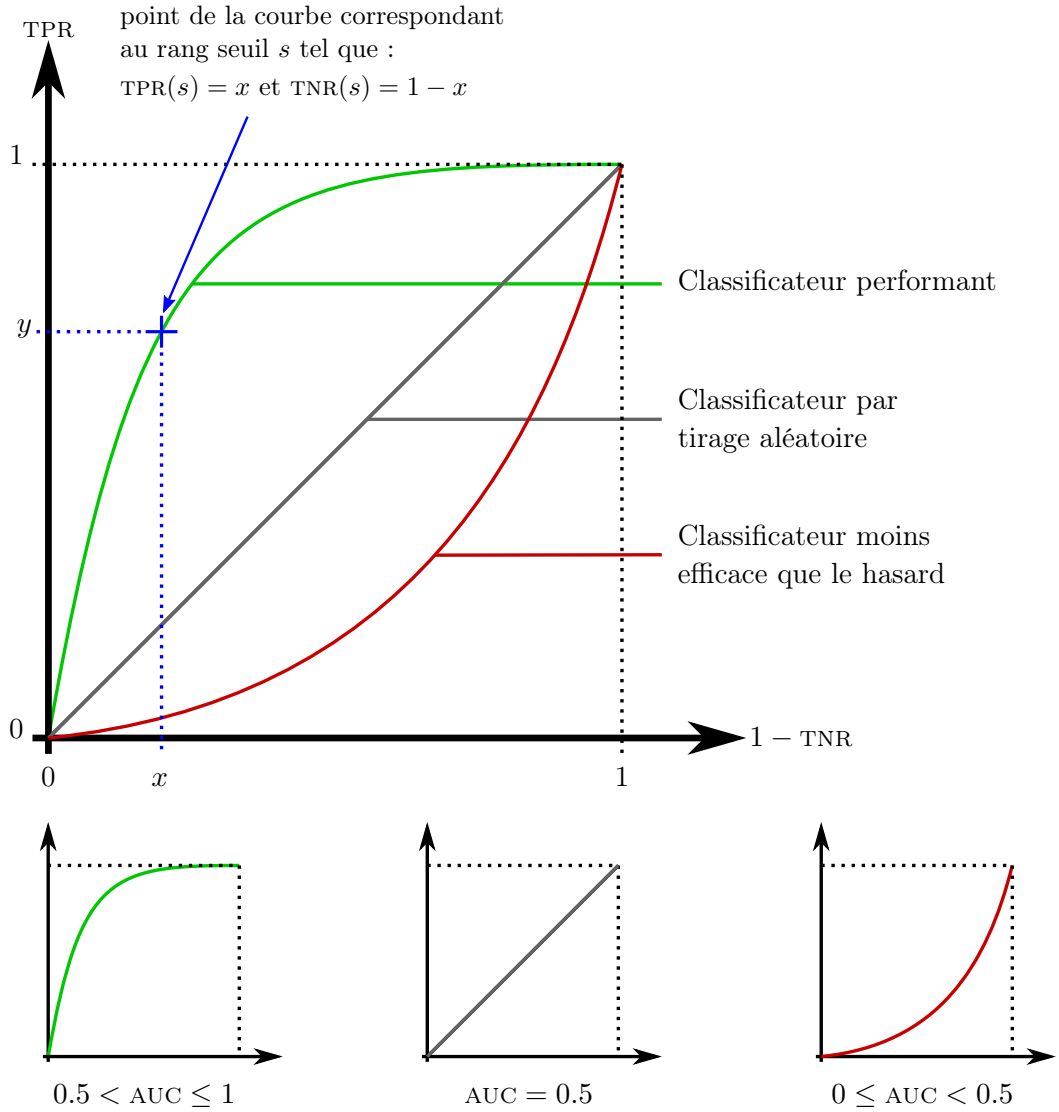



FIGURE 2 – Évaluation d'un classificateur par sa courbe ROC et l'aire sous la courbe AUC. Une aire de 0.5 correspond à l'espérance d'un tirage aléatoire, ainsi pour être pertinent un classificateur doit produire une AUC supérieure.

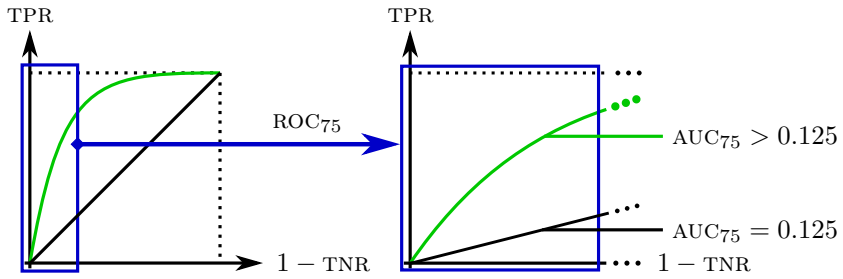



FIGURE 3 – Courbe ROC restreinte à une spécificité supérieure à 75 %. On note que même si l'$\text{AUC}_X$ est normalisée dans l'intervalle $[0, 1]$, dans ce contexte l'AUC de référence d'un classificateur aléatoire n'est plus 0.5 mais $\frac{100-X}{200}$, soit 0.125 pour $X = 75$.

On définit également les paramètres $\text{PPV}_X, \text{TPR}_X, \text{TNR}_X$ en fonction du seuil, ainsi que la courbe $\text{ROC}_X$ et son aire associée $\text{AUC}_X$.

$$
\begin{aligned}
s_X &= \max\{s : \text{TNR}(s) \geq X\%\} & \text{PPV}_X &= \text{PPV}(s_X) \\
\text{ROC}_X &= \{(\text{TPR}(s), 1 - \text{TNR}(s)) : s \in \mathbb{N}, \text{TNR}(s) \geq X\%\} & \text{TPR}_X &= \text{TPR}(s_X) \\
\text{AUC}_X &= \int \text{ROC}_X \times 100/(100 - X) & \text{TNR}_X &= \text{TNR}(s_X)
\end{aligned}
$$

Afin d'illustrer la nécessité d'une telle mesure focalisée sur les premiers rangs, on peut citer l'exemple d'un classificateur qui récupère la moitié des cibles positives très bien classées, et l'autre moitié très mal classées. Un tel classificateur aurait une AUC proche de 0.5, qui ne le distingue pas d'un classificateur aléatoire. Cependant un tel classificateur a un intérêt réel dans la pratique car il apporte une information dès les tous premiers éléments du classement. Cette situation est courante avec les différentes méthodes de prédiction de cibles, c'est pourquoi nous privilégierons les $\text{AUC}_{75}$ ou $\text{AUC}_{90}$ pour évaluer les différents résultats. La figure 4 schématise un tel exemple.

## 1.2 Métrique définie sur une union d'instances

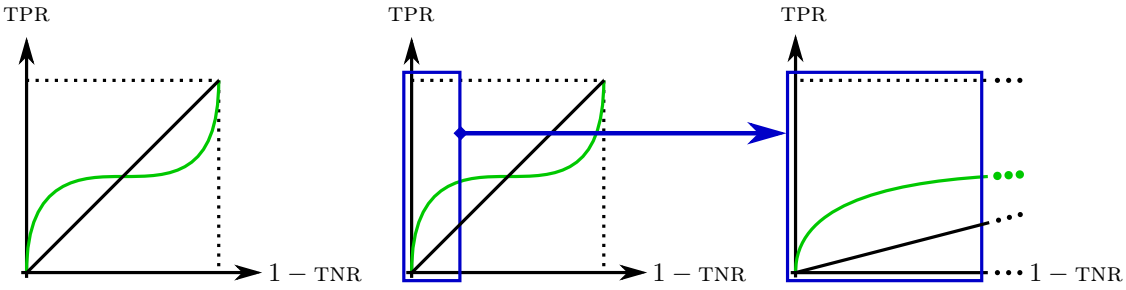
Le problème de prédiction de cibles est défini pour un ligand et un ensemble de cibles candidates. Afin d'évaluer une approche, il est nécessaire de réaliser plusieurs expériences, en considérant plusieurs problèmes de prédiction différents. On considère notamment plusieurs cibles requêtes pour chaque ligand et de manière générale plusieurs ligands requêtes. Afin d'évaluer simultanément les résultats de plusieurs instances, on définit une notion d'*union d'instances du problème de classification binaire* sur laquelle les métriques sont adaptées. Un exemple de cette construction est présenté en figure 5.

On considère la donnée de $n$ instances du problème de prédiction, pour les ligands $\{L_i\}_{i=1}^n$, les cibles positives étant respectivement $\{\mathcal{P}_i\}_{i=1}^n$, et les cibles négatives étant $\{\mathcal{N}_i\}_{i=1}^n$. Les espaces de recherches sont donc les $\mathcal{C}_i = \mathcal{P}_i \cup \mathcal{N}_i$. On définit alors l'union des instance du problème de classification binaire par :

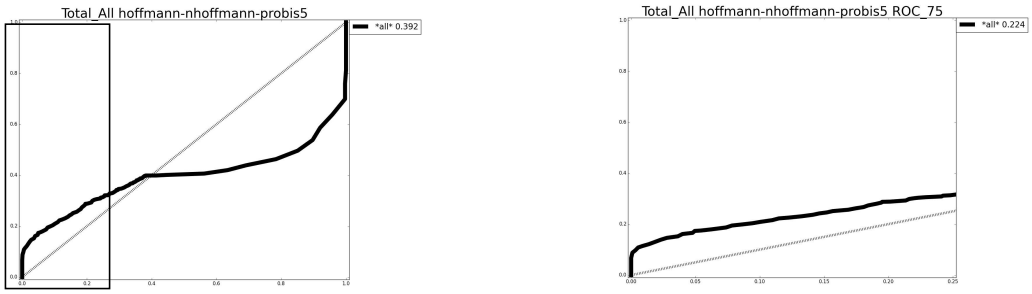
$$
\mathcal{C} = \bigcup_{i=1}^{n} (\{i\} \times \mathcal{C}_i) \qquad \mathcal{P} = \bigcup_{i=1}^{n} (\{i\} \times \mathcal{P}_i)
$$

$$
\mathcal{N} = \bigcup_{i=1}^{n} (\{i\} \times \mathcal{N}_i)
$$

Les *prédits positifs* et *prédits négatifs* sont déterminés par un n-uplet de rangs seuils $s = (s_i)_{i=1}^n$, définissant un seuil par instance. On définit également le n-uplet $s_X = (s_{i,X})_{i=1}^n$ en considérant chaque $s_{i,X}$ indépendamment dans le cadre de l'instance $i$.




Exemple schématique d'un classificateur dont l'AUC globale de 0.5 ne reflète pas l'intérêt qu'il peut avoir par rapport à un classificateur aléatoire, car la différence de comportement entre le début et la fin du classement est moyennée par l'AUC. En revanche l'$AUC_{75}$, restreinte aux cas où la spécificité reste supérieure à 75%, permet de rendre compte de cette caractéristique.




Exemple concret sur une instance réalisée avec l'approche ProBiS. A gauche l'AUC globale est de 0.392, moins bonne que la valeur pour un classificateur aléatoire de 0.5. A droite l'$AUC_{75}$ est de 0.224, meilleure que la valeur pour un classificateur aléatoire de 0.125.

Figure 4 – Exemples schématique et réel d'un classificateur, montrant l'intérêt d'une métrique prenant en compte uniquement la tête du classement.

$$P\mathcal{P}(s) = \bigcup_{i=1}^{n}(\{i\} \times P\mathcal{P}_i(s_i)) \qquad P\mathcal{N}(s) = \bigcup_{i=1}^{n}(\{i\} \times P\mathcal{N}_i(s_i))$$

À partir de ces définitions, on utilise alors les mêmes outils et les mêmes notations que les définitions précédentes. En particulier les courbes ROC, et les différentes AUC et $\text{AUC}_X$ sont bien définies, permettant de comparer des approches sur l'ensemble d'un jeu de données comportant plusieurs instances du problème de prédiction.

# 2 Jeux de données de validation

La construction d'un jeu de données consiste à proposer plusieurs instances du problème de prédiction de cibles, dont les cibles positives et négatives sont connues. Certains jeux de données sont référencés dans la littérature, mais ils sont parfois peu pertinents du point de vue de la prédiction de cibles et en particulier dans la représentativité de molécules d'intérêt thérapeutique (médicament) qui sont la motivation de l'approche qui vise à prédire des cibles secondaires pour des molécules à potentiel thérapeutique. Nous proposons un nouveau jeu de données, à partir de complexes ligand-cibles sélectionnés dans la PDB pour leur intérêt bio-chimique.

## 2.1 Objectifs d'un jeu de données de qualité

### 2.1.1 Accessibilité et qualité des fichiers structuraux

Pour être utile, un jeu de données doit être facilement accessible afin de pouvoir reproduire les *benchmarks* avec différentes approches pour les comparer entre elles. La PDB fournit une source très riche de données structurales, publique et donc accessible par tous, qui est naturellement privilégiée dans la construction de jeu de données. Elle contient notamment, en plus de macromolécules cibles, de nombreux complexes ligand-cibles pour une variété de ligands.

Définir un tel jeu de données revient ainsi à fournir la liste des références aux fichiers de la PDB, permettant de déterminer sans équivoque les molécules et conformations considérées. En particulier il est nécessaire de reporter exactement l'*identifiant* du fichier structural qui peut contenir une ou plusieurs macromolécules en complexe ou non, un identifiant appelé *chaîne* qui détermine la macromolécule considérée au sein du fichier structural, et un *index* qui référence la petite molécule ligand considérée.

La PDB contient des fichiers structuraux de qualités variables. Il est essentiel de vérifier que les structures sélectionnées sont de bonne qualité, par exemple résolues par rayons X et d'une résolution suffisante (typiquement inférieure à 3 ou 5 angström). Il peut être utile de vérifier également que la séquence représentée, présente dans le fichier, correspond bien à la séquence réelle de la protéine au moins dans le voisinage des sites de liaisons car un biais expérimental peut introduire des mutations ou des suppressions de résidus.

### 2.1.2 Intérêt bio-chimique ou pharmacologique des ligands étudiés

Dans l'objectif d'évaluer ou comparer des approches de prédiction de cibles, les ligands pour lesquels la prédiction des cibles est évaluée doivent être représentatifs de problèmes réels qui concernent typiquement la recherche de cibles de molécules bio-chimiquement actives. Certains ligands, comme l'adénosine triphosphate (ATP), sont par exemple sur-représentés dans les jeux de données proposés dans la littérature alors que leur comportement est tout à fait singulier, en se liant à un nombre de protéines différentes très important. Par ailleurs l'étude des cibles de molécules très petites comme le sulfate (SO4) ou le phosphate

Illustration de la définition de l'union des instances.

On considère deux instances du problème de prédiction de cibles.

$L_1$ : première requête
$\mathcal{P}_1 = \{abc\}$
$\mathcal{N}_1 = \{xyz\}$
$\mathcal{C}_1 = \mathcal{P}_1 \cup \mathcal{N}_1 = \{abcxyz\}$

$L_2$ : deuxième requête
$\mathcal{P}_2 = \{adef\}$
$\mathcal{N}_2 = \{xw\}$
$\mathcal{C}_2 = \mathcal{P}_2 \cup \mathcal{N}_2 = \{adextu\}$

À partir des deux instances, on définit l'instance union :

$$
\begin{aligned}
L &= (L_1, L_2) \\
\mathcal{P} &= (\{1\} \times \mathcal{P}_1) \cup (\{2\} \times \mathcal{P}_2) &&= \{(1,a),(1,b),(1,c),(2,a),(2,d),(2,e),(2,f)\} \\
\mathcal{N} &= (\{1\} \times \mathcal{N}_1) \cup (\{2\} \times \mathcal{N}_2) &&= \{(1,x),(1,y),(1,z),(2,x),(2,w)\} \\
\mathcal{C} &= (\{1\} \times \mathcal{C}_1) \cup (\{2\} \times \mathcal{C}_2) &&= \{(1,a),(1,b),(1,c),(2,a),(2,d),(2,e),(2,f),(1,x),(1,y),(1,z),(2,x),(2,w)\} \\
&= \mathcal{P} \cup \mathcal{N}
\end{aligned}
$$

On suppose que les deux instances obtiennent les deux résultats respectifs suivants. On s'intéresse à la partie du classement où la spécificité reste au dessus de 50%, en déterminant les deux rangs seuils correspondant pour chaque instance.

$\text{résultat}(1) = (a,b,x,y,c,z)$
$s_{1,50\%} = 3$
$P\mathcal{P}_1(s_{1,50\%}) = \{a,b,x\}$
$P\mathcal{N}_1(s_{1,50\%}) = \{y,c,z\}$

$\text{résultat}(2) = (a,w,x,d,e,f)$
$s_{2,50\%} = 2$
$P\mathcal{P}_2(s_{2,50\%}) = \{a,w\}$
$P\mathcal{N}_2(s_{2,50\%}) = \{x,d,e,f\}$

Ainsi le n-uplet seuil pour l'union des instances afin de considérer une spécificité supérieure à 50% est $s_{50\%} = (3,2)$.

$$P\mathcal{P}(s_{50\%}) = \{(1,a),(1,b),(1,x),(2,a),(2,w)\} \quad P\mathcal{N}(s_{50\%}) = \{(1,y),(1,c),(1,z),(2,x),(2,d),(2,e),(2,f)\}$$

Figure 5 – Exemple d'union d'instances.

(PO4) présente peu d'intérêt pour notre problématique car leurs propriétés physico-chimiques sont très éloignées des ligands d'intérêt pharmacologique. Les ligands ayant un potentiel thérapeutique peuvent par exemple être reconnus selon les règles de 5 Linpiski [Lipinski 2001], qui propose des fourchettes dans le nombre optimal d'atomes donneurs et accepteurs d'hydrogène, la masse de la molécule, et le coefficient de partage octanol-eau. Ensuite les cofacteurs, qui sont des ligands particuliers qui permettent à un autre ligand de se lier à une macromolécule, jouent un rôle dans la formation du site de liaison pour cet autre ligand mais n'ont pas d'activité biologique *per se*. Enfin certains ligands, comme l'hème (HEM), ont un mode de liaison particulier comportant une liaison covalente ne correspondant pas au mode de liaison principal des ligands d'intérêts.

Il s'agit essentiellement de privilégier la pertinence des ligands dans les mécanismes bio-chimiques, par opposition à la fréquence d'apparition dans la PDB. En effet les ligands très fortement représentés dans la PDB sont souvent ceux qui ne sont pas spécifiques, et peu pertinents dans le contexte de notre problématique appliquée au design rationnel de médicament.

### 2.1.3 Représentativité et redondance des cibles

Dans la liste des cibles positives proposées, la redondance doit être minimisée. En effet si deux cibles positives partagent des séquences trop similaires, on peut supposer que toute méthode permettant d'en retrouver une pourra également retrouver l'autre, ce qui apporte ainsi peu d'information supplémentaire. Par ailleurs des méthodes de prédiction de cibles basées sur la séquence existent, et l'intérêt des méthodes structurales est justement de déterminer celles qui n'auraient pas pu être prédites par les approches basées sur la séquence.

## 2.2 Jeux de données de la littérature

### 2.2.1 Intérêt et limites des jeu de données de la littérature

Contrairement au problème du *docking* pour lequel des jeux de données de qualité sont disponibles, comme l'*Astex diverse set* [Hartshorn 2007], les jeux de données permettant d'évaluer le problème de la prédiction de cibles sont plus rares.

L'objectif de tels jeux de données est de proposer des *benchmarks* permettant de comparer objectivement différentes méthodes pour un même problème de la prédiction. Cela permet notamment de comparer directement une approche à toutes celles dont un *benchmark* sur un jeu de données a déjà été publié. Cependant, il n'existe aucun jeu de données faisant un consensus suffisant pour avoir été utilisé par un nombre significatif de méthodes de prédiction de cibles.

Par ailleurs les jeux de données publiés, comme ceux dont sont issus les jeux de données Kahraman et Hoffmann décrits ci-après partagent certains défauts vis-à-vis de la problématique de prédiction de cibles, qui n'est pas la motivation initiale de ces jeux de données, comme :

- Le choix des ligands : beaucoup de jeux de données utilisent par exemple l'adénosine triphosphate (ATP) dont on sait qu'il se lie à un nombre très important de cibles, c'est-à-dire dont l'activité chimique est sans rapport avec les molécules plus spécifiques d'intérêt pharmacologique. Par ailleurs de nombreux jeux de données considèrent comme différents des ligands pourtant très proches, comme l'ATP et AMP, ou NAD et FAD.
- Le choix des fichiers structuraux : la notion de qualité de la structure tertiaire tridimensionnelle, induite par le choix de la méthode expérimentale (rayon X ou RMN notamment), ou la qualité de la mesure expérimentale n'est pas toujours prise en compte, contrairement aux jeux de données pour le problème de docking qui sont généralement très vérifiés.
- Le choix des cibles négatives : les cibles négatives sont généralement les cibles positives des autres ligands du même jeu de données. Cela induit en particulier une redondance, car par construction plusieurs ensembles de cibles fonctionnellement similaires sont présentes, même si elles ne le sont

pas du point de vue de la séquence. De plus si des ligands sont choisis trop similaires entre eux ils peuvent effectivement se lier aux cibles correspondantes.

- Les imprécisions dans les références : certaines publications ne reportent que le code du fichier de la PDB, le nom du ligand, et parfois l'identifiant de la chaîne de la protéine, une information généralement insuffisante pour retrouver le site de liaison exact dans le cas où les molécules sont représentées en plusieurs exemplaires dans le fichier structural dans des conformations différentes. Parfois l'information reportée est erronée concernant l'identifiant de la chaîne de la protéine ou l'index ligand.

### 2.2.2 Jeu de données Kahraman

Le jeu de données Kahraman, détaillé en annexe B section 1, est construit à partir du jeu données originalement publié dans [Kahraman 2007].

Cet article a pour objectif de comparer la géométrie de plusieurs ligands avec la géométrie de leurs multiples sites de liaison. Plus précisément, ce sont les variations entre les sites de liaison d'un même ligand qui sont étudiées, et l'étude réalisée montre que cette variation dans la géométrie des sites est très importante. En effet les différences de conformation d'un ligand ne suffiraient pas à expliquer cette variété de formes capables de lier ce ligand. Le jeu de données originalement publié est constitué de 9 ligands différents, chacun présenté en complexe avec 5 à 20 cibles positives différentes totalisant 100 cibles. Les ligands présentent une grande variété de forme, taille, et flexibilité. Et pour chaque ligand, les cibles positives appartiennent à des familles différentes, en particulier leurs séquences sont très différentes deux à deux.

Plusieurs modifications ont été réalisées par rapport au jeu de données original. Les ligands phosphate et hème sont retirés, car ces deux petites molécules ne sont pas considérées comme des ligands d'intérêt pour notre problématique de prédiction de cibles. Le phosphate est une très petite molécule ayant de trop nombreux sites de liaisons, et l'hème est une plus grosse molécule ayant un mode de liaison covalent. Les références originales de certaines cibles sont erronées ou obsolètes et sont remplacées. Le jeu de données résultant, Kahraman, est ainsi composé de 7 ligands pour un total de 64 cibles. Pour chaque ligand, les cibles positives des autres ligands sont considérées comme les cibles négatives.

La diversité des sites de liaisons dans ce jeu de données implique qu'il s'agit d'un jeu de données *a priori* difficile pour les approches par similarité des cibles. Il reste cependant intéressant pour comparer les approches entre elles, et a notamment été utilisé comme validation pour *IsoMIF* [Chartier 2015], *IsoCleft* [Najmanovich 2008], *PatchSurfer* [Chikhi 2010, Sael 2010], ou dans [Spitzer 2011].

### 2.2.3 Jeu de données Hoffmann

Le jeu de données Hoffmann, détaillé en annexe B section 2, est construit à partir du jeu données originalement publié dans [Hoffmann 2010].

Cet article propose ce jeu de données notamment pour palier aux inconvénients du précédent, Kahraman, dans l'optique de l'utiliser comme *benchmark* d'une méthode de prédiction de cibles. En effet les auteurs reportent qu'en utilisant le précédant jeu de données, c'est la mesure simple du volume du site qui apparaît comme la meilleure méthode de prédiction, ce qui motive la construction d'un jeu de données. Ce jeu de données appelé *homogène* est construit en choisissant des ligands de tailles similaires où la mesure du volume n'est pas plus efficace que le hasard pour la prédiction de cibles. Il contient 10 ligands pour un total de 100 cibles.

Comme pour le jeu de données Kahraman, certaines modifications sont apportées aux références originales. Le jeu de données résultant contient ainsi les 10 ligands, pour un total de 96 cibles, 4 ayant été supprimées. De manière analogue au jeu de données Kahraman c'est l'ensemble des cibles positives pour les autres ligands qui constituent les cibles négatives d'un ligand requête.

Ce jeu de données avait été utilisé pour valider la méthode de prédiction *IsoMIF* [Chartier 2015]. Par ailleurs, les auteurs de [Nisius 2012] recommandent l'usage de ce jeu de données pour comparer dans le futur les approches de prédiction de cibles plutôt que le jeu de données Kahraman. Même si les molécules choisies comme ligands dans ce jeu de données sont toutes de taille similaire et cohérente par rapport à la taille typique d'un médicament, on note que certains ligands sont peu pertinents du point de vue de leur représentativité des molécules biologiquement actives.

## 2.3 Notre jeu de données Lam-On

Le jeu de données Lam-On a été développé par Bionext SA [1], il est présenté dans l'annexe B section 3. Il s'agit, comme les deux jeux de données précédents, de ligands et cibles extraits de la PDB. Une différence conceptuelle se situe dans la donnée des cibles négatives. Contrairement aux jeux de données Kahraman et Hoffmann les cibles négatives sont données indépendamment des cibles positives des autres ligands. Cet ensemble, également appelé *bruit*, est référencé comme n-195 et détaillé en annexe B section 3.2. Ainsi pour chaque ligand $L_i$ on dispose des cibles positives $\mathcal{P}_i$, et l'ensemble des cibles négative $\mathcal{N}$ reste fixé pour tous les ligands.

Le jeu de données est constitué de 16 ligands, ayant chacun entre 3 et 7 cibles positives, fournies en complexe avec le ligand, pour un total de 85 complexes. Les ligands sont choisis suivant deux critères : des propriétés physico-chimiques similaires à d'autres ligands d'intérêt ayant un potentiel thérapeutique, et des propriétés physico-chimiques variés entre chaque couple de ligands du jeu de données. Les cibles positives sont déterminées parmi les structures disponibles en complexe avec le ligand dans la PDB dont la qualité du modèle est suffisante : résolues par rayons X d'une résolution inférieure à 4 Angström. Par ailleurs, chaque ensemble de cibles positives pour un ligand donné propose une importante variation dans les séquences et fonctions des cibles. Chaque complexe du jeu de données a été vérifié manuellement par un chimiste de Bionext SA, afin de proposer un jeu de données diversifié permettant une meilleure évaluation et comparaison des méthodes de prédiction de cibles.

Afin de proposer des cibles négatives pertinentes, un ensemble de 195 protéines non redondantes du point de vue de la séquence et de la fonction est proposé. Il est construit à partir de listes non-redondantes proposées par la PDB, et la diversité des fonctions à l'aide des classifications Pfam [Finn 2014] et InterPro [Mitchell 2015].

# 3 Comparaison entre BioBind, ProBiS, et Vina

Afin de valider la pertinence de notre algorithme BioBind, nous évaluons la qualité des résultats au problème de classification binaire de prédiction des cibles en utilisant les métriques précédemment présentées, variantes de l'AUC. Les trois jeux de données Kahraman, Hoffmann, et Lam-On sont considérés, et pour chacun d'entre eux les résultats de BioBind sont comparés à ceux d'une autre approche par similarité des cibles, ProBiS [Konc 2010], et une approche par docking, Vina [Trott 2010].

## 3.1 Méthode

### 3.1.1 Choix des approches comparées

Le choix de ProBiS comme point de comparaison est motivé par plusieurs éléments. Tout d'abord il s'agit d'une approche relativement récente, qui propose de nombreuses améliorations par rapport à d'autres approches antérieures. Ensuite la disponibilité du programme et du code source ainsi que la présence d'une documentation claire, permettent de l'utiliser dans le contexte spécifique des *benchmarks* pour la prédiction de cibles sur nos jeux de données.

1. Jeu de données conçu par Lam Nguyen nlam.nguyen@yahoo.fr et Pascal Muller pascal.muller@bionext.com

L'approche eF-seek [Kinoshita 2002] a également été étudiée, afin d'être comparée à BioBind. Il s'agit de l'une des toutes premières approches de recherche de similarités locales basées sur un modèle mathématique précis, avec CavBase [Schmitt 2002]. Cependant le programme n'est disponible qu'à travers un serveur internet [2] qui est orienté pour la problématique biologique de l'annotation fonctionnelle. Ainsi il est seulement possible de fournir une macromolécule complète dont les similarités locales sont recherchées dans une base de données de sites pré-déterminés. Il est possible de construire une intersection des jeux de données avec les bases de recherches disponibles pour comparer les approches, cependant les résultats que nous avons obtenus sont inférieurs à ceux proposés par ProBiS, et ainsi il n'est pas apparu utile d'inclure cette approche dans notre comparatif présenté dans la suite.

### 3.1.2 Paramètres et versions des logiciels

Le programme ProBiS est compilé à partir des sources de la dernière version disponible sur le site du projet en mai 2016 [3]. Les détails des choix réalisés dans la configuration sont présentés dans l'annexe B section 4.2. Nous avons défini une première version, référencée dans la suite simplement comme ProBiS, qui modifie légèrement les paramètres de la version par défaut proposée par les auteurs, en ajoutant uniquement des résultats supplémentaires en fin de classement sans modifier le rang des résultats communs. Une seconde variante, ProBiS-5, a été réalisée qui modifie elle l'ordre des résultats, mais propose de meilleurs résultats selon notre mesure de performance

Le programme AutoDock/Vina proposant l'approche Vina est directement récupéré sous forme binaire à partir du package des distributions GNU/Linux *Debian* et *Ubuntu*, disponibles en mai 2016. Les configurations utilisées sont précisées en annexe B section 4.3, une première variante référencée comme Vina consiste à conserver l'ensemble des paramètres par défaut proposés par les auteurs, une seconde variante référencée comme Vina-4 modifie un paramètre qui augmente sensiblement le temps de calcul afin d'évaluer si la qualité des résultats s'en trouve améliorée.

La version de BioBind utilisée est celle de juin 2016, dont la référence précise est explicité en annexe B section 4.1.

### 3.1.3 Classificateurs binaires

Les trois approches considérées fonctionnent en affectant des scores entre un ligand requête et une cible candidate, soit par similarité avec la cible requête soit directement par complémentarité. Cependant quel que soit le fonctionnement interne on peut extraire un classement des cibles candidates, on obtient ainsi une vision uniforme des méthodes répondant à la spécification suivante :

**Entrée :** Un ligand requête en complexe avec une cible requête, et un ensemble de cibles candidates.

**Sortie :** Un classement des cibles candidates.

On obtient ainsi pour chaque requête du jeu de données, c'est à dire pour la donnée d'un nom de ligand et la référence de la cible requête, un classificateur binaire dépendant du paramètre rang seuil. Une méthode naturelle pour évaluer un classificateur binaire est de calculer l'AUC, cependant dans notre contexte l'aire sous la courbe ROC complète apporte peu d'information. Comme nous l'avions indiqué en section 1.1.3, c'est en effet la tête du classement qui est porteuse de sens, c'est pourquoi l'$\text{AUC}_{90}$ est privilégiée dans la comparaison des approches.

Pour chaque ligand d'un jeu de données chaque cible positive est proposée en complexe avec le ligand requête, on peut ainsi définir un problème de classificateur pour chaque cible positive de chaque ligand. On utilise les méthodes d'union décrites en section 1.2 pour présenter l'union des résultats pour l'ensemble des requêtes d'un ligand, et plus généralement pour l'ensemble de toutes les requêtes d'un jeu de données.

2. ef-site.hgc.jp/eF-seek/top.do
3. probis.cmm.ki.si/?what=parallel

Légende des figures présentées dans la suite.

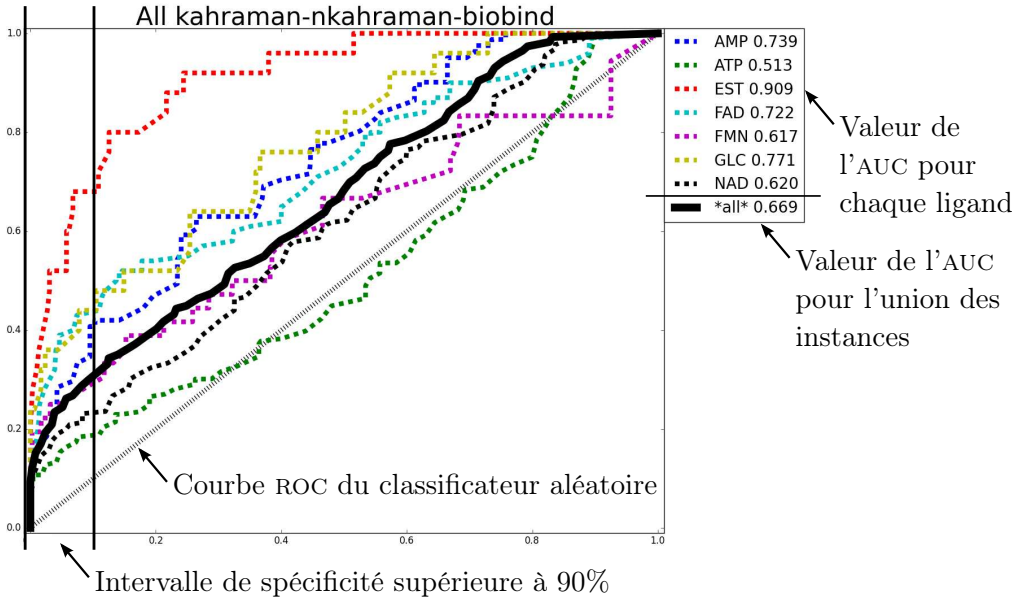



Figure 6 – Description des courbes ROC présentées. Les courbes ROC sont reproduites uniquement dans l'intervalle de spécificité supérieure à 90%. Elles sont tracées pour l'union de toutes les requêtes du jeu de données (ligne noire continue) ainsi que pour les requêtes de chaque ligand (lignes colorées pointillées). De même, les valeurs des AUC correspondantes sont présentées en légende : chaque ligne correspond au problème de prédiction pour un ligand, et la dernière ligne correspond à l'union de toutes les requêtes du jeu de données.

## 3.2 Comparaison des trois approches sur les trois jeu de données

Pour chaque jeu données on précise pour chacune des approches les $\text{AUC}_{90}$ obtenues d'une part pour chacun des ligands requêtes et d'autre part pour l'union de toutes les instances. Les courbes ROC correspondantes, restreintes à cet intervalle de spécificité supérieure à 90% sont également reproduites, selon le modèle décrit dans la figure 6. On rappelle que dans le contexte où les aires sont considérées sur l'intervalle de spécificité supérieure à 90%, la valeur de l'espérance d'un classificateur aléatoire est de 0.05.

### 3.2.1 Résultats pour Kahraman

Le tableau 1 qui suit présente les aires sous les courbes restreintes à la spécificité supérieure à 90% pour Vina, ProBiS, ProBiS-5 et BioBind sur le jeu de données Kahraman. Les courbes ROC sur cet intervalle sont également tracées en figure 7 pour les approches ProBiS et ProBiS-5, et en figure 8 pour les deux autres approches Vina et BioBind.

Les $\text{AUC}_{90}$ des différentes approches sont significativement supérieures à l'espérance d'un classificateur aléatoire. Ces aires restent cependant relativement faibles, car les courbes ROC sont proches de la diagonales, ce qui montre la difficulté du jeu de données. On note en particulier que l'approche ProBiS-5 que nous avons définie en modifiant les paramètres proposés par défaut par les auteurs du logiciel semble proposer de meilleurs résultats sur ce jeu données et pour notre métrique.

On constate notamment que BioBind obtient une aire supérieure pour tous les ligands du jeu de données, sauf GLC (glucose). Ce dernier ligand est par ailleurs l'un des deux seuls pour lesquels la variante ProBiS-5 augmentant la taille du site de liaison requête n'est pas meilleure que la version par défaut, c'est également le ligand pour lequel l'approche par docking, Vina, est la plus performante. Il

| $AUC_{90}$ | Vina | ProBiS | ProBiS-5 | BioBind |
|---|---|---|---|---|
| Total | 0.104 | 0.129 | 0.143 | **0.235** |
| AMP | 0.122 | 0.162 | 0.194 | **0.263** |
| ATP | 0.044 | 0.113 | 0.104 | **0.149** |
| EST | 0.207 | 0.085 | 0.186 | **0.532** |
| FAD | 0.179 | 0.132 | 0.140 | **0.344** |
| FMN | 0.088 | 0.149 | 0.167 | **0.251** |
| GLC | 0.294 | **0.443** | 0.412 | 0.363 |
| NAD | 0.064 | 0.091 | 0.115 | **0.185** |

Table 1 – Comparaison des $AUC_{90}$ sur le jeu de données Kahraman. Chaque ligne correspond au problème de prédiction union pour l'ensemble des complexes requêtes définis sur un ligand. La première ligne correspond à l'union de toutes les requêtes du jeu de données.

s'agit du ligand le plus petit du jeu de données avec seulement 12 atomes lourds (i.e. hors hydrogène), les autres ayant plus de 20 atomes lourds. Une corrélation entre les tailles des ligands et les différences de comportement entre les approches et notamment les variantes de ProBiS semble ainsi cohérente, mais n'a cependant pas encore été étudiée.

### 3.2.2 Résultats pour Hoffmann

Comme pour le jeu de données précédant, le tableau 2 présente les aires sous la courbes sur l'intervalle de spécificité supérieure à 90%, les courbes correspondantes étant tracées en figures 9 et 10. On rappelle encore que dans le contexte où les aires sont considérées sur l'intervalle de spécificité supérieure à 90%, la valeur de l'espérance d'un classificateur aléatoire est de 0.05.

| $AUC_{90}$ | Vina | ProBiS | ProBiS-5 | BioBind |
|---|---|---|---|---|
| Total | 0.134 | 0.139 | 0.163 | **0.310** |
| 1PE | 0.073 | 0.079 | 0.095 | **0.195** |
| BOG | 0.114 | 0.110 | 0.147 | **0.264** |
| GSH | 0.024 | 0.125 | 0.137 | **0.376** |
| LDA | 0.173 | 0.079 | 0.120 | **0.265** |
| LLP | 0.027 | 0.226 | 0.258 | **0.422** |
| PLM | 0.316 | 0.071 | 0.105 | **0.365** |
| PMP | 0.065 | 0.228 | 0.274 | **0.402** |
| SAM | 0.137 | 0.141 | 0.153 | **0.309** |
| SUC | 0.067 | 0.164 | 0.197 | **0.211** |
| U5P | 0.151 | 0.133 | 0.134 | **0.228** |

Table 2 – Comparaison des $AUC_{90}$ sur le jeu de données Hoffmann. Chaque ligne correspond au problème de prédiction union pour l'ensemble des complexes requêtes définis sur un ligand. La première ligne correspond à l'union de toutes les requêtes du jeu de données.

Les résultats sont assez similaires à ceux obtenus sur le jeu de données précédant, c'est-à-dire que les aires sont significativement supérieures à l'espérance pour un classificateur aléatoire, en restant tout de même relativement faibles. On note que BioBind propose des résultats meilleurs pour l'intégralité des ligands considérés dans le jeu de données.

Comme pour Kahraman la variante ProBiS-5 introduite par rapport à la version par défaut ProBiS permet d'améliorer sensiblement les résultats, qui restent toutefois très inférieurs à ceux obtenus par notre

approche BioBind.

### 3.2.3 Résultats pour Lam-On

On présente également les aires sous la courbe ROC sur l'intervalle de spécificité supérieure à 90% dans le tableau 3, les courbes étant tracées en figures 11 et 12. Le ligand 537 n'a pas été traité correctement par Vina, à l'origine de la valeur nulle correspondante, sans que cela ne remette en cause la tendance générale des résultats. Comme pour les jeux de données précédemment évalués, BioBind propose de meilleurs résultats selon notre métrique, que ProBiS et Vina.

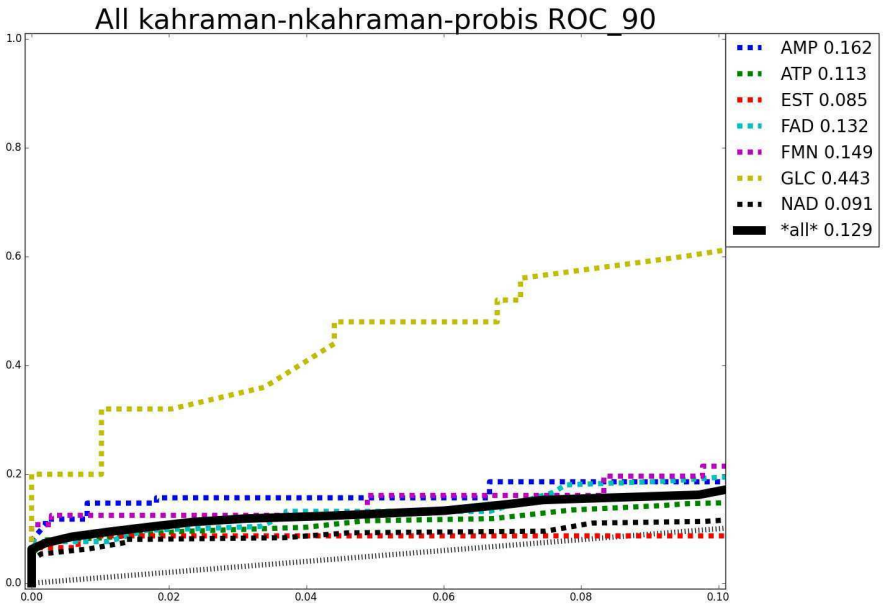
Les résultats de l'ensemble des approches, meilleurs et plus dispersés selon les différents ligands que pour les jeux de données précédents, confirment l'intérêt du jeu de données pour la validation et la comparaison des approches de prédiction de cibles. Certaines requêtes comme RDF illustrent par exemple la supériorité des deux approches par similarité des cibles par rapport au *docking* dans le contexte de la prédiction de cible, alors que d'autres comme 38Z montrent que le *docking* peut tout de même être pertinent dans certains cas.

| $AUC_{90}$ | Vina | ProBiS | ProBiS-5 | BioBind |
|---|---|---|---|---|
| Total | 0.349 | 0.604 | 0.686 | **0.771** |
| 2FA | 0.490 | 0.324 | 0.597 | **0.644** |
| 38Z | 0.712 | 0.756 | 0.733 | **0.798** |
| 3AM | 0.220 | 0.257 | **0.415** | 0.352 |
| 444 | 0.738 | 0.472 | 0.579 | **0.968** |
| 537 | 0.000 | 0.062 | 0.596 | **0.873** |
| 8PR | 0.463 | 0.450 | 0.428 | **0.514** |
| 8XQ | 0.286 | 0.958 | **0.965** | 0.691 |
| AC2 | 0.280 | 0.346 | **0.476** | 0.465 |
| AIX | 0.089 | 0.558 | 0.499 | **0.718** |
| AZR | 0.200 | 0.277 | 0.183 | **0.617** |
| BAT | 0.237 | **0.999** | 0.993 | 0.996 |
| CTS | 0.399 | 0.676 | 0.695 | **0.722** |
| GNH | 0.127 | 0.349 | 0.417 | **0.622** |
| KSA | 0.241 | 0.680 | 0.955 | **0.966** |
| NA7 | 0.142 | 0.295 | 0.397 | **0.633** |
| RDF | 0.089 | 0.999 | **1.000** | **1.000** |

Table 3 – Comparaison des $AUC_{90}$ sur le jeu de données Lam-On. Chaque ligne correspond au problème de prédiction union pour l'ensemble des complexes requêtes définis sur un ligand. La première ligne correspond à l'union de toutes les requêtes du jeu de données.

Jeu de données Kahraman : 1 / 2.

ProBiS $ROC_{90}$




ProBiS-5 $ROC_{90}$

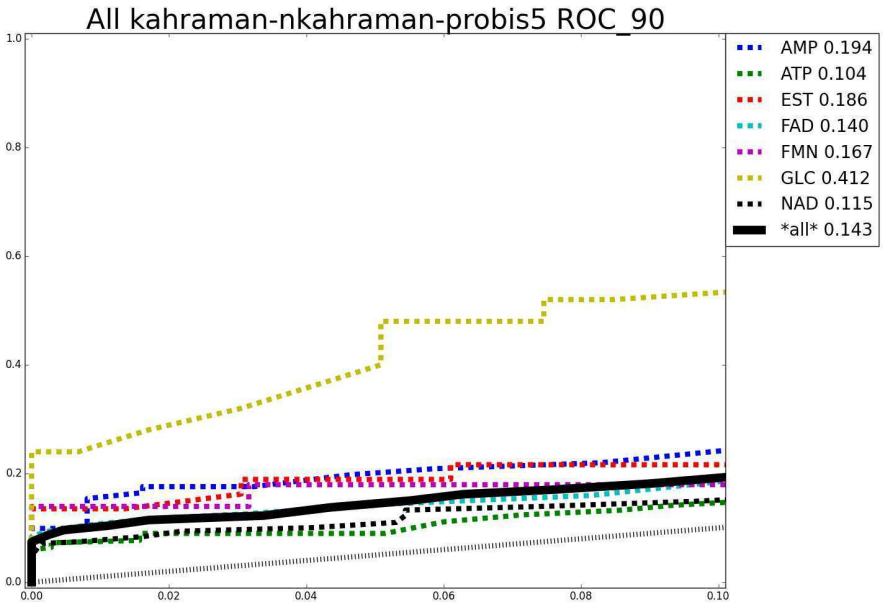
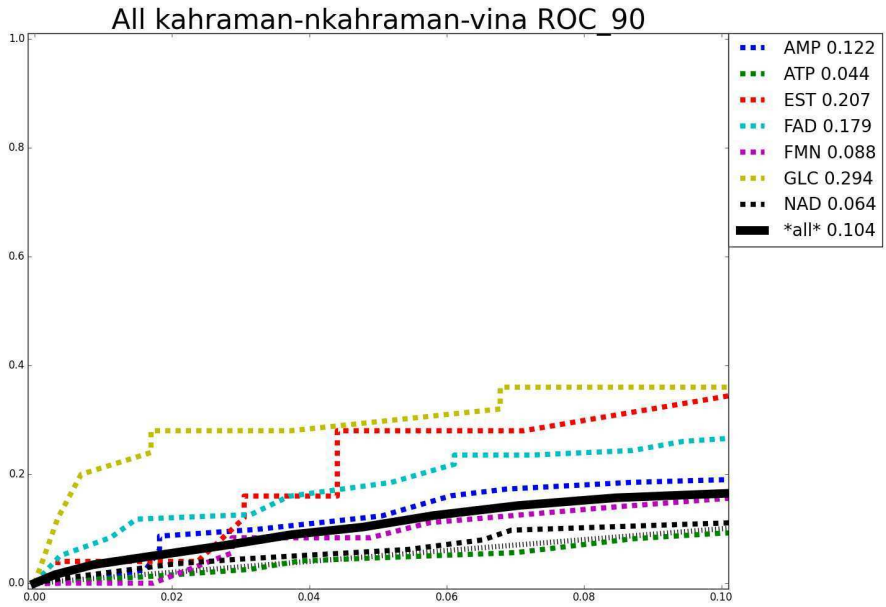



Figure 7 – Comparaison entre ProBiS et ProBiS-5 sur le jeu de données Kahraman. Les courbes ROC sont représentées sur l'intervalle de spécificité supérieure à 90%. Les courbes sont tracées pour l'union de toutes les requêtes du jeu de données (ligne noire continue) ainsi que pour les requêtes de chaque ligand (lignes colorées pointillées). Les AUC correspondantes sont rappelées en légende.

Jeu de données Kahraman : 2 / 2.

Vina $ROC_{90}$




BioBind $ROC_{90}$

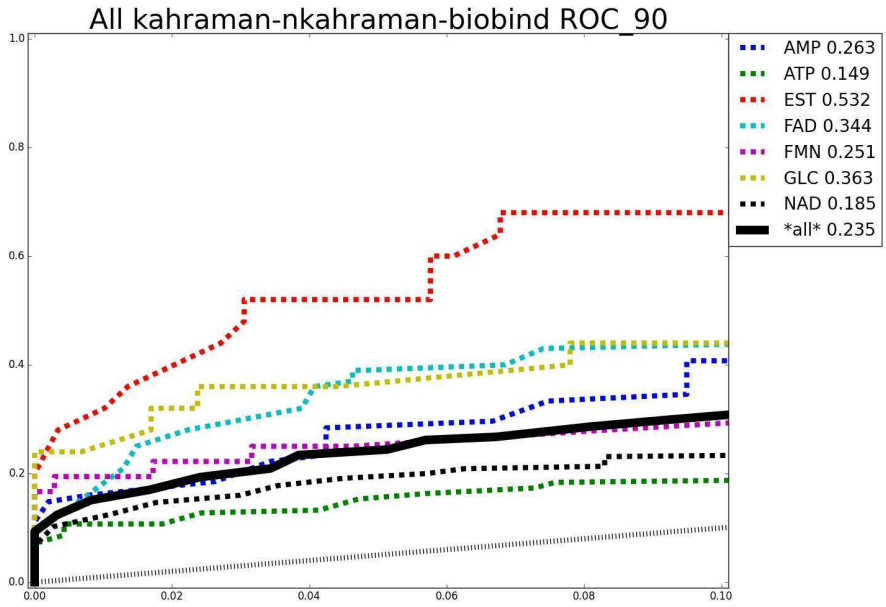
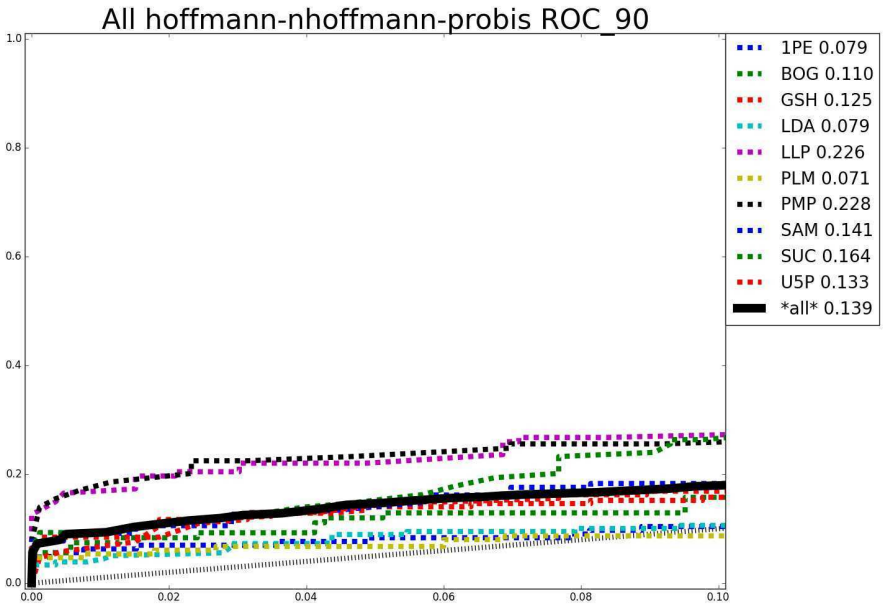



Figure 8 – Comparaison entre Vina et BioBind sur le jeu de données Kahraman. Les courbes ROC sont représentées sur l'intervalle de spécificité supérieure à 90%. Les courbes sont tracées pour l'union de toutes les requêtes du jeu de données (ligne noire continue) ainsi que pour les requêtes de chaque ligand (lignes colorées pointillées). Les AUC correspondantes sont rappelées en légende.

Jeu de données Hoffmann : 1 / 2.

ProBiS $ROC_{90}$




ProBiS-5 $ROC_{90}$

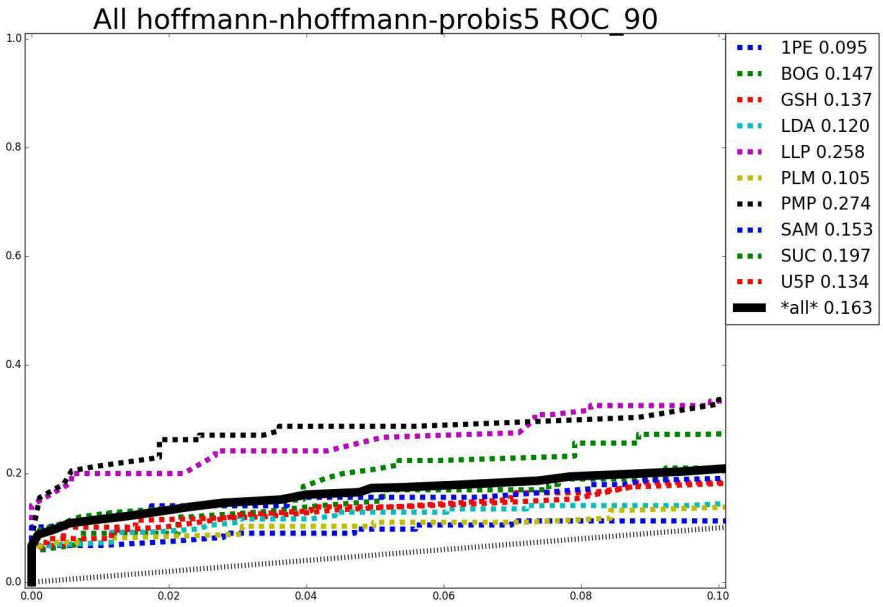
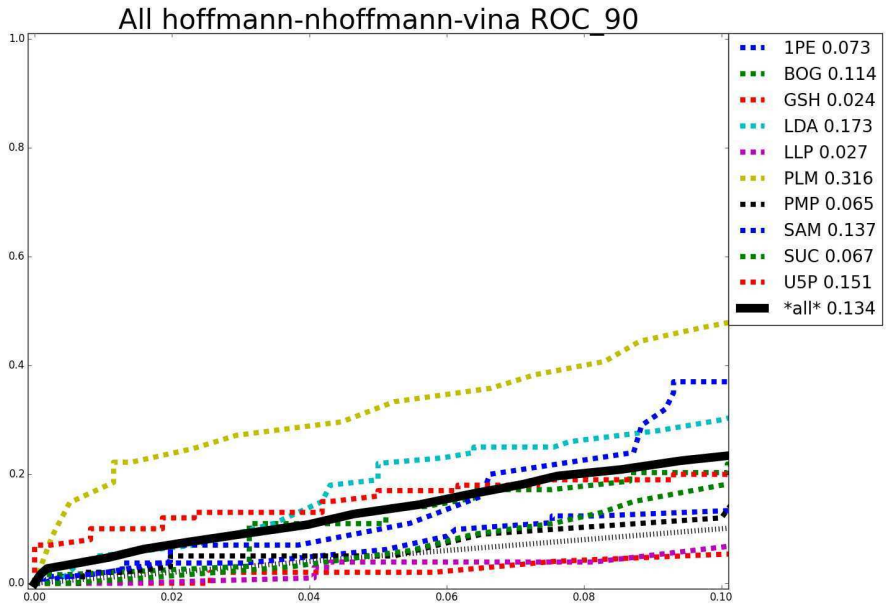



Figure 9 – Comparaison entre ProBiS et ProBiS-5 sur le jeu de données Hoffmann. Les courbes ROC sont représentées sur l'intervalle de spécificité supérieure à 90%. Les courbes sont tracées pour l'union de toutes les requêtes du jeu de données (ligne noire continue) ainsi que pour les requêtes de chaque ligand (lignes colorées pointillées). Les AUC correspondantes sont rappelées en légende.

Jeu de données Hoffmann : 2 / 2.

Vina $ROC_{90}$




BioBind $ROC_{90}$

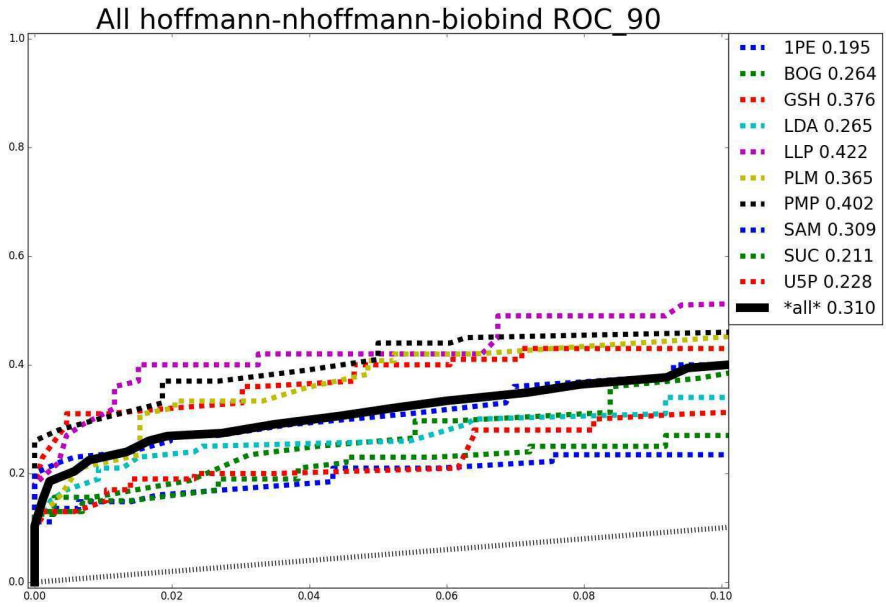
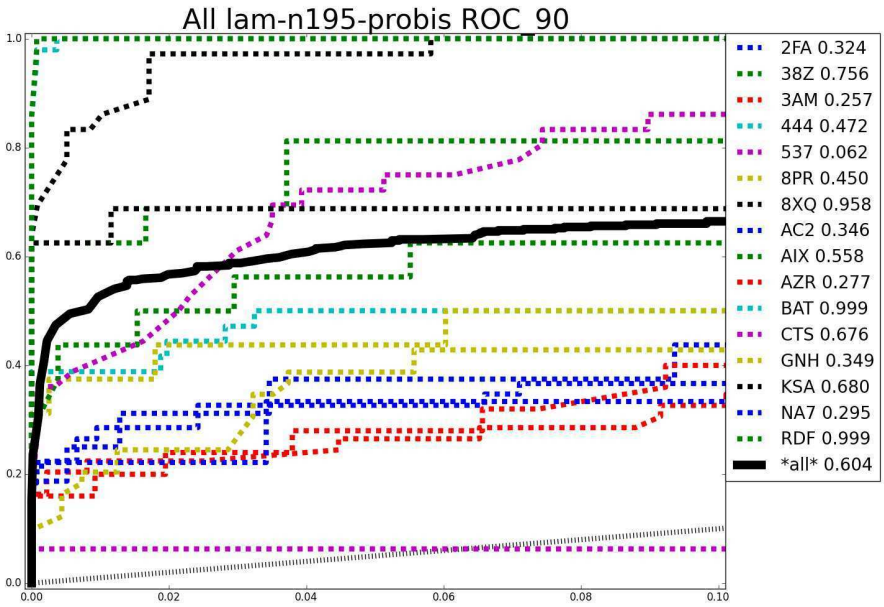



Figure 10 – Comparaison entre Vina et BioBind sur le jeu de données Hoffmann. Les courbes ROC sont représentées sur l'intervalle de spécificité supérieure à 90%. Les courbes sont tracées pour l'union de toutes les requêtes du jeu de données (ligne noire continue) ainsi que pour les requêtes de chaque ligand (lignes colorées pointillées). Les AUC correspondantes sont rappelées en légende.

Jeu de données Lam-On : 1 / 2.

ProBiS $ROC_{90}$




ProBiS-5 $ROC_{90}$

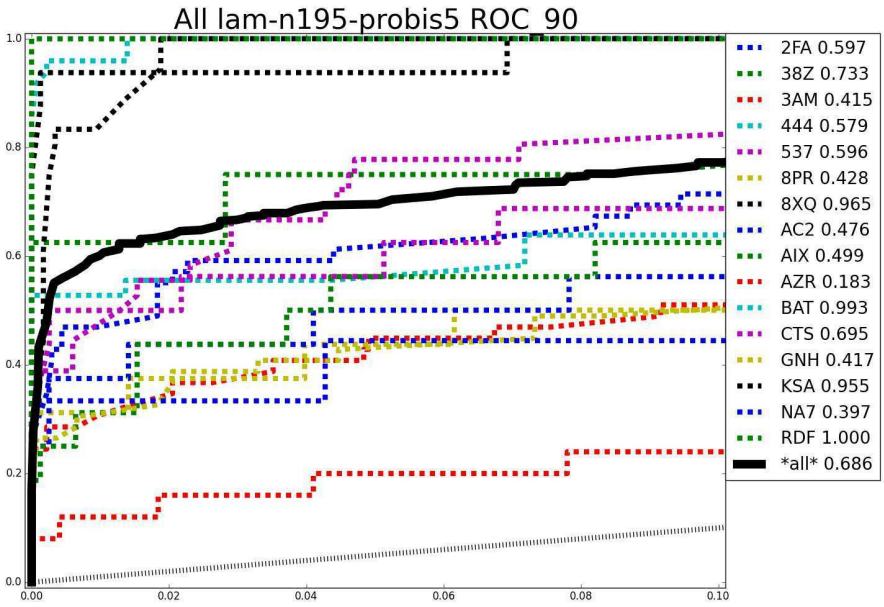
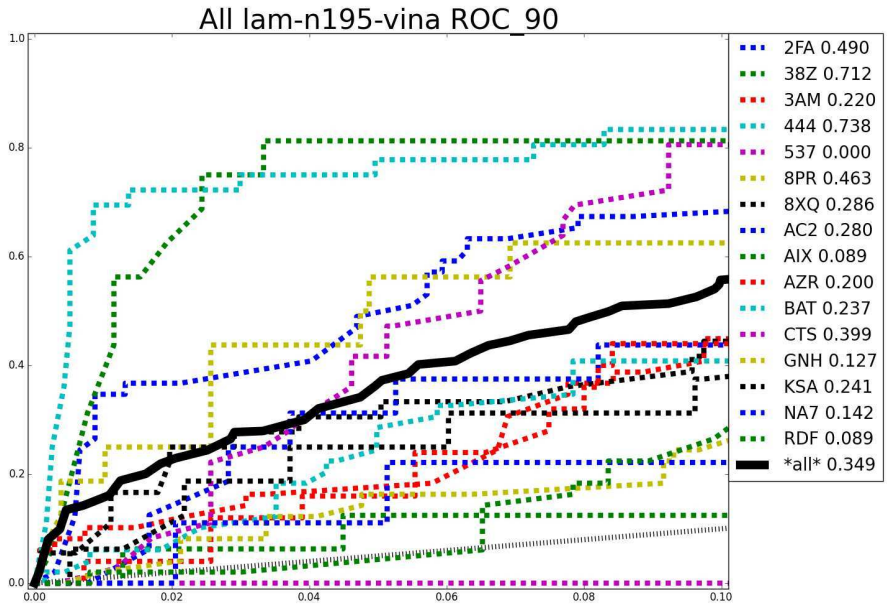



Figure 11 – Comparaison entre ProBiS et ProBiS-5 sur le jeu de données Lam-On. Les courbes ROC sont représentées sur l'intervalle de spécificité supérieure à 90%. Les courbes sont tracées pour l'union de toutes les requêtes du jeu de données (ligne noire continue) ainsi que pour les requêtes de chaque ligand (lignes colorées pointillées). Les AUC correspondantes sont rappelées en légende.

Jeu de données Lam-On : 2 / 2.

Vina $ROC_{90}$




BioBind $ROC_{90}$

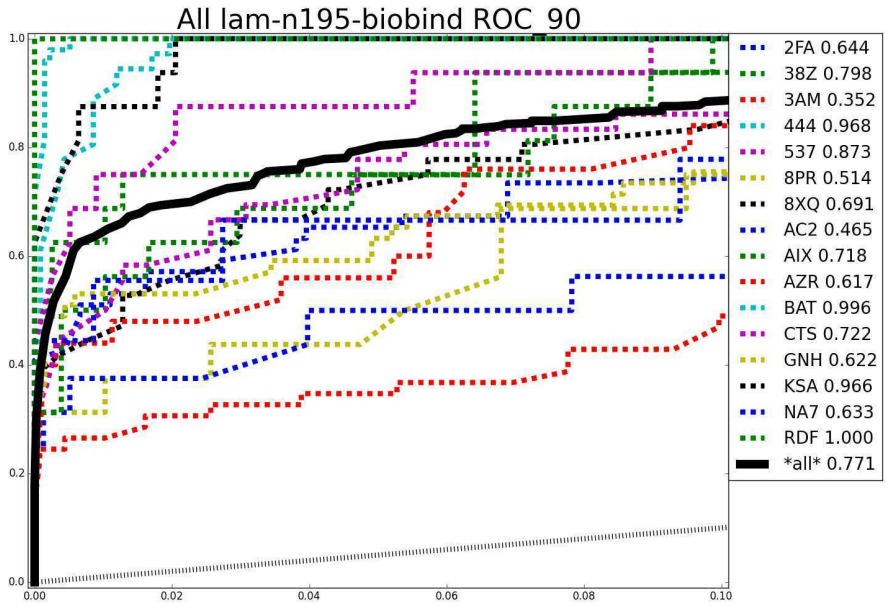



Figure 12 – Comparaison entre Vina et BioBind sur le jeu de données Lam-On. Les courbes ROC sont représentées sur l'intervalle de spécificité supérieure à 90%. Les courbes sont tracées pour l'union de toutes les requêtes du jeu de données (ligne noire continue) ainsi que pour les requêtes de chaque ligand (lignes colorées pointillées). Les AUC correspondantes sont rappelées en légende.

## 3.3 Analyse des résultats

### 3.3.1 Validation de BioBind comme approche par similarité des cibles

Le premier objectif des *benchmarks* réalisés sur les jeux de données est de valider l'intérêt de BioBind pour la prédiction de nouvelles cibles d'un ligand à partir d'un premier complexe connu impliquant celui-ci. On constate que sur les deux jeux de données issus de la littérature, ainsi que sur notre jeu de données original, BioBind obtient des résultats nettement meilleurs que ProBiS et Vina pour notre mesure de performance.

On constate sur les deux jeux de données Kahraman et Hoffmann une même tendance où notre mesure de l'$AUC_{90}$ est assez similaire entre les deux approches Vina et ProBiS, et assez faible, avec une mesure très sensiblement meilleure pour BioBind. Les résultats obtenus sur le jeux de données Lam-On sont eux très différents. En effet sur ce jeu de données les deux approches par similarité BioBind et ProBiS proposent des mesures de l'$AUC_{90}$ proches, et très sensiblement supérieures à la performance de l'approche Vina par *docking*.

Le temps de calcul moyen, mesuré avec le jeu de données Lam-On et reporté dans la table 4, est très variable selon les approches. L'ordre de grandeur est d'un rapport $\times 10$ entre l'approche la plus rapide, ProBiS, et notre approche BioBind. Le temps de calcul pour le *docking* avec Vina est encore très supérieur, de l'ordre d'un facteur $\times 10$ à plus de $\times 100$ selon la configuration utilisée.

On conclut ainsi que BioBind propose des résultats globalement meilleurs que ProBiS au prix d'un temps de calcul plus élevé, ainsi que par rapport à Vina qui est à la fois plus lent et moins performant vis-à-vis de notre problématique considérée et notre mesure de performance.

| Méthode | Durée moyenne (heure) une requête un processeur | Durée mesurée |
|---|---|---|
| ProBiS(-5) | $\sim 0.2$ | 1 heure pour 85 requêtes sur 1 à 32 processeurs |
| BioBind | 2 | 6 heures pour 85 requêtes sur 32 processeurs |
| Vina | 14 | 6 jours pour 85 requêtes sur 8 processeurs |
| Vina-4 | 576 | 12 jours pour 16 requêtes sur 32 processeurs |

Table 4 – Comparaison des temps de calcul entre Vina-4, Vina, BioBind, et ProBiS. Chaque requête est comparée au bruit n-195 (195 cibles) en plus des cibles positives (3 à 7 cibles). Les traitements des fichiers et résultats biaisent légèrement le temps de calcul mesuré, mais reste négligeable devant de temps de « calcul utile » des méthodes. La durée moyenne présentée dans la seconde colonne est ramenée sur un processeur pour une requête afin de pouvoir comparer les valeurs. Pour la première ligne, la valeur approximative 0.2 est proposée car la parallélisation de ProBiS n'a pas été optimisée dans cette expérience, et le temps mesuré ne devrait pas être multiplié par 32 processeurs.

### 3.3.2 *Docking* pour la prédiction de cibles

Les scores attribués par les logiciels de *docking* afin de mesurer l'affinité d'un ligand avec une cible sont connus pour ne pas être pertinents si on les compare entre différentes macromolécules [Wang 2012]. Cependant il n'existe à notre connaissance aucune étude évaluant une approche de *docking aveugle* directement pour le problème de la prédiction de cibles, ce qui motive son inclusion dans notre *benchmark*.

Nous avons utilisé AutoDock/Vina sans aucune modification des paramètres par défaut, sans utiliser d'heuristiques pour normaliser les scores entre différentes cibles (voir par exemple la normalisation proposée dans [Scrima 2014]), et en réalisant un *docking* entièrement aveugle en considérant l'ensemble des macromolécules sans préciser les sites potentiels. Les résultats sont ambivalents, car si les performances

sont nettement inférieures aux approches par similarité des cibles, elles restent très supérieures à un tirage aléatoire.

La connaissance préalable d'une première cible des ligands n'est pas utilisée par les approches par *docking*. Le problème adressé est donc ainsi plus général, ce qui peut justifier son intérêt même si le temps de calcul est supérieur. Afin de savoir si, quitte à utiliser un puissance de calcul très supérieure, le *docking* pourrait répondre à la question de la prédiction de cible, nous avons testé la variante Vina-4 comportant un paramètre augmentant l'exhaustivité de la recherche. Les résultats, présentés dans la figure 13, montrent effectivement une légère amélioration des résultats, au prix d'un temps de calcul très fortement supérieur.

### 3.3.3 Autres métriques pour évaluer la recherche de similarité

Nous proposons une méthode d'évaluation des approches de prédiction de cibles par similarité en considérant le problème de classification binaire associé. Ce point de vue nous permet d'adapter la métrique classique de l'AUC pour évaluer et comparer les approches sur différents jeux de données. Cependant il est possible de considérer d'autres critères d'évaluation d'une approche de recherche de similarité, en prenant en compte la précision de l'alignement proposé du ligand et des sites prédits. Il s'agit par exemple de considérer le RMSD mesurant l'écart entre la position du ligand qui est proposée à partir de la similarité et la position réelle du ligand dans le site prédit.

La définition des cibles vraies positives dans notre *benchmark*, c'est-à-dire celle qui sont prédites positives et qui sont effectivement capable de lier le ligand requête, ne tient pas compte de la position effective de l'interaction prédite. Il arrive par exemple qu'une cible soit à juste titre retenue, mais que la similarité détectée soit dans une région de la macromolécule qui ne correspond pas au site de ce ligand.

Une première manière de tenir compte de cette précision de l'alignement consiste à considérer que les vrais positifs dont le site prédit ne correspond pas au site réel, sont en réalité des faux négatifs. Cette évaluation a été réalisée en utilisant une mesure de la distance entre les sites calculée comme le RMSD entre la position réelle du ligand, et la position prédite associée à la superposition des sites. La valeur seuil a été fixée à 5 Å, les résultats associés sont alors légèrement inférieurs pour les trois approches étudiées, sans modifier les performances relatives.

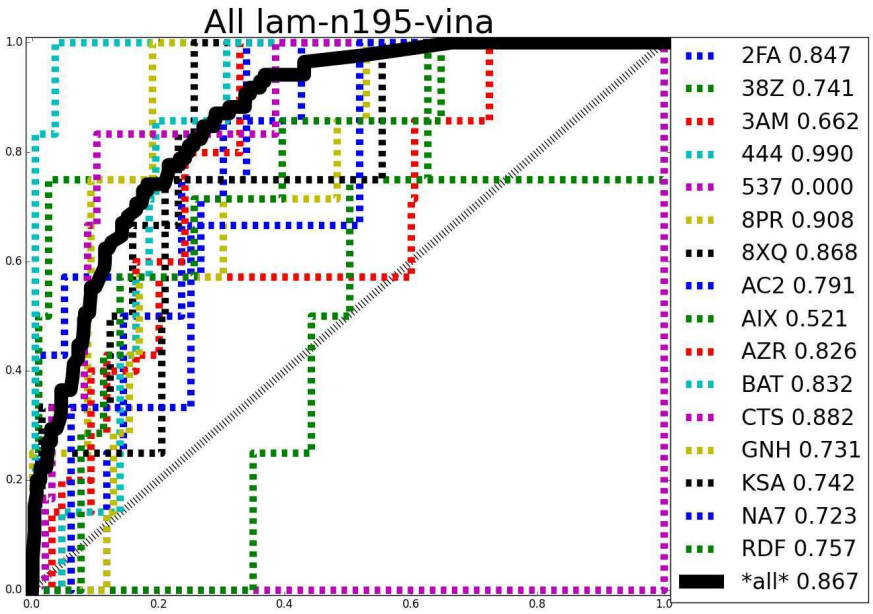
Une seconde manière d'évaluer cette précision, plus proche des méthodes habituelles d'évaluation des logiciels de *docking*, consiste à considérer uniquement les cibles positives en mesurant encore l'écart entre les positions prédite et réelle du ligand. Une analyse similaire est par exemple reportée dans la table 4 de [Konc 2010] qui propose une comparaison de ProBiS avec d'autres approches en considérant les distances entre les résidus des sites réels et prédits, en revanche ce dernier jeu de donnée concernent des ligands ayant peu d'intérêt pharmacologique comme le Zinc, et il n'existe à notre connaissance aucun jeu de données dans la littérature conçu dans ce sens, qui aurait permis une comparaison objective avec les autres approches.

# 4 Conclusion

L'évaluation d'une approche par similarité pour la prédiction de cibles nécessite de définir précisément le problème adressé, afin de proposer un jeu de données assorti d'une métrique pertinente du point de vue de l'usage auquel est destiné une telle approche. Bien que la finalité de notre approche de recherche de similarité soit classique et partagée par de nombreuses autres approches, il n'existe pas de jeu de données faisant consensus qui soit : accessible, précis dans les références, et pertinent dans les ligands et cibles considérés. Il s'agit là d'un point de divergence avec les approches par *docking* qui bénéficient de jeux de données et métriques consensuelles permettant de comparer les différentes approches, par exemple dans le cadre de la compétition CAPRI (*the Critical Assessment of Predicted Interactions* [Janin 2003]).

Nous proposons une définition précise du problème de la prédiction de cibles présenté comme un problème de classification binaire. Cette définition permet essentiellement d'utiliser les outils classiques

Vina




Vina-4

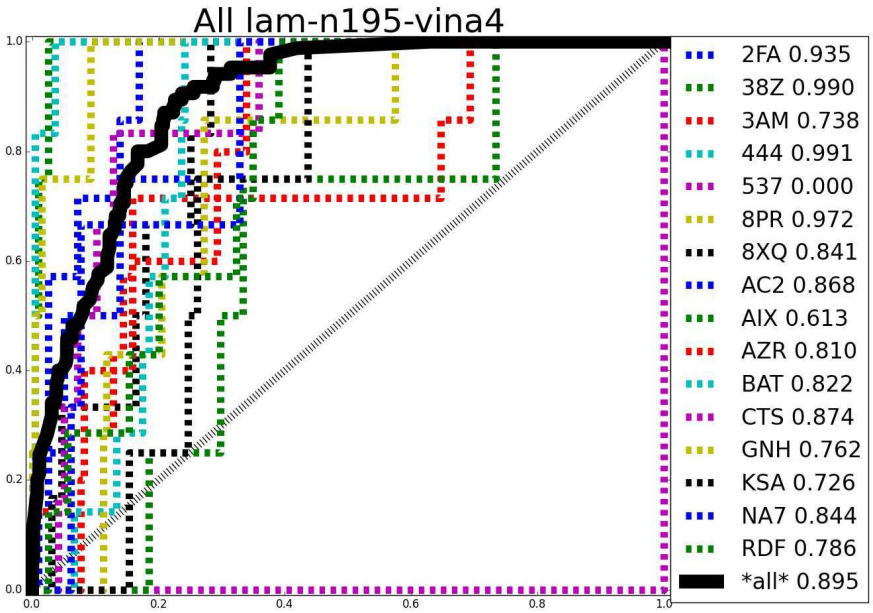



Figure 13 – Comparaison entre Vina et la variante plus exhaustive Vina-4 sur le jeu de données Lam-On. Le temps de calcul pour Vina-4 est très important, permettant une légère amélioration des résultats qui restent cependant très inférieurs aux approches par similarité des cibles.

d'évaluation comme l'aire sous la courbe ROC qui permet de comparer différentes approches entre elles et d'évaluer une approche par rapport à une méthode aléatoire et une méthode parfaite. Cette métrique est adaptée pour proposer les mêmes avantages, mais en ne considérant que les résultats de la tête du classement d'une approche de prédiction, qui correspond aux résultats pouvant être de manière réaliste analysés par un utilisateur réel. On rappelle que la PDB propose plusieurs centaines de milliers de structures tridimensionnelles de macromolécules, alors qu'une expertise par un chimiste représente un travail de l'ordre de plusieurs semaines pour seulement une centaine de cibles candidates proposées, et ne pourra être réalisée que sur très petite portion de l'espace des cibles.

Malgré certains inconvénients dans leur construction, vis-à-vis de notre problématique, nous avons utilisé deux jeux de données issus de la littérature. Le choix des ligands dont les cibles sont étudiées, et parfois le manque de précision dans les références des fichiers structuraux, ont été corrigés pour proposer les jeux de données Kahraman et Hoffmann. Ces deux jeux de données s'avèrent effectivement difficiles dans le contexte de la prédiction de cibles, comme cela était prévisible car par construction ils proposent des sites de liaison particulièrement variés pour chaque ligand considéré. Ces considérations ont motivé la construction d'un jeu de données nouveau spécialement conçu pour l'évaluation d'approches de prédiction de cibles, qu'elles soient par similarité ou non. Ce jeu de données, Lam-On, a comme objectif de proposer des instances plus pertinentes de problème de prédiction, notamment par le choix des ligands ayant des propriétés chimiques proches de composés d'intérêt pharmacologique ainsi que par la sélection des cibles positives et de l'espace de recherche.

Afin de comparer notre algorithme, nous avons sélectionné une autre approche par similarité des cibles, ProBiS [Konc 2010], ainsi qu'une approche par *docking*, Vina [Trott 2010]. ProBiS est une approche récente, performante, et populaire de détection de similarités entre protéines. L'accessibilité du logiciel, du code source, et de la documentation ont également motivé ce choix. Nous avons inclus une approche par *docking* dans notre comparaison, malgré de nombreuses critiques sur la capacité du *docking* et en particulier du *docking* à proposer des scores pertinents entre plusieurs cibles distinctes. En effet même si le comportement du score a déjà été étudié, le pouvoir de prédiction dans le cadre de notre problématique n'a jamais été étudié en tant que tel. AutoDock/Vina est le logiciel de *docking* le plus populaire au vu du nombre de citations dans les publications scientifiques, il s'agit également d'un logiciel récent, librement accessible.

Les résultats obtenus avec les jeux de données Kahraman et Hoffmann sont relativement faibles pour toutes les approches. On note d'importantes différences entre les différents ligands considérés, on vérifie par ailleurs qu'il y a également une dépendance au choix de la cible requête utilisée pour un même ligand. En considérant l'$\text{AUC}_{90}$ on observe une distinction claire entre ProBiS et Vina qui proposent des aires très inférieures à BioBind.

Le jeu de données Lam-On propose quant à lui des résultats très supérieurs selon notre métrique d'évaluation, en particulier pour les deux approches par similarité ProBiS et BioBind. On observe de même des différences importantes selon les ligands et le choix de la cible requête pour un ligand. Cela confirme en particulier l'intérêt de jeu de données dans l'évaluation des approches de prédictions par similarité des cibles.

Le *docking* est une approche qui n'est pas conçue ni optimisée pour répondre à la prédiction de cibles, comme l'a confirmé notre étude. De plus le temps de calcul important rend cette approche impraticable à grande échelle. Cependant il nous semble important de mettre en valeur le fait que les performances restent d'une part très supérieures à un classificateur aléatoire, et d'autre part très proches de celle de ProBiS sur deux des trois jeux données. L'utilisation de cet outil peut ainsi rester pertinente si on ne connaît pas de première cible pour le ligand, ou si l'espace de recherche est réduit.

Notre approche BioBind se révèle ainsi globalement supérieure aux deux autres approches, sur chacun des jeux de données. On note cependant, en particulier sur notre jeu de données Lam-On que pour un certain nombre d'instances, ProBiS propose des résultats plus intéressants. Ces exemples devront servir par la suite pour améliorer encore notre approche, mais nous n'avons pour l'heure pas d'hypothèse sur

les raisons du comportement différents, par exemple par une hypothétique influence de la taille du site de liaison, ou la nature de certains motifs en surface des cibles. On conclut cependant que BioBind, au prix d'un temps de calcul supérieur mais restant réaliste pour des études sur l'ensemble de PDB, propose des résultats intéressants.

Enfin, des métriques complémentaires concernant en particulier la précision des superpositions proposées ainsi que la qualité de notre fonction de score de similarité doivent être développés afin de détailler les voies d'amélioration nécessaire. Notre jeu de données a déjà été enrichi dans ce sens en proposant des mesures de similarité « par inspection visuelle » selon la superposition induite par les ligands.

# Conclusion

L'algorithme BioBind développé dans le cadre de cette thèse contribue à la recherche d'outils informatiques permettant de mettre à profit la quantité de données mises à disposition par la biologie structurale. Il s'agit de proposer des méthodes d'aide à la conception de médicament, de prédiction des effets secondaires, et de compréhension des mécanismes biologiques à l'échelle des interactions moléculaires entre ligands et protéines, ARN, ou ADN. Ces méthodes entrent dans le cadre plus général d'un changement de paradigme dans la pharmacologie, d'une part concernant la rationalisation du fonctionnement des médicaments par la compréhension des interactions moléculaires, et d'autre part en prenant en compte la multiplicité des cibles d'un unique médicament qui au lieu d'être évité doit être comprise et maîtrisée pour définir de nouvelles thérapies.

La problématique spécifique adressée est la prédiction de cibles secondaires d'un ligand, à partir d'une première interaction connue, par similarité entre les cibles. Le principe d'inférence de l'interaction est utilisé pour transposer une information structurale, la similarité géométrique et des propriétés physico-chimiques, en une information bio-chimique, l'interaction prédite. Le problème est ainsi ramené à la recherche de similarité entre les macromolécules, qui sont les cibles connues et prédites de ligands du type des médicaments. Sans définition précise, la similarité entre macromolécules, constitue un vaste champ d'investigation motivé par notre problème initial de recherche de cibles secondaires, mais également par d'autres problèmes connexes comme l'annotation fonctionnelle des protéines ou bien l'étude des interactions entre macromolécules. Par ailleurs d'autres approches affichent également l'objectif de prédire les interactions entre ligands et cibles, notamment par docking. Le problème de docking, très largement étudié est entré depuis longtemps dans les processus de conception de médicaments, offrant un outil inégalé pour mettre à profit les bases de données de ligands dont disposent les industries pharmaceutiques afin de déterminer les meilleurs composés pour une cible donnée. Nous contribuons à montrer que cet outil s'avère en revanche très inefficace dans la problématique complémentaire de l'exploration des cibles, alors que la présence de cibles secondaires toxiques constitue l'une des principales causes d'échec dans les phases plus avancées et plus chère du processus de conception d'un médicament.

Définir la similarité, et donc définir et justifier les modèles utilisés pour représenter la bio-chimie moléculaire, constitue ainsi le point de départ de toute approche par similarité. Nous proposons un modèle de la surface des molécules à l'aide de la théorie des formes alpha, adaptée pour mieux rendre compte de la topologie de la surface. Ce point de vue est motivé par l'existence et l'unicité du modèle, indépendamment de tout système de coordonnées contrairement à toutes les approches basées sur des grilles. Ce modèle permet de définir précisément la représentation d'un site de liaison, ainsi que la surface d'une macromolécule qui doit être explorée. Contrairement à de nombreuses approches nous commençons par définir la mesure de similarité, afin de traduire le problème de recherche en un problème d'optimisation correctement défini. Nous proposons alors une heuristique de résolution pour adresser l'impossibilité d'un parcours exhaustif de l'espace de recherche associé, qui utilise les propriétés de régularité du modèle de surface pour définir une notion de région circulaire permettant un échantillonnage exhaustif, un filtrage rapide, et un point d'ancrage pour l'alignement rapide des sites prédits. Cette méthode nous permet de conserver un temps de calcul raisonnable pour parcourir des bases de données de plusieurs centaines de milliers de macromolécules de manière complète, c'est-à-dire sans définir *a priori* les zones d'interactions.

Notre approche est validée sur un benchmark dont la métrique correspond à l'évaluation d'un classificateur binaire. Nous proposons une mesure de l'AUC complémentée d'une mesure restreinte aux premiers résultats qui correspond mieux à la réponse effectivement attendue dans une utilisation réelle de notre algorithme. Également dans l'objectif de proposer une validation la plus proche possible d'une utilisation réelle, nous complémentons l'usage de deux jeux de données précédemment publiés par un nouveau jeu de données mettant l'accent sur la pertinence des ligands par rapport aux molécules d'intérêts à potentiel

pharmaceutique. Notre approche est comparée à ProBiS, un logiciel récent et performant de recherche de similarité entre protéines, et Vina, un logiciel de docking très populaire. Nous avons souhaité utiliser un logiciel de docking, car bien que la méthode soit connue pour ne pas proposer des scores pertinents entre protéines distinctes aucune étude n'existe à notre connaissance pour vérifier le pouvoir prédictif réel du docking dans notre problématique. Sur chacun de ces jeux de données, BioBind est plus performant que les deux autres approches.

De nombreuses perspectives à court et moyen terme sont envisagées et font actuellement l'objet de développements au sein de l'entreprise Bionext SA à différents stades d'avancement, entre l'idée théorique et la preuve de concept. Elles concernent à la fois le modèle de représentation des molécules, la méthode de résolution du problème d'optimisation, les applications biologiques, ainsi que l'intégration d'autres méthodes structurales comme le docking dans un processus plus complet. Dans chacune de ces directions, nous pensons que la définition de jeux de données fiables et la formulation d'un problème biologique suffisamment précis pour permettre une validation claire sont essentielles. Une telle méthode de validation doit permettre d'une part de comparer les méthodes entre elles du point de vue de la question biologique, et d'autre part de valider l'algorithme de recherche *per se* dans le cadre du modèle. La faiblesse et l'imprécision des jeux de données, ainsi que l'absence de métrique faisant consensus semblent en effet en contradiction avec les enjeux pharmacologiques, en particulier en comparaison avec le problème de docking.

Notre modèle moléculaire a de nombreux avantages, en proposant une structure topologique de surface, et en capturant de manière exacte l'ensemble des atomes en contact avec le solvant. Cependant, la grande sensibilité aux coordonnées atomiques, ainsi que la non prise en compte des atomes légèrement enfouis mais pouvant jouer un rôle lors d'une liaison chimique, motivent une conception prenant en compte ces atomes sous la surface. Il y a essentiellement deux directions pouvant être envisagées : en ajoutant ces atomes enfouis dans le modèle quitte à perdre certaines propriétés géométriques du modèle de surface, ou bien en projetant certaines propriétés chimiques des atomes enfouis sur les sommets représentant les atomes de surface sans ajouter de points dans la représentation.

De nombreuses voies d'amélioration sont également envisagées dans l'algorithme de recherche et l'heuristique des régions circulaires de surface, afin d'en améliorer à la fois la sensibilité et la rapidité. Ces améliorations concernent par exemple l'algorithme d'alignement géométrique, ou bien le choix des paramètres optimaux pour la création des régions circulaires. Enfin la méthode privilégiée actuellement pour la reconstruction du site prédit consiste à étendre une unique région circulaire en un site complet, en effet la technique consistant à considérer plusieurs régions simultanément, chaque région constituant non plus un simple point d'ancrage mais bien le modèle local de la macromolécule n'est pour l'instant pas aussi performante sur nos benchmarks. Ce point de vue continue d'être investigué car il ouvre la voie à une méthode permettant de gérer de manière approchée la flexibilité des molécules indispensable pour traiter des sites entre macromolécules, plus grands et plats.

En dehors de l'algorithme de recherche à proprement dit, il nous semble également important d'adresser une problématique inhérente aux approches de prédictions qui utilisent de très grandes bases de données structurale, la gestion de la redondance et l'incomplétude des données. En effet en considérant la PDB qui constitue l'exemple classique de l'espace de recherche des approches de prédiction de cibles, on observe une forte sur-représentation de certaines macromolécules qui sont souvent présentées dans des conformations légèrement différentes. De plus certains fichiers structuraux sont plus ou moins précis, peuvent contenir des résidus modifiés ou absents, en raison de la méthode expérimentale de résolution des structures. Certaines approches proposent une version dite « non-redondante » de la PDB afin d'adresser une partie de ces problématiques, en considérant uniquement les structures de meilleures qualités pour une séquence donnée. Les deux problématiques essentielles sont de définir précisément ce qui constitue la qualité de la structure, ainsi que de définir la similarité globale des structures. En particulier le choix qui est généralement fait de réduire la redondance par étude des similarités des séquences pose le problème que deux macromolécules

ayant la même séquence peuvent avoir des structures tertiaires très variées, par exemple selon qu'un ligand est présent ou non. Ce choix supprime ainsi une partie de l'exhaustivité de l'espace de recherche.

Un premier prototype d'étude de la similarité globale des macromolécules avait été défini au cours des travaux de la thèse, consistant à considérer la surface des molécules suivant notre modèle, afin de déterminer la variabilité entre deux structures tertiaire uniquement selon ce modèle de surface et non sur la séquence. Nous avons effectivement observé que de nombreuses structures très similaires du point de vu de la séquence exhibent des différences locales importantes, ainsi pour conserver toute la variabilité des structures accessibles la réduction de l'espace de recherche ne pouvait se faire que de manière très limitée. Cependant il apparaît que dans de nombreux cas plusieurs macromolécules, qui peuvent d'ailleurs être très différentes du point de vue global, partagent de larges régions exactement similaires. La factorisation des ces larges régions, afin de proposer une réduction pertinente de l'espace de recherche est ainsi une voie intéressante dans l'amélioration des approches par similarité, par ailleurs il nous parait pertinent d'annoter les régions des structures qui sont au voisinage de résidus supprimés, afin de prévenir certains biais qui ne sont issus que de la méthode de résolution des structures et non de l'approche de comparaison.

Annexe A

# Formes alpha, pondérées, et beta

## Introduction

Les formes alpha ont été introduites en 1983 par H. Edelsbrunner pour définir « la forme d'un ensemble de points dans le plan » [Edelsbrunner 1983]. Elles ont été étendues en dimension supérieure [Edelsbrunner 1994] et pour modéliser un rayon sur les points [Edelsbrunner 1992]. Des implémentations stables et efficaces existent maintenant pour les calculer sur des ensembles de taille importante, de l'ordre de plusieurs millions de points, par exemple dans la librairie Cgal [Project 2016].

Une approche similaire est proposée par D. Kim, les formes beta [Kim 2006, Kim 2010]. Celles-ci ont été développées en particulier dans l'objectif de fournir une meilleure modélisation des molécules telles que les protéines. Des applications ont été proposées concernant la caractérisation des « espaces vides » et la création d'algorithmes de *docking*, comme les programmes BetaVoid [Kim 2014] et BetaDock [Taylor 2011].

L'objectif qui est suivi dans cette annexe est de présenter les formes alpha, les formes alpha pondérées, et les formes beta d'une manière intuitive. Les définitions et concepts présentés sont généralisables en toute dimension, et peuvent s'appliquer à différents types de données pour de nombreuses applications. On se restreint cependant ici à la dimension 3 et on peut toujours supposer qu'on étudie un ensemble d'atomes d'une molécule telle qu'une protéine dont on connaît les positions et les rayons. Par ailleurs une partie de la formalisation des constructions géométriques et les preuves sur les propriétés des structures ne sont par reprises ici, voir [Edelsbrunner 2010] pour les formes alpha, et [Kim 2006, Kim 2010] pour les formes beta. Enfin les algorithmes et leurs implémentations, qui posent souvent le problème du traitement cohérent de certains cas limites, ne seront pas traités, voir par exemple la documentation de la librairie Cgal [Project 2016] et en particulier le traitement des triangulations [Boissonnat 2002].

## Sommaire

# 1 Formes alpha

On se place dans l'espace euclidien de dimension 3, c'est à dire simplement $\mathbb{R}^3$ muni de la norme et de la distance usuelle. On considère un ensemble fini $S$ de points en position générale, c'est à dire que 4 points de $S$ ne sont jamais coplanaires, et 5 points ne peuvent se trouver sur une même sphère. L'abandon de cette hypothèse entraînerait notamment une perte de l'unicité de certaines constructions et une complexification du raisonnement.

**Définition 1.1.** *On définit un $k$-simplexe comme l'enveloppe convexe de $k+1$ points affinement indépendants. Pour $T \subset S$ avec $|T| \leq d+1$ on note $\Delta_T$ le simplexe associé.*

En dimension $d = 3$ il s'agit simplement des points, segments, triangles, et tétraèdres. Avec l'hypothèse de position générale des points, tout ensemble de $k$ points avec $k \leq 4 = d+1$ est affinement indépendant, et définit donc un simplexe. On remarque également que toute face d'un $k$-simplexe est également un ($k$-1)-simplexe pour $k > 0$ : par exemple les trois faces d'un triangle sont les 3 arêtes qui sont également des simplexes.

**Définition 1.2.** *On définit un complexe simplicial comme un ensemble $C$ de simplexes vérifiant :*
- *toute face d'un simplexe de $C$ est également un simplexe de $C$,*
- *l'intersection de deux simplexes de $C$ est soit vide soit une face commune.*

En dimension 2, c'est à dire dans le plan, les 0-simplexes sont les points, les 1-simplexes sont les segments, et les 2-simplexes sont les triangles non-aplatis. Il n'y a pas de 3-simplexes car 4 points du plan sont toujours affinement dépendants. Voir la figure 1 présentant un exemple de complexe simplicial ainsi qu'un contre-exemple d'une famille de simplexes ne constituant pas un complexe simplicial.

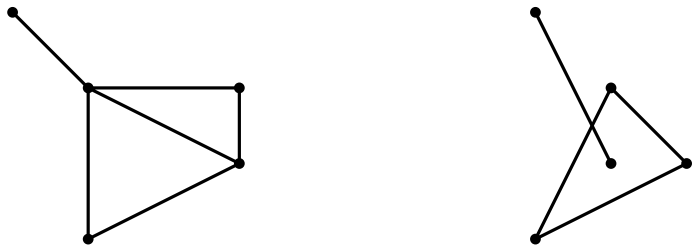


FIGURE 1 – La première figure est un complexe simplicial contenant 2 triangles, 6 segments, et 5 points. La seconde n'en est pas un, par exemple l'intersection du triangle et de la 4$^{\text{e}}$ arête n'est pas une arête du triangle.

## 1.1 Intuition de l'effaceur omniprésent

En paraphrasant la présentation de l'« effaceur omniprésent » dans du polystyrène parsemé de cailloux [Edelsbrunner 1994], on peut décrire la forme alpha de la façon suivante. On imagine l'espace $\mathbb{R}^3$ rempli de glace à la vanille, et contenant un nombre fini de morceaux de chocolat solides, fixes, et de taille négligeable. On suppose qu'on dispose d'une cuillère parfaitement sphérique d'un rayon alpha qui permet d'enlever des sphères de glace tant qu'elle ne contiennent pas de chocolat. En particulier on peut atteindre des parties « internes » de la glace en créant des cavités. On redresse enfin le résultat en remplaçant les encoches sphériques par des triangles. Les formes alpha sont simplement une formalisation de cette intuition.

## 1.2 Triangulation de Delaunay, formes et complexes alpha

On donne une première définition constructive de la forme alpha et du complexe alpha basée sur la triangulation de Delaunay. Pour cela on construit tout d'abord la triangulation de Delaunay qui est un

complexe simplicial avec certaines propriétés, puis on sélectionne un sous-ensemble qui constitue un sous complexe simplicial, selon les définitions suivantes.

**Définition 1.3.** *On dit qu'un simplexe $\Delta_T$ de $S$ vérifie la propriété de Delaunay s'il existe une sphère passant par tous les points de $T$ dont l'intérieur ne contient aucun autre point de $S$. On définit la triangulation de Delaunay* DT *comme l'ensemble des simplexes vérifiant cette propriété.*

En particulier tous les points de $S$ vérifient cette propriété. Pour un tétraèdre, il n'existe qu'une sphère circonscrite et il suffit de vérifier si elle contient d'autres points de $S$. De plus on peut montrer que DT est un complexe simplicial contenant exactement les tétraèdres vérifiant la propriété de Delaunay et leurs facettes. Ce dernier complexe DT est unique si les points sont en position générale.

**Définition 1.4.** *On dit qu'un simplexe $\Delta_T$ de la triangulation de Delaunay* DT *est alpha exposé s'il existe une sphère de rayon $\alpha$ passant par tous les points de $T$ et ne contenant aucun autre point de $S$. L'union de ces simplexes est appelé* forme alpha *de $S$, noté* $\mathrm{A}(\alpha)$*, c'est un polytope union d'un sous complexe simplicial de la triangulation de Delaunay, appelé* complexe alpha *et noté* $\mathrm{K}(\alpha)$.

La figure 2 illustre les concepts en dimension 2.

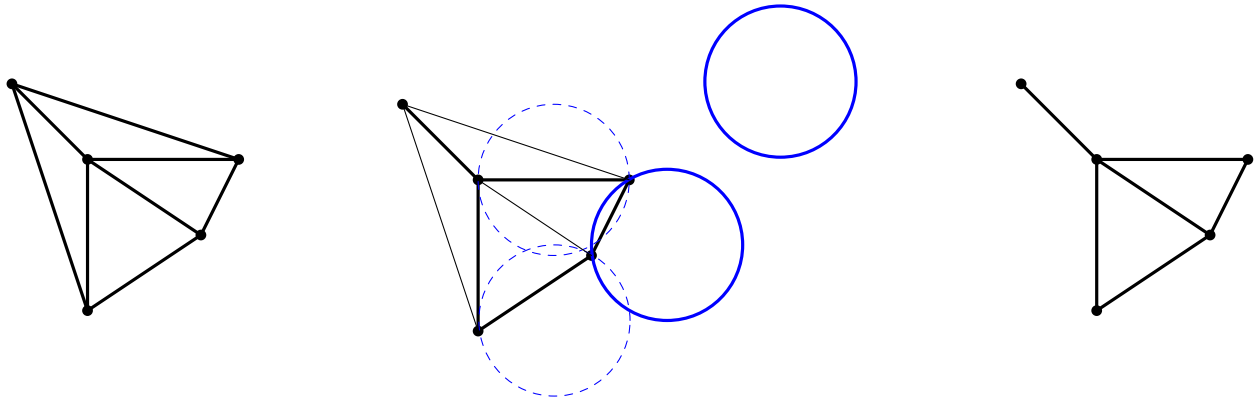


FIGURE 2 – La première figure représente la triangulation de Delaunay de 5 points du plan. La seconde figure détaille les segments alpha exposés définissant la forme alpha à la frontière du complexe alpha. La dernière présente le complexe alpha résultant, dont la frontière est la forme alpha.

## 1.3 Diagramme de Voronoï et lien avec la triangulation de Delaunay

**Définition 1.5.** *Le diagramme de Voronoï de $S$, noté* VD*, est une partition de l'espace $\mathbb{R}^3$ en régions, chacune contenant les points les plus proches d'un point donné de $S$ que de tous les autres points de $S$. On note $\mathcal{V}_s \in \mathrm{VD}$ la région engendré par $s \in S$.*

$$\mathrm{VD} = \left\{\mathcal{V}_s \subset \mathbb{R}^3 : s \in S\right\}$$
$$\mathcal{V}_s = \left\{x \in \mathbb{R}^3 : \forall t \in S,\ \|x - s\| \leq \|x - t\|\right\}$$

Chaque région $\mathcal{V}_s$ est ainsi un polytope, éventuellement non borné, qui est l'intersection de $|S| - 1$ demi-espaces ; chaque demi-espace étant défini par le plan médiateur entre $s$ et chaque $t \in S \setminus \{s\}$. On considère alors l'ensemble des familles de régions qui s'intersectent, le nerf du recouvrement, dont on peut montrer qu'il contient uniquement : des singletons (chaque région), des paires de régions s'intersectant sur une face commune, des triplets de régions s'intersectant sur une arête commune, et des quadruplets de régions s'intersectant sur un point.

$$\mathcal{I} = \left\{ T \subset S : \bigcap_{t \in T} \mathcal{V}_t \neq \emptyset \right\}$$

On construit alors la réalisation géométrique de ce nerf en identifiant chaque région avec le point de $S$ qui la génère. Pour chaque $T \subset \mathcal{I}$ on associe le simplexe $\Delta_T$.

**Propriété 1.1.** *Les simplexes ainsi construits constituent la triangulation de Delaunay de $S$.*

$$\mathrm{DT} = \left\{ \Delta_T : \bigcap_{t \in T} \mathcal{V}_t \neq \emptyset \right\}$$

La figure 3 présente un exemple en dimension 2. Les régions sont des polygones, dont certains sont non bornés. La réalisation du nerf constitue la triangulation de Delaunay.

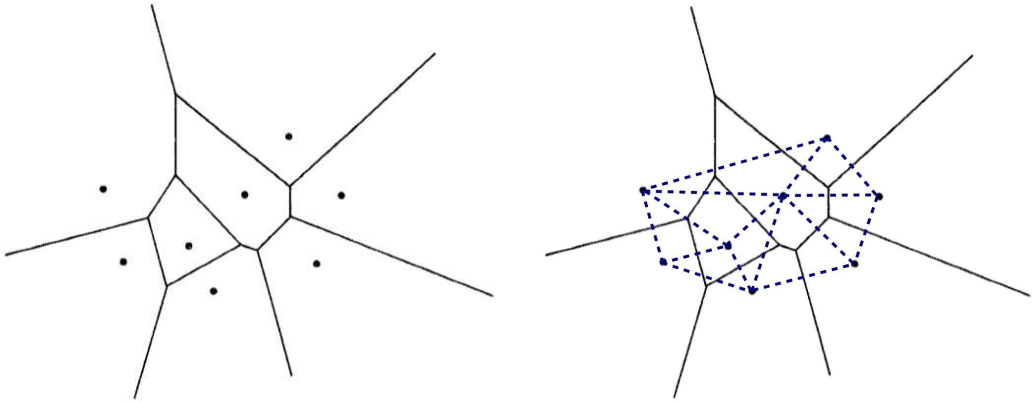


FIGURE 3 – Diagramme de Voronoï en dimension 2 (image de gauche reproduite à partir de [Aurenhammer 1991]). La triangulation de Delaunay associée a été ajoutée sur la seconde figure : on vérifie que pour chaque triangle le cercle circonscrit ne contient aucun autre point de $S$.

## 1.4 Union des sphères et nouvelle définition du complexe alpha

On définit $\mathrm{U}(\alpha)$ l'union des sphères centrées en les points de $S$ et de rayon alpha, puis on définit un diagramme de Voronoï $\mathrm{VD}(\alpha)$ restreint à $\mathrm{U}(\alpha)$.

$$\begin{aligned}
\mathrm{U}(\alpha) &= \left\{ x \in \mathbb{R}^3 : \exists s \in S,\ \|x - s\| \leq \alpha \right\} \\
\mathrm{VD}(\alpha) &= \left\{ \mathcal{V}_s(\alpha) \subset \mathrm{U}(\alpha) : s \in S \right\} \\
\mathcal{V}_s(\alpha) &= \left\{ x \in \mathrm{U}(\alpha) : \forall t \in S,\ \|x - s\| \leq \|x - t\| \right\} \\
\mathrm{K}(\alpha) &= \left\{ \Delta_T : \bigcap_{t \in T} \mathbb{V}(\alpha)_t \neq \emptyset \right\}
\end{aligned}$$

Les intersections des nouvelles régions sont clairement un sous-ensemble des intersections précédentes. Ainsi les simplexes réalisant le nerf de ce recouvrement sont un sous-ensemble de la triangulation de Delaunay, il s'agit en fait exactement du complexe alpha. La figure 4 présente un exemple en dimension 2.

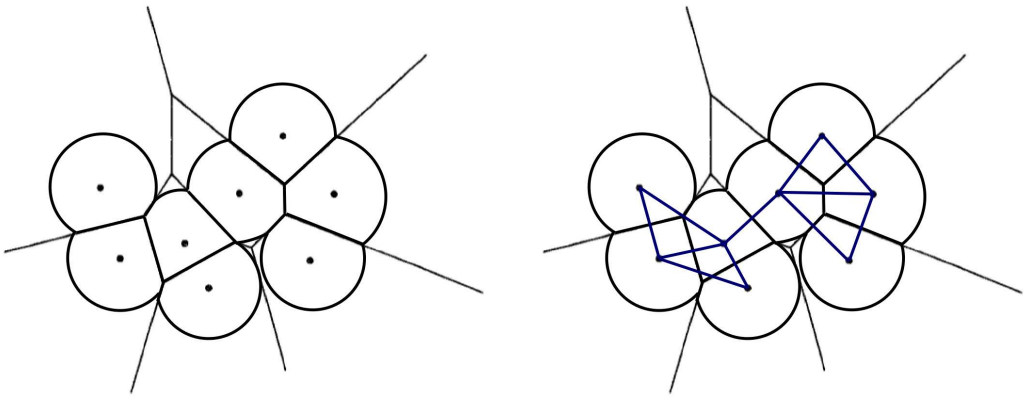


FIGURE 4 – La restriction des régions du diagramme de Voronoï à des sphère de rayon alpha induit un sous-ensemble de la triangulation de Delaunay qui est le complexe alpha.

# 2 Formes alpha pondérées

On présente les formes alpha pondérées d'une manière assez théorique, comme une extension des formes alpha classiques. Ces formes alpha pondérées ont pour objectif principal de modéliser une structure géométrique sur un ensemble de sphères et non plus un ensemble de points. Intuitivement cela signifie, pour reprendre l'analogie de la glace à la vanille, que les morceaux de chocolats peuvent avoir un certain rayon. En particulier en choisissant des rayons nuls, on doit retrouver la forme alpha précédemment décrite.

On pondère ainsi les points de $S$ par des valeurs réelles. On suppose que ces valeurs sont positives, elles modélisent alors des rayons des sphères centrées en des points de $S$. Il est possible de généraliser avec des valeurs négatives, mais ce cas ne sera traité ici car en pratique on s'intéresse principalement à des applications où toutes les pondérations sont positives et modélisent effectivement des rayons.

On note $P : S \to \mathbb{R}$ la fonction de pondération, et $p_s$ la valeur en un point $s \in S$. On définit alors une distance quadratique sur $S$ noté $\pi : S \times S \to \mathbb{R}$ et on associe à chaque point $s$ de $S$ une fonction $\pi_s : \mathbb{R}^3 \to \mathbb{R}$.

$$\forall s, t \in S,\ \pi(s,t) = \|s-t\|^2 - p_s^2 - p_t^2$$
$$\forall s \in S,\ \forall x \in \mathbb{R}^3,\ \pi_s(x) = \|s-x\|^2 - p_s^2$$

## 2.1 Généralisation de la triangulation de Delaunay et du diagramme de Voronoï

On généralise la triangulation de Delaunay en introduisant les pondérations dans la définition de la propriété de Delaunay pondérée, ainsi que dans la définition des simplexes alpha exposés pondérés. On note $\mathrm{DT}(P)$ la triangulation de Delaunay pondérée et $\mathrm{A}(P,\alpha)$ la forme alpha pondérée.

$$\Delta_T \in \mathrm{DT}(P) \text{ si } \exists \omega \in \mathbb{R}^3,\ \lambda \in \mathbb{R},\ \forall t \in T,\ \pi_t(x) = \lambda^2 \text{ et } \forall s \in S \setminus T,\ \pi_s(x) > \lambda^2$$
$$\Delta_T \in \mathrm{A}(P,\alpha) \text{ si } \exists \omega \in \mathbb{R}^3,\ \forall t \in T,\ \pi_t(x) = \alpha^2 \text{ et } \forall t \in S \setminus T,\ \pi_t(x) > \alpha^2$$

Le diagramme de Voronoï se généralise de manière similaire. On note $\mathrm{VD}(P)$ le diagramme de Voronoï

pondéré, et $\mathcal{V}(P)_s$ chaque région.

$$\begin{aligned}
\mathrm{VD}(P) &= \left\{\mathcal{V}(P)_s \subset \mathbb{R}^3 : s \in S\right\} \\
\mathcal{V}(P)_s &= \left\{x \in \mathbb{R}^3 : \forall t \in S,\ \pi_s(x) \leq \pi_t(x)\right\} \\
\mathrm{DT}(P) &= \left\{\Delta_T : \bigcap_{t\in T} \mathcal{V}(P)_t \neq \emptyset\right\}
\end{aligned}$$

De même la définition de l'union des sphères est adaptée, on construit $\mathrm{U}(P,\alpha)$ l'union des sphères pondérées, afin de donner une nouvelle définition du complexe alpha pondéré $\mathrm{K}(P,\alpha)$ qui coïncide encore avec la construction à partir de la triangulation de Delaunay pondérée.

$$\begin{aligned}
\mathrm{U}(P,\alpha) &= \left\{x \in \mathbb{R}^3 : \exists s \in S,\ \pi_s(x) = \alpha^2\right\} \\
\mathrm{VD}(P,\alpha) &= \left\{\mathcal{V}(P,\alpha)_s \subset \mathrm{U}(P,\alpha) : s \in S\right\} \\
\mathcal{V}(P,\alpha)_s &= \left\{x \in \mathrm{U}(P,\alpha) : \forall t \in S,\ \pi_s(x) \leq \pi_t(x)\right\} \\
\mathrm{K}(P,\alpha) &= \left\{\Delta_T : \bigcap_{t\in T} \mathcal{V}(P,\alpha) \neq \emptyset\right\}
\end{aligned}$$

Pour comprendre les définitions précédentes, on commence par remarquer que si toutes les pondérations sont nulles alors on retrouve les définitions originales du cas non pondéré. Ensuite on remarque que les régions du diagramme de Voronoï pondéré sont encore des polytopes intersection de demi-espaces, cependant certaines régions peuvent être vides ce qui signifie en particulier que tous les points de $S$ ne font pas forcément partie de la triangulation de Delaunay pondérée. Enfin, on conserve la propriété selon laquelle le complexe alpha pondéré est toujours un sous complexe simplicial de la triangulation de Delaunay pondérée.

## 2.2 Familles de formes alpha et interprétation du paramètre alpha

En fixant l'ensemble de points $S$ et les pondérations $R$, on fixe la triangulation de Delaunay $\mathrm{DT}(P)$. On considère alors la famille des complexes alpha pour $\alpha \in \mathbb{R}_+$. Par construction on a toujours $\mathrm{A}(P,\alpha) \subset \mathrm{DT}(P)$. De plus pour $\alpha_1 < \alpha_2$, les inclusions suivantes découlent de la définition d'un simplexe alpha exposé.

$$S \subset \mathrm{K}(P,0) \subset \mathrm{K}(P,\alpha_1) \subset \mathrm{K}(P,\alpha_2) \subset \mathrm{K}(P,\infty) = \mathrm{DT}(P)$$

On note en particulier que même si $\alpha$ peut prendre sa valeur dans l'ensemble indénombrable $\mathbb{R}_+$, la famille des complexes alpha est elle finie car $\mathrm{DT}(P)$ est un ensemble fini de simplexes. Cela donne également un sens à la notation $\mathrm{K}(P,\infty)$ car il existe bien $\alpha_0$ tel que pour tout $\alpha > \alpha_0$ on a $\mathrm{K}(P,\alpha) = \mathrm{K}(P,\alpha_0)$ par existence de la limite d'une suite croissante bornée. On précise à nouveau qu'en revanche on a pas toujours $\mathrm{K}(P,0) = S$ en général, contrairement au complexe alpha classique non pondéré.

Lorsque les pondérations sont toutes nulles, c'est à dire pour la forme alpha non pondérée, on a vu que le paramètre alpha modélise le rayon de l'« effaceur omniprésent » qui délimite les contour de la forme des points de $S$. Dans le cas pondéré, l'interprétation est assez différente. Pour mieux comprendre, on étudie de manière exhaustive un très petit exemple dans le plan avec seulement deux points $a = (0,0)$ et $b = (2,0)$.

### 2.2.1 Pondération nulle

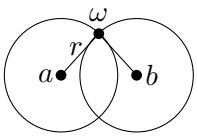
Dans le cas d'une pondération nulle, la valeur d'alpha représente bien le rayon de l'effaceur.

$\alpha < 1$ $\qquad$ $\alpha \geq 1$

$\mathbb{K}(\alpha) = S$  $\mathbb{K}(\alpha) = \mathrm{DT}$ 

### 2.2.2 Très forte pondération

On considère maintenant une pondération homogène ayant une valeur suffisamment grande, $p_a = p_b = r \geq 1$. Dans ce cas, pour toute valeur d'alpha le complexe alpha contient l'arête $[a, b]$ et $\mathrm{K}(P, \alpha) = \mathrm{DT}(P)$, il suffit de vérifier que pour $\alpha = 0$ et $r \geq 1$ le segment $[a, b]$ est bien alpha exposé en considérant $\omega$ tel que $\|\omega - a\| = \|\omega - b\| = r$ (ce point $\omega$ existe dès que $r \geq 1$).



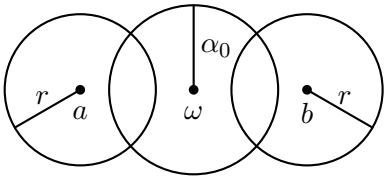

$$\pi_a(\omega) = \|a - \omega\|^2 - p_a^2 = r^2 - r^2 = 0 = \alpha^2$$
$$\pi_b(\omega) = \|b - \omega\|^2 - p_b^2 = r^2 - r^2 = 0 = \alpha^2$$

### 2.2.3 Pondération modérée

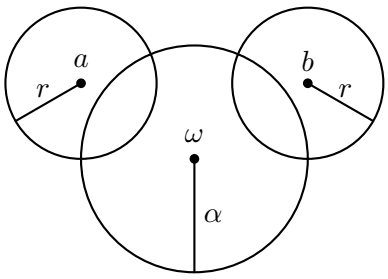
On considère maintenant une pondération homogène non nulle mais plus faible $0 < p_a = p_b = r < 1$ de telle sorte qu'il existe un seuil $\alpha_0$ tel que pour toute valeur d'alpha plus petite la forme alpha ne contient par le segment $[a, b]$ et pour toute valeur d'alpha plus grande elle le contient. On se propose de déterminer la valeur de $\alpha_0$ en étudiant la propriété d'exposition alpha pondérée pour notre exemple $S = \{a, b\}$.

$$\begin{aligned}
[a,b] \text{ alpha exposé si } \exists \omega \in \mathbb{R}^3,\ & \pi_a(\omega) = \pi_b(\omega) = \alpha^2 \\
& \Leftrightarrow \|\omega - a\|^2 - r^2 = \|\omega - b\|^2 - p_b^2 = \alpha^2 \\
& \Leftrightarrow \|\omega - a\|^2 = \|\omega - b\|^2 = \alpha^2 + r^2 \\
& \Leftrightarrow \|\omega - a\| = \|\omega - b\| = \sqrt{\alpha^2 + r^2} \qquad \text{(a)} \\
\text{comme } \|a - b\| = 2 \text{ on doit avoir } & \|\omega - a\| = \|\omega - b\| \geq 1 \\
& \Leftrightarrow \alpha^2 + r^2 \geq 1^2 \\
& \Leftrightarrow \alpha^2 \geq 1 - r^2 \qquad \text{(b)} \\
& \Rightarrow \alpha_0 = \sqrt{1 - r^2} \text{ (avec l'hypothèse } 0 < r < 1) \qquad \text{(c)}
\end{aligned}$$

On remarque que l'équation (b) décrit correctement les cas traités précédemment où $r = 0$ qui implique $\alpha_0 = 1$ et où $r > 1$ qui implique que pour tout $\alpha$ le segment est dans le complexe alpha. On illustre la situation limite $\alpha = \alpha_0$ à l'aide des équations (a), (c) et un second cas où $\alpha > \alpha_0$.

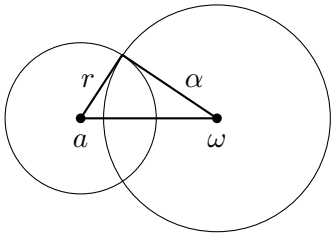



on fixe $r = \frac{2}{3}$
alors $\alpha_0 = \sqrt{1 - r^2} = \frac{\sqrt{5}}{3}$
et $\|\omega - a\| = \|\omega - b\| = \sqrt{\alpha^2 + r^2} = 1$

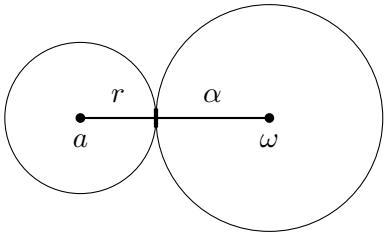



on fixe $r = \frac{2}{3}$ et on choisi $\alpha = 1 > \alpha_0$

La configuration des cercles dans les figures précédentes peut se décrire en définissant l'orthogonalité entre les cercles. On dit que deux cercles sont orthogonaux s'il se coupent en deux points en lesquels leurs tangentes sont orthogonales. La condition d'orthogonalité vient directement de l'équation (a), la figure suivante illustre cette orthogonalité en la comparant à des cercles tangents.




Les deux cercles sont orthogonaux car le triangle est rectangle.




Deux cercles tangents.

### 2.2.4 Bilan sur le paramètre alpha

Après l'étude exhaustive de ce très petit exemple, on constate que l'interprétation géométrique du paramètre alpha parait assez naturelle dès que la pondération est nulle : on peut voir alpha comme le rayon d'une sphère qui vient se plaquer contre les simplexes de $S$ qui seront sélectionnés dans la forme alpha.

En revanche, quand la pondération n'est pas nulle, l'interprétation « naturelle », qui serait de dire qu'on sélectionne les simplexes lorsqu'une sphère de rayon alpha peut se plaquer sur les sphères centrées en les sommets du simplexe, s'avère erronée. En effet cette interprétation s'appuierait sur la tangence de la sphère alpha sur les sphères d'un simplexe, alors que la propriété d'alpha exposition fait intervenir l'orthogonalité de ces sphères qui apparaît comme une « collision ». Pour avoir ce comportement, il faut modifier la pondération en ajoutant alpha puis en considérant la forme zéro ($\alpha = 0$) sur cet ensemble.

On termine cette partie par une formule facilement vérifiable qui lie le paramètre alpha à la pondération, mais qui n'a pas de réciproque évidente :

$$\mathrm{A}(P, \alpha + \lambda) = \mathrm{A}\left(\sqrt{\lambda^2 + P^2}, \alpha\right) \text{ en définissant } \sqrt{\lambda^2 + R^2} : s \mapsto \sqrt{\lambda^2 + p_s^2}$$
$$\mathrm{A}(P + \lambda, \alpha) = \mathrm{A}(P, ?)$$

# 3 Formes beta

## 3.1 Introduction et motivations

Les formes beta ont pour objectif d'étendre les formes alpha pondérées, notamment dans le contexte de la représentation de structures moléculaires telles que les protéines [Kim 2006]. Elle sont également définies à partir d'un ensemble de points $S$, d'une pondération $R : S \to \mathbb{R}$ et d'un paramètre $\beta$. Par analogie avec les complexes et les formes alpha, on notera $\mathrm{B}(R, \beta)$ les formes beta et $\mathrm{C}(R, \beta)$ les complexes beta.

La construction des formes beta est motivée [Kim 2006] par la nécessité de prendre mieux en compte les variations dans la géométrie des données, c'est à dire les variations des rayons des atomes. Les auteurs ont principalement insisté sur les améliorations par rapport aux formes alpha classiques, non pondérées. Les améliorations par rapport aux formes alpha pondérées sont plus subtiles, voir 3.3. La construction repose cependant sur une tesselation de l'espace différente à la fois plus naturelle par rapport à l'intuition de la glace à la vanille et plus complexe que des intersections de demi-espaces, voir 3.2.

## 3.2 Description

Les formes beta sont définies à partir d'un diagramme de Voronoï sur les sphères. Il s'agit d'une extension du diagramme de Voronoï classique puisqu'on considère pour construire la partition de l'espace en régions les distances aux sphères et non les distances aux centres. Comme on le voit dans l'exemple en dimension 2 de la figure 5, la principale différence est que les frontières entre les régions ne sont plus planes (droites en dimension 2) mais coniques (hyperboliques en dimension 2).

Cette partition de l'espace définit une triangulation de l'espace qui est identique à la triangulation de Delaunay pondérée quand les atomes s'intersectent, et qui peut être différentes quand les atomes ne s'intersectent pas. Cela signifie que formes alpha et formes beta sont identiques quand $\alpha = \beta = 0$.

Cette construction permet de définir une notion d'exposition beta des simplexes plus naturelle que l'exposition alpha. Alors qu'un simplexe est exposé alpha si la sphère test est orthogonale aux sphères centrées en les sommets du simplexe (donc avec une certaine « collision »), un simplexe est exposé beta si la sphère test est tangente aux sphères centrées en les sommets du simplexe.

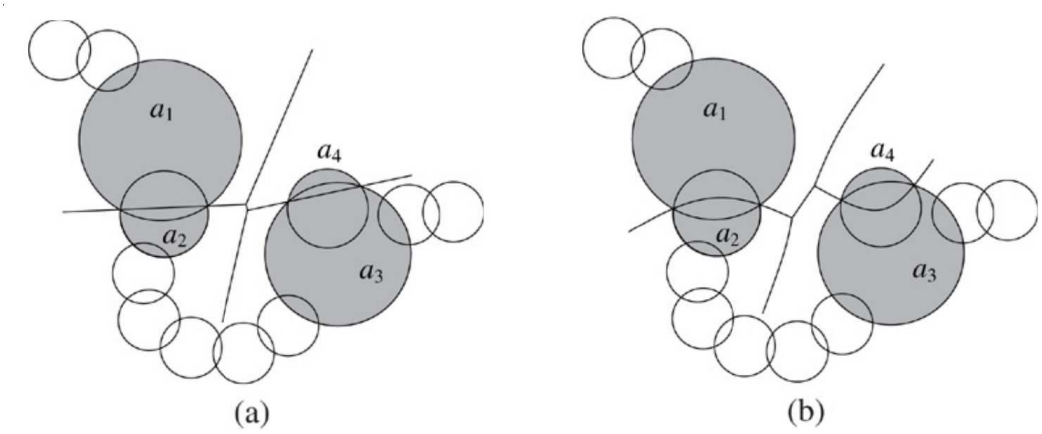



FIGURE 5 – Diagramme de Voronoï pondéré (gauche) et à partir des disques (droite). Les régions entre les atomes $a_2$ et $a_4$ sont adjacentes dans le premier cas, alors que la distance entre les atomes $a_1$ et $a_3$ est plus petite ce qui est pris en compte dans la forme beta. Images reproduites à partir de [Kim 2006].

## 3.3 Familles de formes alpha et de formes beta

De la même manière qu'on peut toujours se ramener au cas où alpha est nul dans les formes alpha pondérées, on peut se ramener au cas où beta est nul dans les formes beta. Cela permet également de montrer que toute forme beta peut se construire comme une forme alpha pour une pondération modifiée, et inversement. Voir notamment les propriétés 8 et 9 de [Kim 2006].

$$\begin{aligned} \mathrm{A}(P,\alpha) &= \mathrm{B}(\sqrt{P^2+\alpha^2},0) \\ \text{et } \mathrm{B}(R,\beta) &= \mathrm{A}(R+\beta,0) \end{aligned}$$

Cela signifie que l'intérêt de considérer les formes beta consiste à étudier une famille entière de formes beta en faisant varier beta. En particulier on peut les construire sans recalculer la tesselation (voir la propriété 10 de [Kim 2006]) alors que le changement de la pondération dans les formes alpha nécessite un nouveau calcul de la triangulation.

## 3.4 Conclusion

Si la pondération des atomes et le rayon de la sphère test sont fixés (ce qui parait naturel dans la modélisation moléculaire), le calcul de la forme beta peut se ramener au cas connu des formes alpha. En revanche l'étude du diagramme de Voronoï pondéré sur les sphères permet d'étudier l'ensemble du partitionnement et de la triangulation (et non seulement les simplexes exposés) pour mieux caractériser les espaces vides.

Annexe B

# Jeux de données pour l'évaluation des performances

## Introduction

Cette annexe reproduit les différents jeux de données utilisés, dont les deux issus de la littérature (Kahraman et Hoffmann) et le jeu de données Lam-On construit par Bionext SA. Les différentes modifications apportées par rapport aux jeux de données originaux sont précisés. Ensuite, les versions et paramètres des logiciels utilisés sont également précisés.

## Sommaire

# 1 Jeu de données KAHRAMAN

## 1.1 Construction du jeu de données

Le jeu de données KAHRAMAN présenté dans la table 1 est construit à partir de [Kahraman 2007] qui décrit pour chaque cible :

- L'identifiant du fichier structural.
- L'identifiant de la chaîne de la protéine cible.
- L'identifiant de la chaîne contenant le ligand.
- L'index du ligand.

On note que dans de nombreux cas, les informations reportées semblent incomplètes ou incorrectes. Le cas échéant, l'identifiant de la chaîne de la protéine est privilégié, puis le site le plus grand pour le ligand donné est considéré.

Par ailleurs les molécules HEM et PO4 ne sont pas considérées comme des ligands et sont simplement retirés. Le fichier 1k87 n'est plus disponible sur la PDB, car il est marqué comme obsolète, le fichier 4o8a est donné en remplacement suivant la recommandation de la PDB.

## 1.2 Les ligands

Les sept ligands du jeu de données sont illustrés dans la figure 1.

| Nom | Formule | Nom complet | Cibles |
|---|---|---|---|
| AMP | $C_{10}H_{14}N_5O_7P$ | ADENOSINE MONOPHOSPHATE | 9 |
| ATP | $C_{10}H_{16}N_5O_{13}P_3$ | ADENOSINE-5'-TRIPHOSPHATE | 14 |
| EST | $C_{18}H_{24}O_2$ | ESTRADIOL | 5 |
| FAD | $C_{27}H_{33}N_9O_{15}P_2$ | FLAVIN-ADENINE DINUCLEOTIDE | 10 |
| FMN | $C_{17}H_{21}N_4O_9P$ | FLAVIN MONONUCLEOTIDE | 6 |
| GLC | $C_6H_{12}O_6$ | ALPHA-D-GLUCOSE | 5 |
| NAD | $C_{21}H_{27}N_7O_{14}P_2$ | NICOTINAMIDE-ADENINE-DINUCLEOTIDE | 15 |

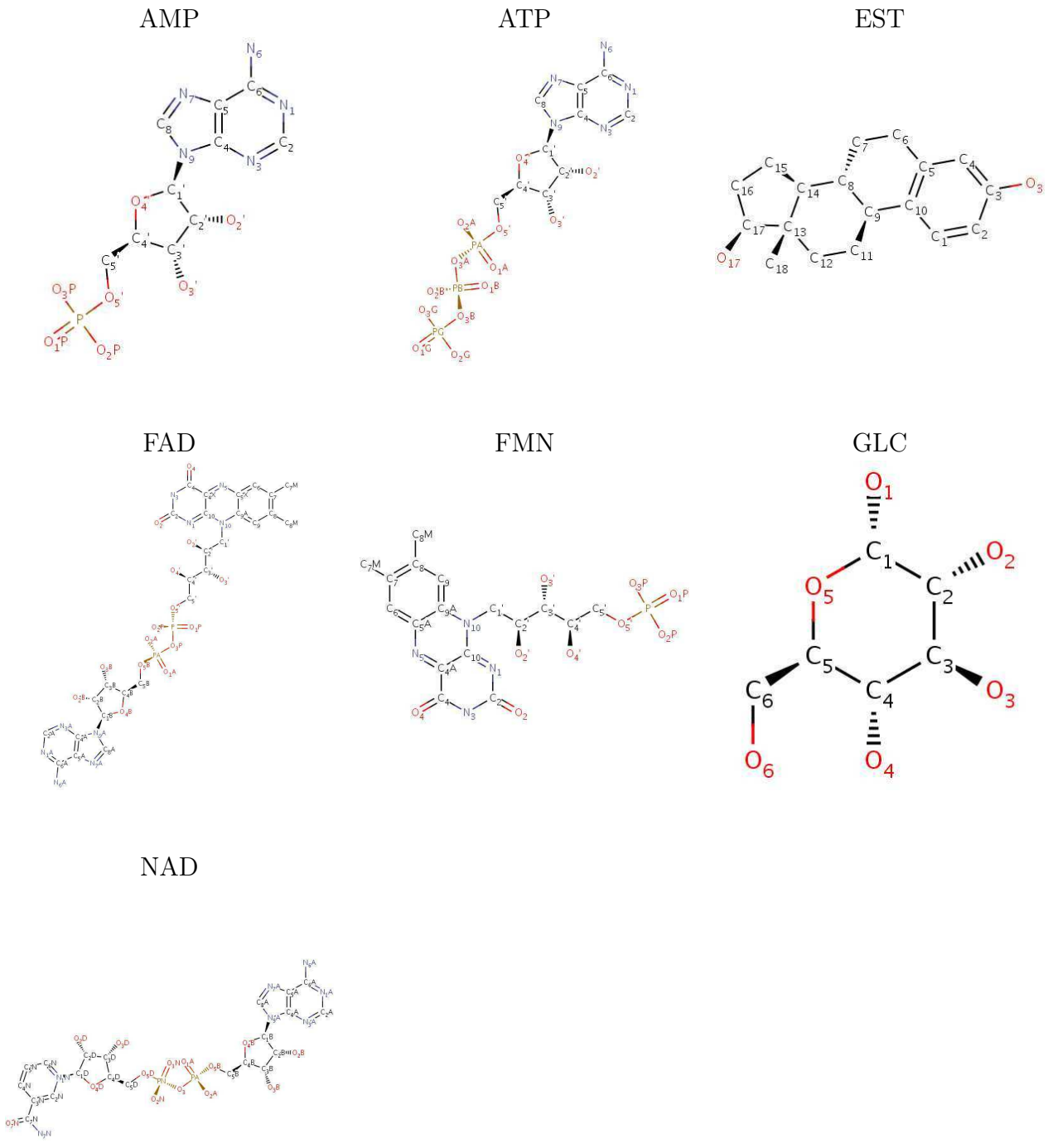



FIGURE 1 – Les ligands du jeu de données KAHRAMAN.

| Ligand | idcode-chain (residue number) | | | |
|---|---|---|---|---|
| AMP | 12as-A(332) | 1amu-A(567) | 1c0a-A(800) | 1ct9-A(1100) |
| | 1jp4-A(601) | 1kht-B(2193) | 1qb8-A(300) | 1tb7-B(1003) |
| | 8gpb-A(930) | | | |
| ATP | 1a0i-A(1) | 1a49-A(535) | 1ayl-A(544) | 1b8a-A(500) |
| | 1dv2-A(1000) | 1dy3-A(200) | 1e2q-A(302) | 1e8x-A(3000) |
| | 1esq-A(300) | 1gn8-B(600) | 1kvk-A(535) | 1o9t-B(1397) |
| | 1rdq-E(600) | 1tid-A(200) | | |
| EST | 1e3r-B(801) | 1fds-A(350) | 1j99-A(401) | 1lhu-A(301) |
| | 1qkt-A(600) | | | |
| FAD | 1cqx-A(405) | 1e8g-B(600) | 1evi-B(353) | 1h69-A(1274) |
| | 1hsk-A(401) | 1jqi-A(399) | 1jr8-B(334) | 4o8a-A(2001) |
| | 1pox-A(612) | 3grs-A(479) | | |
| FMN | 1dnl-A(250) | 1f5v-A(360) | 1ja1-A(1751) | 1mvl-A(1001) |
| | 1p4c-A(490) | 1p4m-A(401) | | |
| GLC | 1bdg-A(501) | 1cq1-A(455) | 1k1w-A(660) | 1nf5-D(527) |
| | 2gpb-A(910) | | | |
| NAD | 1ej2-A(1339) | 1hex-A(400) | 1ib0-A(1994) | 1jq5-A(401) |
| | 1mew-A(987) | 1mi3-A(1350) | 1o04-A(6501) | 1og3-A(1227) |
| | 1qax-B(1001) | 1rlz-A(700) | 1s7g-B(701) | 1t2d-A(323) |
| | 1tox-A(536) | 2a5f-B(1536) | 2npx-A(818) | |

TABLE 1 – Jeu de données KAHRAMAN, adapté à partir de [Kahraman 2007].

# 2 Jeu de données HOFFMANN

## 2.1 Version complète

On considère la table proposée dans [Hoffmann 2010] en considérant :

- L'identifiant du fichier structural.
- L'identifiant de la chaîne de la protéine.
- Le nom du ligand.

À partir de ces informations, on reprend le nom du ligand, les identifiants du fichier et de la chaîne de la protéine, et on détermine le numéro d'index du ligand sur cette chaîne. On note en particulier que ce dernier index n'est pas toujours unique quand le ligand est présent en plusieurs exemplaires sur la même protéine.

L'ensemble du jeu de données proposé dans [Hoffmann 2010] est repris, à l'exception de 4 complexes qui sont retirés :

- Comme précédemment, le fichier structural obsolète 1k87 a été remplacé dans la PDB par le fichier 4o8a. Cependant ce dernier ne contient pas d'occurrence du ligand 1PE.
- Les trois complexes référencés dans les fichiers 2czv, 2hd0, et 1pq2 sur les chaînes respectives C, E, et B dont les ligands n'appartiennent pas à la même chaîne.

## 2.2 Version régularisée

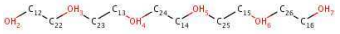
Les répétitions d'un même ligand sur une même protéine sont supprimées, en choisissant le site contenant le nombre d'atomes en interaction avec le ligand le plus important. Le jeu de données résultant est celui qui est référencé comme HOFFMANN dans les *benchmarks*. Il est présenté dans la table 2, reprenant les dix ligands en complexe avec quatre-vingt seize cibles.

## 2.3 Les ligands

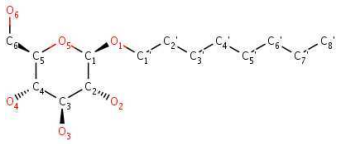
Les dix ligands du jeu de données sont illustrés dans la figure 2.

| Nom | Formule | Nom complet | Cibles |
|---|---|---|---|
| 1PE | $C_{10}H_{22}O_6$ | PENTAETHYLENE GLYCOL | 9 |
| BOG | $C_{14}H_{28}O_6$ | B-OCTYLGLUCOSIDE | 8 |
| GSH | $C_{10}H_{17}N_3O_6S$ | GLUTATHIONE | 10 |
| LDA | $C_{14}H_{31}NO$ | LAURYL DIMETHYLAMINE-N-OXIDE | 10 |
| LLP | $C_{14}H_{22}N_3O_7P$ | N'-PYRIDOXYL-LYSINE-5'-MONOPHOSPHATE | 10 |
| PLM | $C_{16}H_{32}O_2$ | PALMITIC ACID | 10 |
| PMP | $C_8H_{13}N_2O_5P$ | PYRIDOXAMINE-5'-PHOSPHATE | 10 |
| SAM | $C_{15}H_{22}N_6O_5S$ | S-ADENOSYLMETHIONINE | 10 |
| SUC | $C_{12}H_{22}O_{11}$ | SUCROSE | 10 |
| U5P | $C_9H_{13}N_2O_9P$ | URIDINE-5'-MONOPHOSPHATE | 10 |

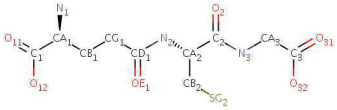
1PE



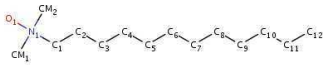
BOG



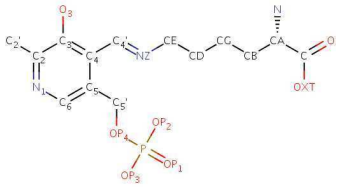
GSH



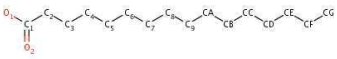
LDA



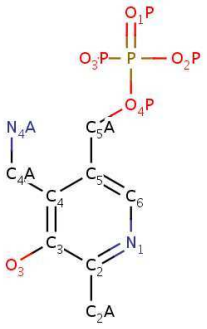
LLP



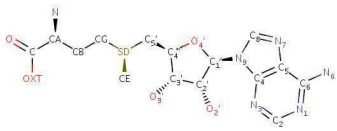
PLM



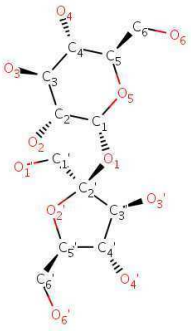
PMP



SAM



SUC



U5P

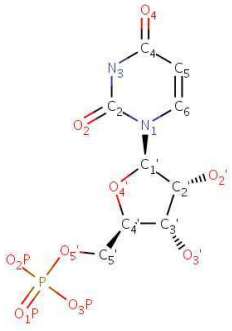


FIGURE 2 – Les ligands du jeu de données HOFFMANN.

| Ligand | idcode-chain (residue number) | | | |
|---|---|---|---|---|
| 1PE | 1g8i-A(951) | 1o57-C(576) | 1q0r-A(2712) | 1s7g-A(504) |
| | 1y10-A(601) | 1zx8-B(125) | 2byn-C(401) | 2haw-B(1001) |
| | 2idb-A(506) | | | |
| BOG | 1aua-A(2) | 1b4w-A(200) | 1fx8-A(475) | 1i78-B(600) |
| | 1k8q-A(1006) | 2p4b-A(500) | 2z73-A(1005) | 3b6h-A(701) |
| GSH | 1dug-B(1587) | 1eem-A(999) | 1fw1-A(217) | 1iyh-A(1200) |
| | 1jlv-A(701) | 1r4w-A(301) | 1y1a-B(601) | 2fls-A(125) |
| | 2imd-A(301) | 2pbj-A(477) | | |
| LDA | 1aij-M(315) | 1ar1-B(274) | 1c8u-A(800) | 1dxr-M(1324) |
| | 1f7s-A(900) | 1kmo-A(742) | 1ojd-A(601) | 1thq-A(200) |
| | 1umx-H(1251) | 1xkw-A(2004) | | |
| LLP | 1a8i-A(680) | 1ax4-A(266) | 1bjw-B(234) | 1bw0-A(253) |
| | 1cl1-A(210) | 1cs1-A(198) | 1d7k-A(69) | 1iug-A(185) |
| | 1j04-A(209) | 1jg8-A(203) | | |
| PLM | 1b56-A(136) | 1eh5-A(430) | 1m66-A(402) | 1mgp-A(314) |
| | 1o6u-A(1398) | 1sz7-A(221) | 2fik-A(502) | 2iu8-B(1349) |
| | 2nwl-A(801) | | | |
| PMP | 1a0g-A(285) | 1aia-A(411) | 1fg7-A(357) | 1kta-A(400) |
| | 1mdo-A(394) | 1uu1-A(1335) | 1zc9-A(502) | 2c81-A(1416) |
| | 2cjg-A(1450) | 2e7u-A(1001) | | |
| SAM | 1cmc-B(105) | 1eiz-A(301) | 1hmy-A(328) | 1i9g-A(301) |
| | 1msk-A(1301) | 1nt2-A(301) | 1nw3-A(500) | 1p91-A(1401) |
| | 1qzz-A(635) | 1r30-A(501) | | |
| SUC | 1a0t-P(1) | 1jgi-A(2064) | 1jj0-A(2380) | 1l0g-A(363) |
| | 1m98-A(401) | 1pt2-A(501) | 1tj4-A(245) | 1uc2-A(1001) |
| | 1w2t-A(1434) | 1ylj-A(1050) | | |
| U5P | 1dbt-A(250) | 1fgx-B(101) | 1g8o-A(474) | 1i5e-A(250) |
| | 1wlj-A(300) | 2b56-A(601) | 2bln-A(1306) | 2bmu-A(1227) |
| | 2c37-B(403) | 2j4j-A(227) | | |

TABLE 2 – Jeu de données HOFFMANN adapté à partir de [Hoffmann 2010].

# 3 Notre nouveau jeu de donnees Lam-On

## 3.1 Les ligands

Les seize ligands du jeu de données sont précisés ci-dessous et illustrés dans la figure 3. Les complexes du jeu de données sont listés dans la table 3.

| Nom | Formule | Nom complet | Cibles |
|---|---|---|---|
| 2FA | $C_{10}H_{12}FN_5O_4$ | 2-fluoroadenosine | 7 |
| 38Z | $C_{33}H_{33}N_9O_2$ | (3R)-. . .-carboxamide | 4 |
| 3AM | $C_{10}H_{14}N_5O_7P$ | 3'-adenylic acid | 7 |
| 444 | $C_{17}H_{12}F_9NO_3S$ | N-. . . BENZENESULFONAMIDE | 6 |
| 537 | $C_{14}H_8N_2O$ | 2,6-DIHYDROANTHRA/1,9-CD/PYRAZOL-6-ONE | 4 |
| 8PR | $C_{19}H_{20}FNO_3$ | Paroxetine | 4 |
| 8XQ | $C_{10}H_7NO_3$ | 8-hydroxyquinoline-5-carboxylic acid | 6 |
| AC2 | $C_8H_{11}N_5O_3$ | 9-HYROXYETHOXYMETHYLGUANINE | 4 |
| AIX | $C_{16}H_{21}N_3O_4S$ | AMPICILLIN (open form) | 4 |
| AZR | $C_{13}H_{19}N_5O_8S_2$ | AZTREONAM (open form) | 5 |
| BAT | $C_{23}H_{31}N_3O_4S_2$ | BATIMASTAT | 7 |
| CTS | $C_8H_{15}NO_4$ | CASTANOSPERMINE | 6 |
| GNH | $C_{10}H_{16}N_6O_{10}P_2$ | AMINOPHOSPHONIC ACID-GUANYLATE ESTER | 7 |
| KSA | $C_{27}H_{21}N_3O_5$ | K-252A | 4 |
| NA7 | $C_{24}H_{25}N_7O_3$ | NAPHTYRIDINE-AZAQUINOLONE | 3 |
| RDF | $C_{23}H_{34}N_3O_{10}P$ | PHOSPHORAMIDON | 7 |

## 3.2 Bruit n-195

Le bruit n-195 est présenté dans la table 4.

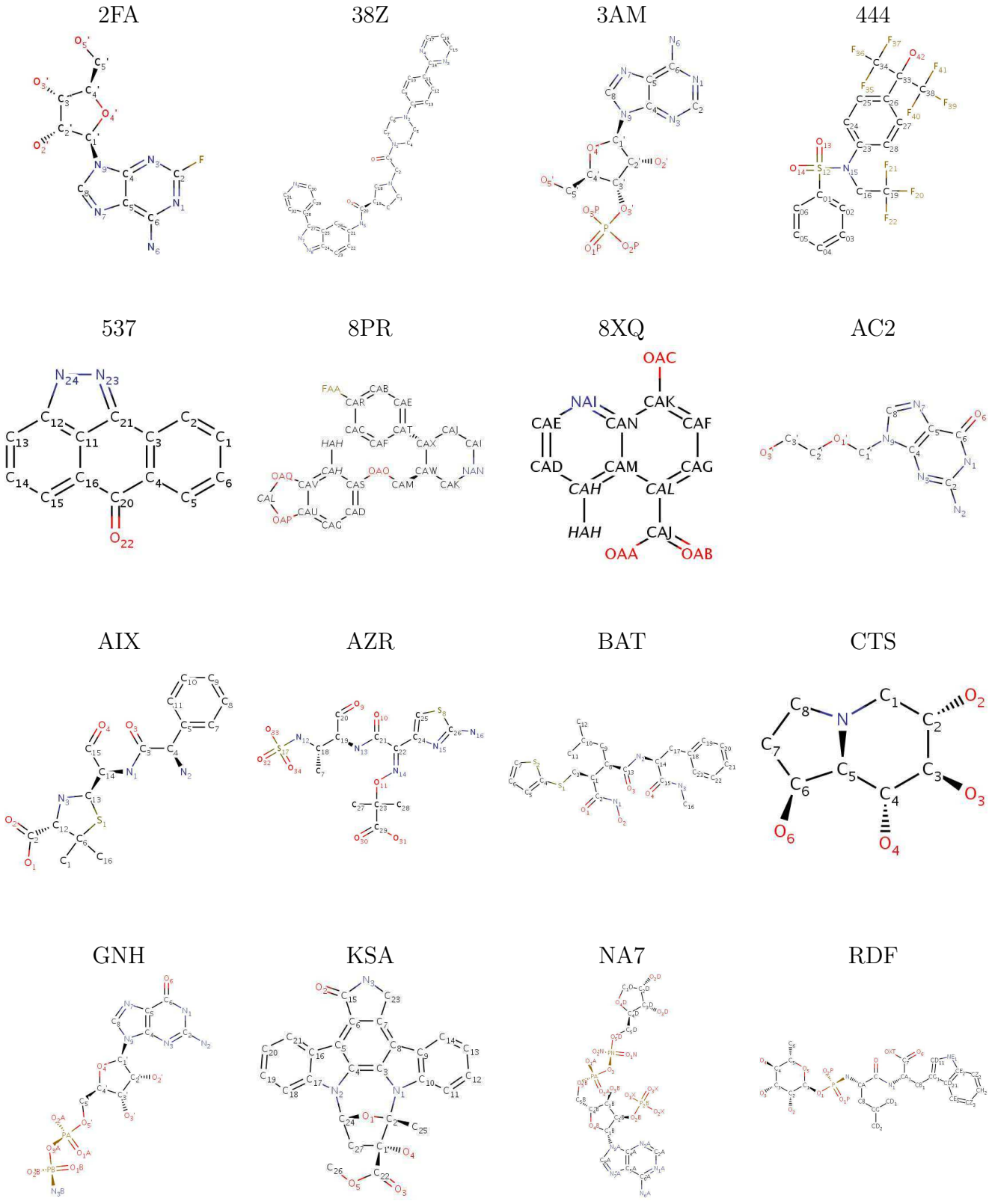



Figure 3 – Les ligands du jeu de données Lam-On.

| Ligand | idcode-chain (residue number) | | | |
|---|---|---|---|---|
| 2FA | 2zj0-A(500) | 2pkk-A(501) | 4ube-A(401) | 4dc3-B(401) |
| | 4dan-A(301) | 1pk9-A(306) | 1z35-A(300) | |
| 38Z | 4qtc-A(804) | 4qtb-A(418) | 4qta-A(411) | 4qtd-A(424) |
| 3AM | 3ago-A(115) | 3agn-A(115) | 3it3-A(343) | 3w2v-A(902) |
| | 3ocw-A(264) | 3c0i-A(338) | 3c0g-A(338) | |
| 444 | 1pqc-A(1500) | 1upv-A(462) | 1upw-A(1462) | 1uhl-B(1002) |
| | 4nb6-A(501) | 2o9i-A(2) | | |
| 537 | 1pmv-A(501) | 4feu-A(302) | 2zmd-A(900) | 1uki-A(0) |
| 8PR | 3v5w-A(701) | 4l9i-A(601) | 4jlt-A(505) | 4mm4-A(603) |
| 8XQ | 2xxz-A(3001) | 3njy-A(360) | 4bio-A(601) | 3od4-A(350) |
| | 4jht-A(301) | 4ie4-A(603) | | |
| AC2 | 1pwy-E(290) | 4da7-A(301) | 3mjr-A(301) | 2ki5-A(1) |
| AIX | 3a3i-A(501) | 3n8l-A(1) | 3ita-A(500) | 3zg8-B(800) |
| AZR | 3ue0-A(998) | 3pbs-A(999) | 4x53-A(301) | 1fr6-A(362) |
| | 2zqc-A(301) | | | |
| BAT | 1mmb-A(1) | 1dth-A(972) | 1jk3-A(900) | 1rm8-A(800) |
| | 4dd8-A(1000) | 2j83-A(996) | 2rjq-A(559) | |
| CTS | 4iif-A(941) | 2cbu-A(1447) | 2vl8-A(1544) | 2jkp-A(1727) |
| | 1eqc-A(401) | 2pwg-A(8000) | | |
| GNH | 1a4r-B(400) | 2efe-B(1200) | 3b1y-A(300) | 3w6n-A(801) |
| | 1hon-A(432) | 1hoo-A(432'A) | 4r98-A(300) | |
| KSA | 1r0p-A(0) | 4wsq-A(405) | 4kik-A(800) | 3eqf-A(1) |
| NA7 | 2uyy-A(1555) | 3dgz-A(501) | 2wzm-A(1284) | |
| RDF | 1dmt-A(750) | 4cth-A(1759) | 3zuk-A(1665) | 3dbk-A(2001) |
| | 4b52-A(501) | 3dwb-A(817) | 1tlp-E(317) | |

TABLE 3 – Jeu de données LAM-ON. Pour chaque ligand, les cibles positives sont définies par le code du fichier structural, l'identifiant de la chaîne de la protéine, et l'index du résidu correspondant au ligand en complexe avec la cible.

| | | | | | |
|---|---|---|---|---|---|
| 4m1x-A | 3lkv-A | 1p7b-A | 3uv5-A | 1zq1-C | 3h6p-B |
| 4d5a-A | 2hzk-D | 4dqq-D | 4kh9-A | 4hl8-A | 4cmy-A |
| 2xov-A | 3i5q-A | 3l18-A | 1qto-A | 2w1n-A | 2yzt-A |
| 3rc6-A | 1z3e-A | 1ym3-A | 3grh-A | 2vsw-A | 2ags-A |
| 3utk-A | 3gqu-A | 1sdo-A | 3wzm-A | 1ko6-A | 3sri-B |
| 3zih-B | 3ifu-A | 1xw3-A | 4f2l-A | 3hoe-A | 4hd1-A |
| 2ra8-A | 3nrk-A | 1wsu-A | 4nn5-B | 3rgu-A | 3aaf-A |
| 4r3f-A | 1pjm-A | 3clm-A | 4nn5-C | 3nke-A | 3txa-A |
| 1am2-A | 4qrn-A | 3iab-B | 4hd4-A | 3bz5-A | 3a1f-A |
| 3ij3-A | 2h7o-A | 4e72-A | 2p9b-A | 3p2h-A | 1f1m-A |
| 4a1r-A | 1mzg-A | 1k8t-A | 1af7-A | 3f8x-B | 2b9d-A |
| 4dcn-C | 4h9d-A | 9pai-B | 3lyg-A | 3ejj-X | 3tbi-A |
| 3akb-A | 2wfv-A | 1zv8-H | 3cip-G | 4yii-A | 3g8r-A |
| 4i61-A | 2vxg-A | 2yjg-A | 2vdu-B | 3g21-A | 1vdr-A |
| 3oa1-A | 1weh-A | 3m50-P | 3phx-A | 4rkk-A | 2doq-D |
| 3i7u-C | 2h9l-A | 4wis-A | 3ep0-A | 2q0o-D | 3ocq-A |
| 3djl-A | 4maq-B | 2qjw-A | 4p94-B | 4hlr-A | 4hny-B |
| 3kf8-B | 2ybx-A | 3dd7-A | 3w88-A | 1dcf-A | 4czw-A |
| 2d2s-A | 1ux5-A | 3j6b-D | 4b7y-A | 3vza-E | 4uqx-A |
| 2zqk-C | 4r4x-A | 3soj-B | 2r5u-B | 3ci0-J | 1v70-A |
| 2fq3-A | 2h9d-D | 3wv6-A | 4qft-A | 1n3l-A | 3oru-A |
| 2clv-A | 3b9j-B | 2b4w-A | 3kio-A | 3s88-J | 3adg-A |
| 3t7z-A | 3lnb-A | 1ryp-I | 2fmm-A | 3dnh-A | 4asc-A |
| 5ame-A | 4rdq-A | 4u10-A | 2pl2-A | 3plx-A | 2beq-D |
| 4cry-G | 2gia-B | 2yln-A | 2hqt-J | 4kib-B | 1ash-A |
| 3mgk-B | 4d0k-C | 1ayo-A | 4pv2-B | 3e23-A | 2qzt-B |
| 2iuh-A | 4bsz-B | 3pq1-A | 3qzp-B | 1ygs-A | 2y9m-B |
| 1t11-A | 4jpr-A | 3nv0-A | 4rdb-A | 2b3g-B | 3mdq-A |
| 1nvm-A | 1qle-D | 3piw-A | 4gu4-A | 1zel-B | 2q18-X |
| 1y96-A | 4g4i-A | 3lgb-B | 2y5q-A | 1sp3-A | 2hs5-A |
| 4bhr-B | 4lws-B | 3zdf-C | 2hyt-A | 1b9l-A | |
| 4ayb-A | 2wgh-A | 2vxz-A | 1tov-A | 2ycl-B | |
| 4jpb-B | 3c1v-B | 4a0p-A | 4hq1-A | 1qw2-A | |

Table 4 – Bruit n-195. Les cent quatre-vingt quinze cibles, considérées comme négatives dans les *benchmarks*.

# 4 Paramètres des logiciels

## 4.1 Paramètres de BioBind

La version de BioBind utilisée est celle de juin 2016, référencée comme suit :

```
program.build.number=7a736739068b3f478878d78938168201f0bc1bb4
config.file=default
```

## 4.2 Paramètres de ProBiS

Le programme ProBiS est compilé à partir des sources dans la version 2.4.2 du 31 août 2012, correspondant à la dernière version disponible sur le site en mai 2016 [1].

```
$ probis -h
Protein Binding Sites by Local Structural Alignments -- Ver. 2.4.2 -- Aug 31 2012 --
```

Dans un premier temps, le logiciel ProBiS a été utilisé en laissant l'ensemble des paramètres par défaut. Il apparaît cependant que dans de nombreux cas cela conduit à avoir un classement très restreint, c'est à dire que la liste des prédits positifs est petite car un grand nombre de cibles candidates ne se voit attribuer aucun score. Pour cette raison deux paramètres ont été modifiés. Le premier paramètre `-nofp` supprime une première étape de filtrage (*Do not filter by fingerprint residues*), et le second paramètre `-zscore -255` abaisse le seuil d'une seconde étape de filtrage (*The cutoff value for z_ score. Low z_ score (<2) means that more insignificant alignments will be outputted.*). Ces modifications permettent de fournir une liste de prédits positifs plus grande sans que les scores et donc le classement de ceux qui étaient déjà retrouvés ne soit changé. C'est cette version qui est référencée comme l'approche ProBiS.

Un autre paramètre a été étudié, parmi les choix laissés à l'utilisateur via l'interface web proposée par les auteurs de l'approche. Il s'agit du paramètre contrôlant la construction du site de liaison requête, et plus précisément le seuil de distance dans la sélection des résidus. Une valeur de 5 Angström est proposée dans l'interface web pour remplacer la valeur par défaut de 3 Angström. Le paramètre `-dist 5` (*The distance between ligand and protein.*) est ainsi ajouté. Ce changement de paramètre modifie le classement, et semble produire de meilleurs résultats selon notre mesure de performance. Cette variante est référencée comme l'approche ProBiS-5.

## 4.3 Paramètres de Vina

Le programme AutoDock/Vina proposant l'approche Vina est directement récupéré sous forme binaire à partir du package des distributions GNU/Linux *Debian* et *Ubuntu*. Il s'agit de la dernière version disponible en mai 2016 [2].

```
$ vina --version
AutoDock Vina 1.1.2 (May 11, 2011)
```

Le logiciel propose une interface simple avec peu de paramètres à définir. Le seul paramètre devant obligatoirement être fourni concerne la donnée d'un parallélépipède rectangle délimitant l'espace de recherche pour le placement du ligand sur une cible candidate. Ce paramètre est déterminé par le parallélépipède contenant l'ensemble de la macromolécules candidate augmenté de la taille du ligand, c'est-à-dire que le docking est réalisé de manière aveugle sans qu'un site de liaison candidat ne soit déterminé *a priori*. Cette approche est simplement référencée comme Vina dans la suite.

Un paramètre optionnel d'*exhaustivité* de la recherche peut être modifié. Il s'agit du paramètre `--exhaustiveness` (*exhaustiveness of the global search - roughly proportional to time*). Il contrôle essentiellement le nombre de valeurs testées par l'algorithme de recherche, de sorte que l'augmentation de

1. probis.cmm.ki.si/?what=parallel, documentation disponible ici : probis.cmm.ki.si/download/ProBiS-2012-Users-Guide.pdf
2. vina.scripps.edu/download.html

ce paramètre accroît la probabilité d'obtenir un alignement optimal du ligand au détriment du temps de calcul. La variante référencée comme Vina-4 consiste à modifier le paramètre en calculant cette valeur comme le quart du nombre total d'atomes de la cible, au lieu de la valeur par défaut 8.

# Glossaire

**ADN** Les acides désoxyribonucléiques sont des macromolécules biologiques, supports de l'information génétique. 3, 4, 6, 7, 9, 91

**ARN** Les acides ribonucléiques sont des macromolécules biologiques, assurant notamment le transfert de l'information génétique pour la construction des protéines, ainsi que d'autres fonctions biologiques dans la cellule. 3, 4, 7, 9, 91

**AUC** (de l'anglais *Area Under the Curve*, Aire Sous la Courbe) Aire délimitée sous la courbe ROC. Cette mesure permet d'estimer la qualité d'un classificateur binaire. 2, 67–69, 71, 75–77, 80–85, 87, 91

**benchmark** (de l'anglais *benchmark*, banc d'essais) Ensemble des jeux de données et méthodes permettant d'évaluer les performances d'une approche informatique, ou de comparer plusieurs méthodes entre elles. 27, 58, 71, 73–75, 86, 87, 109, 115, 125

**cible** Macromolécule biologique susceptible de former un complexe ligand-cible avec une autre petite molécule. Il s'agit des protéines, ADN, et ARN. Pour un ligand donné, une cible est dite positive si il est connu qu'elle peut interagir avec ce ligand, et négative dans le cas contraire. Dans le contexte de la prédiction de cible, une cible est dite candidate si sa capacité à former un complexe est destinée à être évaluée, et sera le cas échéant une cible prédite pour ce ligand. 1–3, 7–12, 14–18, 20, 22–24, 34, 35, 37–41, 45, 50–52, 56, 58, 61–63, 65–67, 69, 71–76, 79, 86–92, 106, 109, 114–117, 123, 125

**classificateur binaire** Algorithme répondant à un problème de classification binaire, c'est-à-dire capable de séparer un ensemble en entrée en deux sous-ensembles distincts. 66–70, 76–78, 89, 91, 123

**complexe ligand-cible** Une molécule ligand et une molécule cible qui interagissent, liées entre elles par un ensemble d'interactions chimiques. 2, 7, 9–11, 22

**conformation** Un conformère, ou la conformation d'une molécule, est un plongement d'une molécule dans l'espace, défini par l'ensemble des coordonnées de chacun de ses atomes et les interactions entre ceux-ci. 4–6, 8–12, 14, 16, 23, 40, 41, 71, 74, 92

**docking** (de l'anglais *docking*, amarrage) Technique consistant à déterminer la meilleure conformation d'un ligand sur une cible donnée pour former le complexe ligand-cible le plus stable. 2, 10–12, 14, 15, 23, 73, 79, 86, 87, 89, 95

**docking aveugle** (de l'anglais *blind docking*, amarrage aveugle) Variante de la technique du docking où le site de liaison n'est pas déterminé préalable sur le récepteur. C'est généralement le cas dans l'application au docking inverse. 11, 12, 14, 86

**docking inverse** Application du docking à la recherche de cibles pour un ligand donné. 10–14, 16, 123

**forme alpha** Polytope basé sur la triangulation de Delaunay d'un ensemble de points dans l'espace. Lorsque ces points représentes les atomes d'une molécule, la forme alpha pondérée par les rayons des atomes modélise la surface accessible au solvant des macromolécules. 2, 17, 18, 32, 33, 35, 37, 41–44, 62, 63, 91, 95–97, 99–104, 123

**in silico** (néologisme littéralement du latin *in silico*, en silice) En utilisant une méthode informatique, par oppositions aux méthodes expérimentales *in vitro* ou *in vivo*. 1–3, 8, 10, 16, 125

**in vitro** (du latin *in vitro*, en éprouvette) En réalisant une expérience sur le vivant, en dehors de l'organisme. 1, 8

**in vivo** (du latin *in vivo*, en vie) En réalisant une expérience sur le vivant, au sein de l'organisme. 1

**interaction chimique** Interaction entre atomes d'une même molécule ou de différentes molécules. Les interactions chimiques sont responsables de la structure des molécules ou des complexes entre différentes molécules. 1–4, 7–9, 11, 12, 14, 15, 17, 19–23, 25, 27, 31, 38, 40, 41, 61, 63, 87, 91, 109, 123

**jeu de données** Ensemble des données représentatives des situations réelles, permettant de réaliser des benchmarks. 2, 14, 18, 23, 31, 59, 63, 65, 71, 73–92, 105–107, 109, 110, 112, 113, 123, 125

**ligand** Petite molécule, il peut s'agir d'un médicament ou d'une autre molécule endogène ou exogène susceptible d'interagir avec d'autres molécules biologiques. 2, 3, 7–12, 14–18, 20, 22, 24, 25, 27, 31, 33–35, 38, 40, 41, 50, 51, 61–63, 65–67, 69, 71, 73–87, 89–91, 93, 106, 107, 109, 110, 112–114, 116, 117, 123

**macromolécule** Les macromolécules biologiques sont les protéines, l'ADN, et l'ARN. Elles se distinguent des autres molécules biologiques par leur taille importante souvent au delà d'un millier d'atomes, et par leur structure constituée d'une répétitions de motifs aussi appelés résidus. 1–11, 14, 16–20, 22–28, 30–35, 37, 38, 40–42, 63, 66, 71, 73, 76, 86, 87, 89, 91–93, 116, 123, 125

**médicament** Ligand exogène ayant un effet thérapeutique de part ses interactions avec les molécules biologiques. 1, 3, 7, 9, 12, 34, 71, 73, 75, 91

**PDB** (en anglais *Protein Data Bank*, Banque de Données de Protéines) Il s'agit de la plus grande base de données publique de structures tridimensionnelles de molécules biologiques, déterminées expérimentalement. 5, 6, 9, 10, 14, 17, 18, 63, 71, 73–75, 89, 90, 92, 106, 109

**polypharmacologie** Étude simultanée de l'ensemble des cibles d'un médicament ou autre composé. 9

**problème de classification binaire** Problème consistant à décider la séparation d'un ensemble donné en deux sous-ensembles selon un critère. Le problème de classification binaire étudié dans ce mémoire consiste à séparer un ensemble de cibles candidates pour distinguer celles prédites pour interagir avec un ligand requête donné des autres cibles. 2, 40, 63, 69

**protéine** Les protéines sont des macromolécules biologiques qui assurent les fonctions biologiques de la cellules. Les protéines sont constitués de chaînes de résidus appelés peptides. 1, 3–7, 9–11, 22, 24, 30–32, 71, 74, 75, 89, 91, 92, 95, 102, 106, 109, 114, 123

**repositionnement** Utilisation d'un médicament précédemment développé, pour une nouvelle indication thérapeutique. 9

**RMSD** (de l'anglais *Root Mean Square Deviation*, Racine de l'Écart Quadratique Moyen) Métrique entre deux ensemble de points, définie comme la racine de la moyenne des distances au carrés. 26, 61, 87

**ROC** (de l'anglais *Receiver Operating Characteristic*, fonction d'efficacité du récepteur) Fonction de la sensibilité par rapport à la spécificité. La courbe est par définition inclue dans le carré unité, et permet de représenter le comportement d'un classificateur binaire. Son aire sous la courbe, AUC, est en particulier une métrique évaluant le classificateur. 2, 67–69, 71, 76, 77, 79–85, 89, 123

**récepteur** Synonyme pour cible, le terme est utilisé dans le contexte du docking. 11, 12

**résidu** Au sein d'une macromolécule, on appelle résidu chaque petite molécule dont la séquence complète forme la macromolécule. Les résidus des protéines sont les peptides, pour l'ARN et l'ADN il s'agit de nucléotides. 4, 6, 11, 17, 22, 24–26, 28, 31, 32, 34, 41, 71, 87, 92, 93, 114, 116

**site de liaison** Région de la macromolécule cible dans un complexe ligand-cible qui permet l'interaction chimique. On parle de site de liaison requête s'il est déterminé par la présence d'un ligand, on appelle site candidat toute région destinée à être évaluée comme site pour un potentiel ligand et sera dans ce cas appelé site prédit. 2, 8, 10–12, 14, 16–20, 22–35, 37, 38, 40, 43, 45, 50–52, 56, 58, 59, 61–63, 71, 73, 74, 76, 77, 86, 87, 89–92, 106, 109, 116, 123, 125

**sous-graphe commun maximal** Graphe isomorphe à un sous graphe de chacun des graphes, de taille maximale. Le problème de la recherche du sous-graphe maximal, NP-complet, est la traduction naturelle du problème de la recherche d'un sous-motif commun entre des objets modélisés par des graphes, telles que des molécules. 28, 30

**structure primaire** Donnée de la séquence des résidus d'une macromolécule. 4, 6, 10

**structure quaternaire** Donnée de la structure formée par l'assemblage de plusieurs macromolécules en interaction. 4, 6

**structure secondaire** Séquence des sous-motifs remarquables d'une macromolécule, tels que les hélices alpha ou feuillets bêta d'une protéine. 4, 6, 30

**structure tertiaire** Donnée de la conformation spatiale d'une macromolécule, définie par la nature et la position de chacun des atomes qui la compose. 4, 6, 7, 10, 22, 73, 93

**triangulation** Une triangulation d'un ensemble de points dans l'espace correspond à un découpage en tétraèdres dont les sommets sont les points de l'ensemble. La triangulation de Delaunay est un exemple de triangulation où les simplexes (tétraèdres) vérifient une propriété dite de Delaunay . 2, 25, 32, 33, 42, 62, 95, 103, 104

**triangulation de Delaunay** Une triangulation est dite de Delaunay si les simplexes (qui sont les tétraèdres) vérifient la propriété de Delaunay, c'est-à-dire que les sphères circonscrites ne contiennent aucun autre sommet de la triangulation. La version pondérée modélise des rayons sur les points et permet de définir la forme alpha pondérée. 96–100, 103, 123

# Table des figures

# Liste des tableaux

# Bibliographie


[Adams 2007] Melanie A. Adams, Michael D L Suits, Jimin Zheng et Zongchao Jia. *Piecing together the structure-function puzzle : Experiences in structure-based functional annotation of hypothetical proteins.* Proteomics, vol. 7, no. 16, pages 2920–2932, 2007. (Cité en page 20.)

[Alberts 2013] Bruce Alberts, Dennis Bray, Karen Hopkin, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts et Peter Walter. Essential cell biology. 2013. (Cité en page 4.)

[An 2005] J. An. *Pocketome via Comprehensive Identification and Classification of Ligand Binding Envelopes.* Molecular & Cellular Proteomics, vol. 4, no. 6, pages 752–761, 2005. (Cité en pages 32 et 34.)

[Angaran 2009] Stefano Angaran, Mary Ellen Bock, Claudio Garutti et Concettina Guerra. *MolLoc : A web tool for the local structural alignment of molecular surfaces.* Nucleic Acids Research, vol. 37, no. SUPPL. 2, pages 565–570, 2009. (Cité en page 24.)

[Apweiler 2004] Rolf Apweiler, Amos Bairoch, Cathy H Wu, Winona C Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, Hongzhan Huang, Rodrigo Lopez, Michele Magrane, Maria J Martin, Darren A Natale, Claire O'Donovan, Nicole Redaschi et Lai-Su L Yeh. *UniProt : the Universal Protein knowledgebase.* Nucleic acids research, vol. 32, no. Database issue, pages D115–9, 2004. (Cité en page 10.)

[Armon 2001] a Armon, D Graur et N Ben-Tal. *ConSurf : an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information.* Journal of molecular biology, vol. 307, no. 1, pages 447–463, 2001. (Cité en page 32.)

[Aurenhammer 1991] Franz Aurenhammer. *Voronoi Diagrams — A Survey of a Fundamental Data Structure.* ACM Computing Surveys, vol. 23, no. 3, pages 345–405, sep 1991. (Cité en page 98.)

[Bajusz 2015] David Bajusz, Anita Racz et Karoly Héberger. *Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations ?* Journal of Cheminformatics, vol. 7, no. 1, pages 1–13, 2015. (Cité en page 16.)

[Baker 2000] N Baker, M Holst et F Wang. *Adaptive Multilevel Finite Element Solution of the Poisson–Boltzmann Equation II . Refinement at Solvent-Accessible Surfaces in Biomolecular Systems.* J. Comput. Chem., vol. 21, no. 15, pages 1343–1352, 2000. (Cité en page 25.)

[Bender 2007] Andreas Bender, Daniel W Young, Jeremy L Jenkins, Martin Serrano, Dmitri Mikhailov, Paul a Clemons et John W Davies. *Chemogenomic data analysis : prediction of small-molecule targets and the advent of biological fingerprint.* Combinatorial chemistry & high throughput screening, vol. 10, pages 719–731, 2007. (Cité en pages 9 et 10.)

[Biasini 2014] Marco Biasini, Stefan Bienert, Andrew Waterhouse, Konstantin Arnold, Gabriel Studer, Tobias Schmidt, Florian Kiefer, Tiziano Gallo Cassarino, Martino Bertoni, Lorenza Bordoli et Torsten Schwede. *SWISS-MODEL : Modelling protein tertiary and quaternary structure using evolutionary information.* Nucleic Acids Research, vol. 42, no. W1, pages 252–258, 2014. (Cité en page 10.)

[Boissonnat 2002] Jean-daniel Boissonnat, Olivier Devillers, Sylvain Pion, Monique Teillaud, Jean-daniel Boissonnat, Olivier Devillers, Sylvain Pion, Monique Teillaud, Mariette Yvinec et Jean-daniel Boissonnat. *Triangulations in CGAL.* Computational Geometry, vol. 22, no. 1-3, pages 5–19, 2002. (Cité en page 95.)

[Bourne 1995] P E Bourne et Others. *The Macromolecular Crystallographic Information File {(mmCIF)}.* Methods in Enzymology, 1995. (Cité en page 41.)

[Brady 2000] G. Patrick Brady et Pieter F W Stouten. *Fast prediction and visualization of protein binding pockets with PASS.* Journal of Computer-Aided Molecular Design, vol. 14, no. 4, pages 383–401, 2000. (Cité en pages 32 et 33.)

[Bredel 2004] Markus Bredel et Edgar Jacoby. *Chemogenomics : an emerging strategy for rapid target and drug discovery.* Nature reviews. Genetics, vol. 5, no. 4, pages 262–275, 2004. (Cité en page 9.)

[Bron 1973] Coen Bron et Joep Kerbosch. *Algorithm 457 : Finding All Cliques of an Undirected Graph.* Commun. ACM, vol. 16, no. 9, pages 575–577, sep 1973. (Cité en pages 30 et 31.)

[Carvalho 2014] L. F. Carvalho, G. Fernandes, M. V O De Assis, J. J P C Rodrigues et M. Lemes Proença. *Digital signature of network segment for healthcare environments support.* Irbm, vol. 35, no. 6, pages 299–309, 2014. (Cité en page 67.)

[Chartier 2015] Matthieu Chartier, Etienne Adriansen et Rafael Najmanovich. *IsoMIF Finder : Online detection of binding site molecular interaction field similarities.* Bioinformatics, vol. 32, no. 4, pages 621–623, 2015. (Cité en pages 24, 25, 31, 74 et 75.)

[Chikhi 2010] Rayan Chikhi, Lee Sael et Daisuke Kihara. *Real-time ligand binding pocket database search using local surface descriptors.* Proteins : Structure, Function and Bioinformatics, vol. 78, no. 9, pages 2007–2028, 2010. (Cité en page 74.)

[Cole 2005] Jason C. Cole, Christopher W. Murray, J. Willem M Nissink, Richard D. Taylor et Robin Taylor. *Comparing protein-ligand docking programs is difficult.* Proteins : Structure, Function and Genetics, vol. 60, no. 3, pages 325–332, 2005. (Cité en page 12.)

[Coleman 2006] Ryan G. Coleman et Kim A. Sharp. *Travel Depth, a New Shape Descriptor for Macromolecules : Application to Ligand Binding.* Journal of Molecular Biology, vol. 362, no. 3, pages 441–458, 2006. (Cité en pages 32 et 33.)

[Connolly 1983] M L Connolly. *Solvent-accessible surfaces of proteins and nucleic acids.* Science (New York, N.Y.), vol. 221, no. 4612, pages 709–713, 1983. (Cité en page 31.)

[Cornell 1995] Wendy D. Cornell, Piotr Cieplak, Christopher I. Bayly, Ian R. Gould, Kenneth M. Merz, David M. Ferguson, David C. Spellmeyer, Thomas Fox, James W. Caldwell et Peter A. Kollman. *A second generation force field for the simulation of proteins, nucleic acids, and organic molecules.* Journal of the American Chemical Society, vol. 117, no. 19, pages 5179–5197, 1995. (Cité en page 12.)

[Crick 1958] F H Crick. *On protein synthesis.* Symposia of the Society for Experimental Biology, vol. 12, pages 138–63, 1958. (Cité en page 6.)

[Crick 1970] Francis Crick. *Central dogma of molecular biology.* Nature, vol. 227, pages 561–563, 1970. (Cité en page 6.)

[Das 2009] Sourav Das, Arshad Kokardekar et Curt M. Breneman. *Rapid comparison of protein binding site surfaces with property encoded shape distributions.* Journal of Chemical Information and Modeling, vol. 49, no. 12, pages 2863–2872, 2009. (Cité en pages 24, 26 et 30.)

[Depolli 2013] Matjaz Depolli, Janez Konc, Kati Rozman, Roman Trobec et Dusanka Janezic. *Exact parallel maximum clique algorithm for general and protein graphs.* Journal of Chemical Information and Modeling, vol. 53, no. 9, pages 2217–2228, 2013. (Cité en page 30.)

[Dill 2012] K. A. Dill et J. L. MacCallum. *The Protein-Folding Problem, 50 Years On.* Science, vol. 338, no. 6110, pages 1042–1046, 2012. (Cité en page 6.)

[Edelsbrunner 1983] Herbert Edelsbrunner, David G. Kirkpatrick et Raimund Seidel. *On the Shape of a Set of Points in the Plane*, 1983. (Cité en pages 41 et 95.)

[Edelsbrunner 1992] H Edelsbrunner. Weighted alpha shapes. 1992. (Cité en page 95.)

[Edelsbrunner 1994] Herbert Edelsbrunner et Ernst P Mücke. *Three-Dimensional Alpha Shapes.* ACM Transactions on Graphics, vol. 13, no. 1, pages 43–72, jan 1994. (Cité en pages 17, 95 et 96.)

[Edelsbrunner 1995] H. Edelsbrunner. *The union of balls and its dual shape.* Discrete & Computational Geometry, vol. 13, no. 1, pages 415–440, 1995. (Cité en pages 41 et 42.)

[Edelsbrunner 2010] Herbert Edelsbrunner. *Alpha shapes-a survey.* Tessellations in the Sciences, pages 1–25, 2010. (Cité en pages 41 et 95.)

[Ewing 2001] T. J A Ewing, Shingo Makino, A. Geoffrey Skillman et Irwin D. Kuntz. *DOCK 4.0 : Search strategies for automated molecular docking of flexible molecule databases.* Journal of Computer-Aided Molecular Design, vol. 15, no. 5, pages 411–428, 2001. (Cité en page 12.)

[Finn 2014] Robert D. Finn, Alex Bateman, Jody Clements, Penelope Coggill, Ruth Y. Eberhardt, Sean R. Eddy, Andreas Heger, Kirstie Hetherington, Liisa Holm, Jaina Mistry, Erik L L Sonnhammer, John Tate et Marco Punta. *Pfam : The protein families database.* Nucleic Acids Research, vol. 42, no. D1, pages 222–230, 2014. (Cité en page 75.)

[Fischer 1894] E. Fischer. *Einfluss der Configuration auf die Wirkung der Enzyme.* Ber. Dtsch. Chem. Ges., vol. 27, pages 2985–2993, 1894. (Cité en page 7.)

[Friesner 2004] Richard A. Friesner, Jay L. Banks, Robert B. Murphy, Thomas A. Halgren, Jasna J. Klicic, Daniel T. Mainz, Matthew P. Repasky, Eric H. Knoll, Mee Shelley, Jason K. Perry, David E. Shaw, Perry Francis et Peter S. Shenkin. *Glide : A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy.* Journal of Medicinal Chemistry, vol. 47, no. 7, pages 1739–1749, 2004. (Cité en page 14.)

[Fuhrmann 2010] Jan Fuhrmann, Alexander Rurainski, Hans Peter Lenhof et D. Neumann. *A new Lamarckian genetic algorithm for flexible ligand-receptor docking.* Journal of Computational Chemistry, vol. 31, no. 9, pages 1911–1918, 2010. (Cité en page 12.)

[Gao 2013] Mu Gao et Jeffrey Skolnick. *APoc : Large-scale identification of similar protein pockets.* Bioinformatics, vol. 29, no. 5, pages 597–604, 2013. (Cité en pages 24 et 25.)

[Glaser 2003] Fabian Glaser, Tal Pupko, Inbal Paz, Rachel E. Bell, Dalit Bechor-Shental, Eric Martz et Nir Ben-Tal. *ConSurf : Identification of functional regions in proteins by surface-mapping of phylogenetic information.* Bioinformatics, vol. 19, no. 1, pages 163–164, 2003. (Cité en page 32.)

[Glaser 2005] Fabian Glaser, Yossi Rosenberg, Amit Kessel, Tal Pupko et Nir Ben-Tal. *The ConSurf-HSSP database : The mapping of evolutionary conservation among homologs onto PDB structures.* Proteins : Structure, Function and Genetics, vol. 58, no. 3, pages 610–617, 2005. (Cité en page 31.)

[Glaser 2006] Fabian Glaser, Richard J. Morris, Rafael J. Najmanovich, Roman A. Laskowski et Janet M. Thornton. *A method for localizing ligand binding pockets in protein structures.* Proteins : Structure, Function and Genetics, vol. 62, no. 2, pages 479–488, 2006. (Cité en page 32.)

[Gold 2014] Victor Gold. International Union of Pure and Applied Chemistry Compendium of Chemical Terminology. 2014. (Cité en page 4.)

[Grindley 1993] Helen M. Grindley, Peter J. Artymiuk, David W. Rice et Peter Willett. *Identification of Tertiary Structure Resemblance in Proteins Using a Maximal Common Subgraph Isomorphism Algorithm.* Journal of Molecular Biology, vol. 229, no. 3, pages 707–721, 1993. (Cité en page 30.)

[Grosdidier 2011] Aurélien Grosdidier, Vincent Zoete et Olivier Michielin. *Fast docking using the CHARMM force field with EADock DSS.* Journal of Computational Chemistry, vol. 32, no. 10, pages 2149–2159, 2011. (Cité en page 12.)

[Halgren 2004] Thomas A. Halgren, Robert B. Murphy, Richard A. Friesner, Hege S. Beard, Leah L. Frye, W. Thomas Pollard et Jay L. Banks. *Glide : A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening.* Journal of Medicinal Chemistry, vol. 47, no. 7, pages 1750–1759, mar 2004. (Cité en page 14.)

[Halgren 2009] Thomas A. Halgren. *Identifying and characterizing binding sites and assessing druggability.* Journal of Chemical Information and Modeling, vol. 49, no. 2, pages 377–389, 2009. (Cité en page 32.)

[Hartmanis 1982] Juris Hartmanis. *Computers and Intractability : A Guide to the Theory of NP-Completeness (Michael R. Garey and David S. Johnson)*, 1982. (Cité en pages 28 et 30.)

[Hartshorn 2007] Michael J. Hartshorn, Marcel L. Verdonk, Gianni Chessari, Suzanne C. Brewerton, Wijnand T M Mooij, Paul N. Mortenson et Christopher W. Murray. *Diverse, high-quality test set for the validation of protein-ligand docking performance.* Journal of Medicinal Chemistry, vol. 50, no. 4, pages 726–741, 2007. (Cité en pages 23 et 73.)

[Hendlich 1997] Manfred Hendlich, Friedrich Rippmann et Gerhard Barnickel. *LIGSITE : Automatic and efficient detection of potential small molecule-binding sites in proteins.* Journal of Molecular Graphics and Modelling, vol. 15, no. 6, pages 359–363, 1997. (Cité en pages 31 et 32.)

[Hoffmann 2010] Brice Hoffmann, Mikhail Zaslavskiy, Jean-Philippe Vert et Véronique Stoven. *A new protein binding pocket similarity measure based on comparison of clouds of atoms in 3D : application to ligand prediction.* BMC bioinformatics, vol. 11, page 99, 2010. (Cité en pages 24, 25, 26, 28, 74, 109 et 111.)

[Hopkins 2008] Andrew L Hopkins. *Network pharmacology : the next paradigm in drug discovery.* Nature chemical biology, vol. 4, no. 11, pages 682–90, 2008. (Cité en pages 1 et 9.)

[Huang 2006] Bingding Huang et Michael Schroeder. *LIGSITEcsc : predicting ligand binding sites using the Connolly surface and degree of conservation.* BMC structural biology, vol. 6, page 19, 2006. (Cité en pages 31 et 32.)

[Hussein 2015] Hiba Abi Hussein, Alexandre Borrel, Colette Geneix, Michel Petitjean, Leslie Regad et Anne Claude Camproux. *PockDrug-Server : A new web server for predicting pocket druggability on holo and apo proteins.* Nucleic Acids Research, vol. 43, no. W1, pages W436–W442, 2015. (Cité en page 34.)

[Hussein 2016] Hiba Abi Hussein, Colette Geneix, Michel Petitjean, Alexandre Borrel, Delphine Flatters et Anne-Claude Camproux. *Global vision of druggability issues : applications and perspectives.* Drug Discovery Today, 2016. (Cité en page 34.)

[Jambon 2003] Martin Jambon, Anne Imberty, Gilbert Deléage et Christophe Geourjon. *A new bioinformatic approach to detect common 3D sites in protein structures.* Proteins : Structure, Function and Genetics, vol. 52, no. 2, pages 137–145, aug 2003. (Cité en pages 24 et 25.)

[Janin 1990] J. Janin et C. Chothia. *The structure of protein-protein recognition sites.* Journal of Biological Chemistry, vol. 265, no. 27, pages 16027–16030, 1990. (Cité en page 22.)

[Janin 2003] J Janin, K Henrick et J Moult. *CAPRI : a critical assessment of predicted interactions.* Proteins : Structure, Function and Genetics, vol. 9, pages 2–9, 2003. (Cité en page 87.)

[Kabsch 1983] Wolfgang Kabsch et Christian Sander. *Dictionary of protein secondary structure : Pattern recognition of hydrogen-bonded and geometrical features.* Biopolymers, vol. 22, no. 12, pages 2577–2637, 1983. (Cité en page 61.)

[Kahraman 2007] Abdullah Kahraman, Richard J. Morris, Roman A. Laskowski et Janet M. Thornton. *Shape Variation in Protein Binding Pockets and their Ligands.* Journal of Molecular Biology, vol. 368, no. 1, pages 283–301, 2007. (Cité en pages 74, 106 et 108.)

[Kasahara 2010] Kota Kasahara, Kengo Kinoshita et Toshihisa Takagi. *Ligand-binding site prediction of proteins based on known fragment-fragment interactions.* Bioinformatics, vol. 26, no. 12, pages 1493–1499, 2010. (Cité en pages 24 et 25.)

[Keiser 2009] Michael J Keiser, Vincent Setola, John J Irwin, Christian Laggner, Atheir I Abbas, Sandra J Hufeisen, Niels H Jensen, Michael B Kuijer, Roberto C Matos, Thuy B Tran, Ryan Whaley, Richard a Glennon, Jérôme Hert, Kelan L H Thomas, Douglas D Edwards, Brian K Shoichet et Bryan L Roth. *Predicting new molecular targets for known drugs.* Nature, vol. 462, no. 7270, pages 175–181, 2009. (Cité en page 10.)

[Kellenberger 2008] Esther Kellenberger, Claire Schalon et Didier Rognan. *How to Measure the Similarity Between Protein Ligand-Binding Sites ?* Current Computer - Aided Drug Design, vol. 4, no. 3, pages 209–220, sep 2008. (Cité en page 26.)

[Kettner 1997] Lutz Kettner et Z Eth. *Designing a Data Structure for Polyhedral Surfaces.* vol. 21957, no. 21957, pages 1–9, 1997. (Cité en page 43.)

[Khanna 2012] Ish Khanna. *Drug discovery in pharmaceutical industry : Productivity challenges and trends.* Drug Discovery Today, vol. 17, no. 19-20, pages 1088–1102, 2012. (Cité en page 1.)

[Kim 2006] Deok Soo Kim, Jeongyeon Seo, Donguk Kim, Joonghyun Ryu et Cheol Hyung Cho. *Three-dimensional beta shapes.* CAD Computer Aided Design, vol. 38, no. 11, pages 1179–1191, nov 2006. (Cité en pages 95, 102 et 103.)

[Kim 2010] Deok Soo Kim, Youngsong Cho, Kokichi Sugihara, Joonghyun Ryu et Donguk Kim. *Three-dimensional beta-shapes and beta-complexes via quasi-triangulation.* CAD Computer Aided Design, vol. 42, no. 10, pages 911–929, oct 2010. (Cité en page 95.)

[Kim 2014] Jae Kwan Kim, Youngsong Cho, Roman A. Laskowski, Seong Eon Ryu, Kokichi Sugihara et Deok Soo Kim. *BetaVoid : Molecular voids via beta-complexes and Voronoi diagrams.* Proteins : Structure, Function and Bioinformatics, vol. 82, no. 9, pages 1829–1849, sep 2014. (Cité en page 95.)

[Kinoshita 2002] Kengo Kinoshita, Jun'ichi Furui et Haruki Nakamura. *Identification of protein functions from a molecular surface database, eF-site.* Journal of Structural and Functional Genomics, vol. 2, no. 1, pages 9–22, 2002. (Cité en pages 23, 24, 25, 26 et 76.)

[Kirchmair 2008] Johannes Kirchmair, Patrick Markt, Simona Distinto, Gerhard Wolber et Thierry Langer. *Evaluation of the performance of 3D virtual screening protocols : RMSD comparisons, enrichment assessments, and decoy selection - What can we learn from earlier mistakes ?* Journal of Computer-Aided Molecular Design, vol. 22, no. 3-4, pages 213–228, 2008. (Cité en page 23.)

[Klabunde 2007] T Klabunde. *Chemogenomic approaches to drug discovery : similar receptors bind similar ligands.* British journal of pharmacology, vol. 152, no. 1, pages 5–7, 2007. (Cité en pages 14 et 20.)

[Klema 1980] Virginia C. Klema et Alan J. Laub. *The Singular Value Decomposition : Its Computation and Some Applications.* IEEE Transactions on Automatic Control, vol. 25, no. 2, pages 164–176, 1980. (Cité en page 59.)

[Konc 2010] Janez Konc et Dušanka Janežič. *ProBiS algorithm for detection of structurally similar protein binding sites by local structural alignment.* Bioinformatics, vol. 26, no. 9, pages 1160–1168, 2010. (Cité en pages 24, 25, 26, 75, 87 et 89.)

[Koshland 1995] Daniel E Koshland. *The Key–Lock Theory and the Induced Fit Theory.* Angewandte Chemie International Edition in English, vol. 33, no. 2324, pages 2375–2378, 1995. (Cité en pages 8 et 22.)

[Koutsoukas 2011] Alexios Koutsoukas, Benjamin Simms, Johannes Kirchmair, Peter J. Bond, Alan V. Whitmore, Steven Zimmer, Malcolm P. Young, Jeremy L. Jenkins, Meir Glick, Robert C. Glen et Andreas Bender. *From in silico target prediction to multi-target drug design : Current databases, methods and applications.* Journal of Proteomics, vol. 74, no. 12, pages 2554–2574, 2011. (Cité en page 10.)

[Kuhn 2010] Harold W. Kuhn. *The Hungarian method for the assignment problem.* In 50 Years of Integer Programming 1958-2008 : From the Early Years to the State-of-the-Art, numéro Logistics Quarterly 2, pages 29–47. 2010. (Cité en page 49.)

[Laskowski 1995] Roman A. Laskowski. *SURFNET : A program for visualizing molecular surfaces, cavities, and intermolecular interactions.* Journal of Molecular Graphics, vol. 13, no. 5, pages 323–330, 1995. (Cité en pages 31 et 32.)

[Laskowski 1996] R A Laskowski, N M Luscombe, M B Swindells et J M Thornton. *Protein clefts in molecular recognition and function.* Protein science : a publication of the Protein Society, vol. 5, no. 12, pages 2438–52, 1996. (Cité en page 31.)

[Laurie 2005] A. T R Laurie et Richard M. Jackson. *Q-SiteFinder : An energy-based method for the prediction of protein-ligand binding sites.* Bioinformatics, vol. 21, no. 9, pages 1908–1916, 2005. (Cité en pages 32 et 34.)

[Lavecchia 2015] Antonio Lavecchia et Carmen Cerchia. *In silico methods to address polypharmacology : Current status, applications and future perspectives.* Drug Discovery Today, vol. 21, no. 2, pages 288–298, 2015. (Cité en pages 1, 9 et 20.)

[Le Guilloux 2009] Vincent Le Guilloux, Peter Schmidtke et Pierre Tuffery. *Fpocket : an open source platform for ligand pocket detection.* BMC bioinformatics, vol. 10, page 168, 2009. (Cité en pages 32, 33 et 34.)

[Levitt 1992] David G. Levitt et Leonard J. Banaszak. *POCKET : A computer graphies method for identifying and displaying protein cavities and their surrounding amino acids.* Journal of Molecular Graphics, vol. 10, no. 4, pages 229–234, 1992. (Cité en page 32.)

[Liang 1998] J Liang, H Edelsbrunner et C Woodward. *Anatomy of protein pockets and cavities : measurement of binding site geometry and implications for ligand design.* Protein science : a publication of the Protein Society, vol. 7, pages 1884–1897, 1998. (Cité en pages 27 et 32.)

[Lipinski 2001] C A Lipinski, F Lombardo, B W Dominy et P J Feeney. *Experimental and computational approaches to estimate solubility and permeability in drug discovery and development setting.* Advanced Drug Delivery Reviews, vol. 46, no. 1–3, pages 3–26, 2001. (Cité en page 73.)

[Loisy 1951] R. Loisy. *Sur la forme des courbes [voir pdf].* Journal de Physique et le Radium, vol. 12, no. 7, pages 735–739, 1951. (Cité en page 42.)

[Lounkine 2012] Eugen Lounkine, Michael J Keiser, Steven Whitebread, Dmitri Mikhailov, Jacques Hamon, Jeremy L Jenkins, Paul Lavan, Eckhard Weber, Allison K Doak, Serge Côté, Brian K Shoichet et Laszlo Urban. *Large-scale prediction and testing of drug activity on side-effect targets.* Nature, vol. 486, no. 7403, pages 361–7, jun 2012. (Cité en page 10.)

[Mamistvalov 1998] Alexander G. Mamistvalov. *N-dimensional moment invariants and conceptual mathematical theory of recognition n-dimensional solids.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 8, pages 819–831, 1998. (Cité en page 29.)

[Milik 2003] Mariusz Milik, Sándor Szalma et Krzysztof A Olszewski. *Common Structural Cliques : a tool for protein structure and function analysis.* Protein engineering, vol. 16, no. 8, pages 543–52, 2003. (Cité en pages 24 et 25.)

[Minot 2015] M Minot, SN Ndiaye et C Solnon. *Recherche d'un plus grand sous-graphe commun par décomposition du graphe de compatibilité.* In jfpc2015.labri.fr, 2015. (Cité en page 30.)

[Mitchell 2015] Alex Mitchell, Hsin Yu Chang, Louise Daugherty, Matthew Fraser, Sarah Hunter, Rodrigo Lopez, Craig McAnulla, Conor McMenamin, Gift Nuka, Sebastien Pesseat, Amaia Sangrador-Vegas, Maxim Scheremetjew, Claudia Rato, Siew Yit Yong, Alex Bateman, Marco Punta, Teresa K. Attwood, Christian J A Sigrist, Nicole Redaschi, Catherine Rivoire, Ioannis Xenarios, Daniel Kahn, Dominique Guyot, Peer Bork, Ivica Letunic, Julian Gough, Matt Oates, Daniel Haft, Hongzhan Huang, Darren A. Natale, Cathy H. Wu, Christine Orengo, Ian Sillitoe, Huaiyu Mi, Paul D. Thomas et Robert D. Finn. *The InterPro protein families database : The classification resource after 15 years.* Nucleic Acids Research, vol. 43, no. D1, pages D213–D221, 2015. (Cité en page 75.)

[Moreira 2007] Irina S Moreira, Pedro A Fernandes et Maria J Ramos. *Hot spots - A review of the protein-protein interface determinant amino-acid residues*, 2007. (Cité en page 22.)

[Morris 2005] Richard J. Morris, Rafael J. Najmanovich, Abdullah Kahraman et Janet M. Thornton. *Real spherical harmonic expansion coefficients as 3D shape descriptors for protein binding pocket and*

*ligand comparisons.* Bioinformatics, vol. 21, no. 10, pages 2347–2355, 2005. (Cité en pages 24, 26 et 28.)

[Murakami 2013] Yoichi Murakami, Kengo Kinoshita, Akira R. Kinjo et Haruki Nakamura. *Exhaustive comparison and classification of ligand-binding surfaces in proteins.* Protein Science, vol. 22, no. 10, pages 1379–1391, oct 2013. (Cité en page 23.)

[Mysinger 2012] Michael M. Mysinger, Michael Carchia, John J. Irwin et Brian K. Shoichet. *Directory of useful decoys, enhanced (DUD-E) : Better ligands and decoys for better benchmarking.* Journal of Medicinal Chemistry, vol. 55, no. 14, pages 6582–6594, 2012. (Cité en page 23.)

[Najmanovich 2008] Rafael Najmanovich, Natalja Kurbatova et Janet Thornton. *Detection of 3D atomic similarities and their use in the discrimination of small molecule protein-binding sites.* Bioinformatics, vol. 24, no. 16, pages 105–111, 2008. (Cité en pages 24, 25, 28 et 74.)

[Nisius 2012] Britta Nisius, Fan Sha et Holger Gohlke. *Structure-based computational analysis of protein binding sites for function and druggability prediction.* Journal of Biotechnology, vol. 159, no. 3, pages 123–134, 2012. (Cité en pages 23 et 75.)

[Osada 2002] Robert Osada, Thomas Funkhouser, Bernard Chazelle et David Dobkin. *Shape distributions.* ACM Transactions on Graphics, vol. 21, no. 4, pages 807–832, 2002. (Cité en pages 29 et 30.)

[Overington 2006] John P Overington, Bissan Al-Lazikani et Andrew L Hopkins. *How many drug targets are there ?* Nature reviews. Drug discovery, vol. 5, no. 12, pages 993–6, 2006. (Cité en page 1.)

[Paolini 2006] Gaia V Paolini, Richard H B Shapland, Willem P van Hoorn, Jonathan S Mason et Andrew L Hopkins. *Global mapping of pharmacological space.* Nature biotechnology, vol. 24, no. 7, pages 805–815, 2006. (Cité en page 9.)

[Paul 2010] Steven M Paul, Daniel S Mytelka, Christopher T Dunwiddie, Charles C Persinger, Bernard H Munos, Stacy R Lindborg et Aaron L Schacht. *How to improve R&D productivity : the pharmaceutical industry's grand challenge.* Nature reviews. Drug discovery, vol. 9, no. 3, pages 203–14, 2010. (Cité en page 1.)

[Peters 1996] K P Peters, J Fauck et C Frömmel. *The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria.* Journal of molecular biology, vol. 256, no. 1, pages 201–213, 1996. (Cité en page 32.)

[Petrek 2006] Martin Petrek, Michal Otyepka, Pavel Banás, Pavlína Kosinová, Jaroslav Koca et Jirí Damborský. *CAVER : a new tool to explore routes from protein clefts, pockets and cavities.* BMC bioinformatics, vol. 7, page 316, 2006. (Cité en pages 32 et 33.)

[Project 2016] The CGAL Project. CGAL User and REference Manual. CGAL Editorial Board, 4.6.2 édition, 2016. (Cité en page 95.)

[Raymond 2002] John W. Raymond et Peter Willett. *Maximum common subgraph isomorphism algorithms for the matching of chemical structures.* Journal of Computer-Aided Molecular Design, vol. 16, no. 7, pages 521–533, 2002. (Cité en page 30.)

[Rowland 1996] R. Scott Rowland et Robin Taylor. *Intermolecular Nonbonded Contact Distances in Organic Crystal Structures : Comparison with Distances Expected from van der Waals Radii.* The Journal of Physical Chemistry, vol. 100, no. 18, pages 7384–7391, 1996. (Cité en page 41.)

[Sael 2010] Lee Sael et Daisuke Kihara. *Binding Ligand Prediction for Proteins Using Partial Matching of Local Surface Patches.* International Journal of Molecular Sciences, vol. 11, no. 12, pages 5009–5026, 2010. (Cité en pages 25 et 74.)

[Sanner 1996] Michel F. Sanner, Arthus J. Olson et Jean-Claude Spehner. *Reduced Surface : An Efficient Way to Compute Molecular Surfaces.* Biopolymers, vol. 38, no. 3, pages 305–320, 1996. (Cité en page 25.)

[Schalon 2008] Claire Schalon, Jean Sébastien Surgand, Esther Kellenberger et Didier Rognan. *A simple and fuzzy method to align and compare druggable ligand-binding sites.* Proteins : Structure, Function and Genetics, vol. 71, no. 4, pages 1755–1778, jun 2008. (Cité en pages 24 et 26.)

[Schmitt 2002] Stefan Schmitt, Daniel Kuhn et Gerhard Klebe. *A new method to detect related function among proteins independent of sequence and fold homology.* Journal of Molecular Biology, vol. 323, no. 2, pages 387–406, 2002. (Cité en pages 24, 25, 26 et 76.)

[Schreiber 2005] Stuart L Schreiber. *Small molecules : the missing link in the central dogma.* Nature chemical biology, vol. 1, no. 2, pages 64–66, 2005. (Cité en page 7.)

[Scrima 2014] Mario Scrima, Gianluigi Lauro, Manuela Grimaldi, Sara Di Marino, Alessandra Tosco, Paola Picardi, Raffaele Riccio, Ettore Novellino, Maurizio Bifulco, Giuseppe Bifulco et Anna Maria D Ursi. *Structural evidence of N-6 isopentenyladenosine as a new ligand of farnesyl pyrophosphate synthase Structural evidence of N-6 isopentenyladenosine as a new ligand of farnesyl pyrophosphate synthase.* Journal of medicinal chemistry, 2014. (Cité en pages 14 et 86.)

[Shulman-Peleg 2004] Alexandra Shulman-Peleg, Ruth Nussinov et Haim J. Wolfson. *Recognition of functional sites in protein structures.* Journal of Molecular Biology, vol. 339, no. 3, pages 607–633, 2004. (Cité en pages 24, 25 et 28.)

[Sommer 2007] Ingolf Sommer, Oliver Muller, Francisco S. Domingues, Oliver Sander, Joachim Weickert et Thomas Lengauer. *Moment invariants as shape recognition technique for comparing protein binding sites.* Bioinformatics, vol. 23, no. 23, pages 3139–3146, dec 2007. (Cité en pages 24, 25, 28 et 29.)

[Sousa 2013] S F Sousa, a J M Ribeiro, J T S Coimbra, R P P Neves, S a Martins, N S H N Moorthy, P a Fernandes et M J Ramos. *Protein-Ligand Docking in the New Millennium – A Retrospective of 10 Years in the Field | BenthamScience.* Current medicinal chemistry, vol. 20, no. 18, pages 2296–314, 2013. (Cité en page 11.)

[Spitzer 2011] Russell Spitzer, Ann E. Cleves et Ajay N. Jain. *Surface-based protein binding pocket similarity.* Proteins : Structure, Function and Bioinformatics, vol. 79, no. 9, pages 2746–2763, 2011. (Cité en page 74.)

[Steffen 2010] Claudia Steffen, Klaus Thomas, Uwe Huniar, Arnim Hellweg, Oliver Rubner et Alexander Schroer. *TmoleX–a graphical user interface for TURBOMOLE.* Journal of computational chemistry, vol. 31, no. 16, pages 2967–2970, 2010. (Cité en page 12.)

[Taylor 2011] Publisher Taylor, Deok-soo Kim, Chong-min Kim, Chung-in Won, Jae-kwan Kim, Joonghyun Ryu, Youngsong Cho, Changhee Lee et Jong Bhak. *BetaDock : Shape-Priority Docking Method Based on BetaDock : Shape-Priority Docking Method Based on.* Journal of Biomolecular Structure and Dynamics, vol. 29, no. June 2014, pages 37–41, 2011. (Cité en page 95.)

[Trott 2010] Oleg Trott et Arthur J. Olson. *Software news and update AutoDock Vina : Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading.* Journal of Computational Chemistry, vol. 31, no. 2, pages 455–461, 2010. (Cité en pages 12, 75 et 89.)

[Volkamer 2012] Andrea Volkamer, Daniel Kuhn, Friedrich Rippmann et Matthias Rarey. *Dogsitescorer : A web server for automatic binding site prediction, analysis and druggability assessment.* Bioinformatics, vol. 28, no. 15, pages 2074–2075, 2012. (Cité en page 32.)

[Vulpetti 2012] Anna Vulpetti, Tuomo Kalliokoski et Francesca Milletti. *Chemogenomics in drug discovery : computational methods based on the comparison of binding sites.* Future medicinal chemistry, vol. 4, no. 15, pages 1971–9, 2012. (Cité en page 16.)

[Wang 2004] Junmei Wang, Romain M. Wolf, James W. Caldwell, Peter A. Kollman et David A. Case. *Development and testing of a general Amber force field.* Journal of Computational Chemistry, vol. 25, no. 9, pages 1157–1174, 2004. (Cité en page 12.)

[Wang 2012] Wei Wang, Xi Zhou, Wanlin He, Yi Fan, Yuzong Chen et Xin Chen. *The interprotein scoring noises in glide docking scores.* Proteins : Structure, Function and Bioinformatics, vol. 80, no. 1, pages 169–183, 2012. (Cité en pages 14 et 86.)

[Wang 2015] Caihua Wang, Juan Liu, Fei Luo, Zixing Deng et Qian-Nan Hu. *Predicting target-ligand interactions using protein ligand-binding site and ligand substructures.* BMC Systems Biology, vol. 9, no. Suppl 1, page S2, 2015. (Cité en pages 24 et 25.)

[Watson 2005] James D. Watson, Roman A. Laskowski et Janet M. Thornton. *Predicting protein function from sequence and structural data.* Current Opinion in Structural Biology, vol. 15, no. 3 SPEC. ISS., pages 275–284, 2005. (Cité en page 20.)

[Weill 2010] Nathanael Weill et Didier Rognan. *Alignment-free ultra-high-throughput comparison of druggable protein-ligand binding sites.* Journal of Chemical Information and Modeling, vol. 50, no. 1, pages 123–135, 2010. (Cité en pages 24, 25 et 28.)

[Weisel 2007] Martin Weisel, Ewgenij Proschak et Gisbert Schneider. *PocketPicker : analysis of ligand binding-sites with shape descriptors.* Chemistry Central journal, vol. 1, page 7, jan 2007. (Cité en pages 32 et 33.)

[Weisel 2009] Martin Weisel, Ewgenij Proschak, Jan M. Kriegl et Gisbert Schneider. *Form follows function : Shape analysis of protein cavities for receptor-based drug design.* Proteomics, vol. 9, no. 2, pages 451–459, 2009. (Cité en page 33.)

[Wishart 2008] David S. Wishart, Craig Knox, An Chi Guo, Dean Cheng, Savita Shrivastava, Dan Tzur, Bijaya Gautam et Murtaza Hassanali. *DrugBank : A knowledgebase for drugs, drug actions and drug targets.* Nucleic Acids Research, vol. 36, no. SUPPL. 1, pages 901–906, 2008. (Cité en page 10.)

[Wold 1987] Svante Wold, Kim Esbensen et Paul Geladi. *Principal component analysis.* Chemometrics and Intelligent Laboratory Systems, vol. 2, no. 1-3, pages 37–52, 1987. (Cité en page 59.)

[Yin 2009] Shuangye Yin, Elizabeth A. Proctor, Alexey A. Lugovskoy et Nikolay V. Dokholyan. *Fast screening of protein surfaces using geometric invariant fingerprints.* Proceedings of the National Academy of Sciences, vol. 106, no. 39, pages 16622–16626, 2009. (Cité en pages 24, 25, 26 et 27.)

[Yuan 2013] Yaxia Yuan, Jianfeng Pei et Luhua Lai. *Binding site detection and druggability prediction of protein targets for structure-based drug design.* Current pharmaceutical design, vol. 19, no. 12, pages 2326–33, 2013. (Cité en pages 27 et 32.)

[Zhong 2007] Shijun Zhong et Alexander D. Mackerell. *Binding response : A descriptor for selecting ligand binding site on protein surfaces.* Journal of Chemical Information and Modeling, vol. 47, no. 6, pages 2303–2315, 2007. (Cité en pages 32 et 33.)

[Zhu 2015] Xiaolei Zhu, Yi Xiong et Daisuke Kihara. *Large-scale binding ligand prediction by improved patch-based method Patch-Surfer2.0.* Bioinformatics, vol. 31, no. 5, pages 707–713, 2015. (Cité en pages 24, 25 et 26.)

université PARIS-SACLAY
ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

**Titre : Algorithmes pour la prédiction *in silico* d'interactions par similarité entre macromolécules biologiques**

**Mots clés :** Formes alpha ; Surface des macromolécules ; Algorithmes géométriques ; Similarité des macromolécules ; Prédiction d'interactions ; Biologie structurale

**Résumé :** Un médicament, ou tout autre petite molécule biologique, agit sur l'organisme via des interactions chimiques qui se produisent avec d'autres macromolécules telles que les protéines qui régissent le fonctionnement des cellules. La détermination de l'ensemble des cibles, c'est à dire de l'ensemble des macromolécules susceptibles de lier une même molécule, est essentielle pour mieux comprendre les mécanismes moléculaires à l'origine des effets d'un médicament. Cette connaissance permettrait en effet de guider la conception d'un composé pour éviter au mieux les effets secondaires indésirables, ou au contraire découvrir de nouvelles applications à des molécules connues. Les avancées de la biologie structurale nous permettent maintenant d'avoir accès à un très grand nombre de structures tridimensionnelles de protéines impliquées dans ces interactions, ce qui motive l'utilisation d'outils in silico (informatique) pour complémenter ou guider les expériences in vitro ou in vivo plus longues et plus chères.

La thèse s'inscrit dans le cadre d'une collaboration entre le laboratoire DAVID de l'Université de Versailles-Saint-Quentin, et l'entreprise Bionext SA qui propose une suite logicielle permettant de visualiser et d'étudier les interactions chimiques. Les travaux de recherches ont pour objectif de développer un algorithme permettant, à partir des données structurales des protéines, de déterminer des cibles potentielles pour un composé donné. L'approche choisie consiste à utiliser la connaissance d'une première interaction entre un composé et une protéine afin de rechercher par similarité d'autres protéines pour lesquelles on peut inférer la capacité à se lier avec le même composé. Il s'agit plus précisément de rechercher une similarité locale entre un motif donné, qui est la région permettant à la cible connue de lier le composé, et un ensemble de protéines candidates.

Un algorithme a été développé, BioBind, qui utilise un modèle des surfaces des macromolécules issu de la théorie des formes alpha afin de modéliser la surface accessible ainsi qu'une topologie sur cette surface permettant la définition de régions en surface. Afin de traiter le problème de la recherche d'un motif en surface, une heuristique est utilisée consistant à définir des motifs réguliers qui sont une approximation de disques géodésiques et permettant un échantillonnage exhaustif à la surface des macromolécules. Ces régions circulaires sont alors étendues à l'ensemble du motif recherché afin de déterminer une mesure de similarité.

Le problème de la prédiction de cibles est ramené à un problème de classification binaire, où il s'agit pour un ensemble de protéines données de déterminer lesquelles sont susceptibles d'interagir avec le composé considéré, par similarité avec la première cible connue. Cette formalisation permet d'étudier les performances de notre approche, ainsi que de la comparer avec d'autres approches sur différents jeux de données. Nous utilisons pour cela deux jeux de données issus de la littérature ainsi qu'un troisième développé spécifiquement pour cette problématique afin d'être plus représentatif des molécules pertinentes du point de vue pharmacologique, c'est-à-dire ayant des propriétés proches des médicaments. Notre approche se compare favorablement sur ces trois jeux de données par rapport à une autre approche de prédiction par similarité, et plus généralement notre analyse confirme que les approches par docking (amarrage) sont moins performantes que les approches par similarité pour le problème de la prédiction de cibles.

**Université Paris-Saclay**
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France

université PARIS-SACLAY
ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

**Title : Similarity based algorithms for the prediction of interactions between biomolecules**

**Keywords :** Alpha shapes ; Macromolecular surface ; Geometric algorithms ; Macromolecular similarities ; Interaction prediction ; Structural biology

**Abstract :** The action of a drug, or another small biomolecule, is induced by chemical interactions with other macromolecules such as proteins regulating the cell functions. The determination of the set of targets, the macromolecules that could bind the same small molecule, is essential in order to understand molecular mechanisms responsible for the effects of a drug. Indeed, this knowledge could help the drug design process so as to avoid side effects or to find new applications for known drugs. The advances of structural biology provides us with three-dimensional representations of many proteins involved in these interactions, motivating the use of in silico tools to complement or guide further in vitro or in vivo experiments which are both more expensive and time consuming.

This research is conducted as part of a collaboration between the DAVID laboratory of the Versailles-Saint-Quentin University, and Bionext SA which offers a software suite to visualize and analyze chemical interactions between biological molecules. The objective is to design an algorithm to predict these interactions for a given compound, using the structures of potential targets. More precisely, starting from a known interaction between a drug and a protein, a new interaction can be inferred with another sufficiently similar protein. This approach consists in the search of a given pattern, the known binding site, across a collection of macromolecules.

An algorithm was implemented, BioBind, which rely on a topological representation of the surface of the macromolecules based on the alpha shapes theory. Our surface representation allows to define a concept of region of any shape on the surface. In order to tackle the search of a given pattern region, a heuristic has been developed, consisting in the definition of a regular region which is an approximation of a geodesic disk. This circular shape allows for an exhaustive sampling and fast comparisons, and any circular region can then be extended to the full original pattern to provide a similarity evaluation with the query binding site.

The target prediction problem is formalised as a binary classification problem, where a set of macromolecules is being separated between those predicted to interact and the others, based on their local similarity with the known target. With this point of view, classic metrics can be used to assess performance, and compare our approach with others. Three datasets were used, two of which were extracted from the literature and the other one was designed specifically for our problem emphasizing the pharmacological relevance of the chosen molecules. Our algorithm proves to be more efficient than another state-of-the-art similarity based approach, and our analysis confirms that docking softwares are not relevant for our target prediction problem when a first target is known, according to our metric.

**Université Paris-Saclay**
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France