UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE DES SCIENCES CHIMIQUES

Laboratoire d'Innovation Thérapeutique, UMR 7200

THÈSE

présentée par :

Jérémy DESAPHY

soutenue le : **09 Octobre 2013**

pour obtenir le grade de : **Docteur de l'université de Strasbourg**

Discipline/ Spécialité : Chimie/Chémoinformatique

L'analyse structurale de complexes protéine/ligand et ses applications en chémogénomique

THÈSE dirigée par :

Dr. ROGNAN Didier	Directeur de recherche, CNRS

RAPPORTEURS :

Pr. BONNET Pascal	Professeur de l'université d'Orléans
Dr. DOUGUET Dominique	Chargé de recherche, CNRS

AUTRES MEMBRES DU JURY :

Dr. DUCROT Pierre	Chercheur, Institut de Recherches SERVIER
Pr. VARNEK Alexandre	Professeur de l'université de Strasbourg

Remerciements

Il y a tant de personnes à remercier, tant d'individus qui m'ont apporté tellement, que pour n'oublier personne je ferais cela par ordre chronologique. Soyez certains que les mots qui suivent ne sont pas assez forts pour exprimer ma gratitude, mais il n'en existe pas de meilleurs.

Ma famille est sûrement celle à qui je dois le plus et je ne sais guère par où commencer. Mes grands-parents, dont l'amour inconditionnel est sans faille. Ma grand-mère, celle pour qui je suis là aujourd'hui, avec le rêve un peu naïf de trouver un jour un médicament pour la revoir marcher. Mes grands-pères, hommes formidables dont l'un est partit trop tôt et l'autre possédant un mental inégalable. J'aimerais avoir leur force et j'espère que mon travail saura être à leurs hauteurs. Mes parents, qui m'ont soutenu durant toutes ces années, mais qui m'ont surtout aimé. À mon père, qui reste mon héros. À ma mère, éternellement présente pour moi. À mon parrain, celui qui m'a transmis sa passion pour la science. À mes cousines, cousins, tantes, oncles, à cette famille que je vois peu mais qui est constamment dans mon cœur.

À Jessica, mon amie, mon soutien, celle qui m'a toujours redonné le sourire dans les moments difficiles et qui continue à me donner de l'espoir dans l'humanité.

À mes frères et sœurs d'armes de prépa, Régis, Stéphane, Chicky, Julien, Fred, Sophie, Luke, Rémi, Sébastien, Soizic, Sophie, Vosgien et tant d'autres. Plus qu'une bande d'amis, j'ai trouvé une seconde famille (bien que la première soit excellente, je vous rassure !).

On dit que la vie est faite d'opportunités, de rencontres, d'évènements. Une en particulière me restera à l'esprit, celle de mon premier jour à Strasbourg en tant que « chercheur ». Merci à Georges Wipff pour m'avoir donné cette chance unique et d'avoir partager son monde, sa passion avec moi. Il en va de même pour Rachel Schurhammer, Etienne Engler et Alain Chaumont.

À Alexandre Varnek et Gilles Marcou, grâce auxquels j'ai pu découvrir ce vaste univers qu'est la chémoinformatique et qui m'ont permis de trouver ma place dans le

monde scientifique. Je tiens aussi à remercier tous les membres du laboratoire de chémoinformatique : Ioana, Laurent, Christophe, Elena, Fanny, Tatiana, Fiorella. À Evgeny, partit trop tôt et à Aurélie, j'espère que tu te portes bien.

Merci à mes collègues et amis du master : Vinca Prana, Florent Chevillard et Cédric Moretti.

Je remercie les laboratoires Servier pour avoir financer ma thèse et à Pierre Ducrot, pour m'avoir donné une opportunité de prouver ma valeur. J'espère ne pas vous avoir déçu. Merci à mon alpha-testeur, Eric Raimbault, pour sa gentillesse et sa capacité à voir l'invisible dans mes programmes. À Arnaud, Magalie, Isabelle et surtout Guillaume, joueur de guitare professionnel.

J'ai aussi découvert durant ces dernières années quelque chose de magnifique, une proximité scientifique et humaine, dès mon stage de licence mais surtout ces trois dernières années dans le laboratoire de Didier Rognan et d'Esther Kellenberger. Pour les embêter, j'adore les appeler papa et maman et c'est tout à fait relevant, d'un point de vue professionnel bien sur : ils m'ont éduqué, m'ont soutenu et ont crû en moi. L'estime que je porte à leurs égards n'a d'égal que ma reconnaissance envers eux. Merci à vous deux.

Cette proximité, je la dois aussi à mes collègues, mes amis. Jamel Meslamani, pour avoir été là pendant 3 ans avec moi, ne change surtout pas. Mlle Anne-Marie Ray, ma chère et tendre Anne-Marie, merci pour ton oreille attentive, ton amitié et ta gentillesse. À Noé Sturm, mon petit protégé, la force est avec toi, ne l'oublie jamais. Ricky Bhajun, merci pour nos constants délires. À mon vieil ami Vladimir Chupakhin, qui a toujours veillé sur moi. À tous mes collègues, dont Anne-Laure Colin, Renaud Issenhart, Karima Azdimousa, Ewa Bielska, le statisticien, Francois Martz, Emilie Blaise, Ana, Marion Jenty, Anne-Florence, pour avoir vécu d'excellents moments avec vous et pour votre constante bonne humeur.

Enfin, je tiens à remercier tous les membres de mon jury, pour avoir accepter de juger mon travail et prendre le temps pour m'écouter.

Merci à vous lecteurs, lisant par curiosité ou par obligation, de porter intérêt à mon travail. Pensez aux forêts, évitez d'imprimer !

Sommaire

Résumé

L'évolution a amené un système biologique à s'adapter à son environnement, à subvenir à ses besoins. Les protéines sont une partie inhérente de cette évolution et, dans le cas des enzymes, elles possèdent toutes une fonction catalytique précise agissant sur un substrat particulier. Ce mode d'action se déroule alors en plusieurs étapes : le substrat s'approche et rentre dans la poche de la protéine, se lie de façon non covalente à celle-ci afin qu'en dernier lieu, la protéine puisse agir. D'autres molécules de faible poids moléculaire (généralement inférieur à 800 g/mol) peuvent se lier à la protéine cible afin d'altérer son fonctionnement en entrant en compétition avec le substrat : soit en ralentissant son fonctionnement, soit en améliorant ses performances. Cependant, ces processus impliquent la mise en jeu de nombreux paramètres, tant au niveau de la molécule que de la protéine cible, rendant ainsi une analyse exhaustive très difficile à réaliser.

Dans ces conditions, il est nécessaire de décomposer le problème en regardant chaque étape de façon indépendante :

- L'entrée de la molécule ou du substrat au sein de la protéine nécessite une poche possédant une géométrie adéquate. Parmi les multiples falaises, fosses, trous, poches, tunnels doit exister au moins une cavité apte à accueillir cette entité moléculaire.
- Une fois l'entité dans la cavité, la molécule doit interagir avec la protéine cible, d'une façon suffisamment favorable pour être compétitive par rapport au substrat endogène et au solvant.

Ces deux étapes sont la base de la reconnaissance moléculaire et la base de ce mémoire. L'objectif est ainsi d'améliorer notre compréhension de cette reconnaissance afin d'affiner la sélection de nouvelles entités moléculaires aptes à devenir des candidats médicaments.

Dans un premier chapitre, nous nous intéresserons tout d'abord aux spécificités structurales et physico-chimiques de ces cavités et des sites de liaison qu'elles forment. Nous discuterons ensuite des interactions non-covalentes qui peuvent être réalisées entre un site de protéine et des molécules de faible poids moléculaire. Les interactions principales ainsi que celles moins usuelles seront étudiées. Enfin, nous analyserons une partie des nombreux outils informatiques disponibles actuellement en s'orientant vers leurs représentations de la protéine et de ses interactions potentielles. On observera ainsi qu'il existe de nombreuses manières de représenter l'information, en utilisant une approche directe, c'est à dire tridimensionnelle et focalisée, ou indirecte, où la représentation sera conceptualisée à des fins de comparaison.

Le second chapitre propose l'élaboration d'une nouvelle méthode de description des cavités des sites de liaisons de protéine. En discrétisant les cavités en un ensemble de points, il nous est alors possible d'associer des propriétés physico-chimiques à ces points en fonction de l'environnement protéique local. En combinaison d'un outil d'alignement, cette stratégie nous permet de comparer les sites de liaisons de protéines et ainsi d'observer la similarité et la conservation, locale ou globale, au sein d'une famille de protéine ou alors de protéines totalement différentes.

Dans le troisième chapitre, une nouvelle représentation des modes d'interaction protéine/ligand a été mise en place. Contrairement aux méthodes existantes qui ne se focalisent que du côté protéine ou du côté ligand, cette méthode prend en considération l'information protéine/ligand à un niveau atomique. Associée à un algorithme de la théorie des graphes pour comparer et aligner ces modes d'interactions, nous avons pu mettre en évidence la corrélation entre deux similarités : celles des paires de modes d'interactions et des paires de sites de liaisons. Cependant, aucune corrélation n'a pu être trouvée entre la similarité de paires de ligands et les paires de modes d'interactions.

Cette conclusion nous a amené à l'hypothèse suivante : si les modes d'interactions sont indépendants de la structure de ligand, existe-t-il des fragments de molécules réalisant le même mode d'interaction ? Dans cette optique, le dernier chapitre se concentrera sur l'analyse locale des modes d'interactions de fragments de molécules actives dans la perspective de trouver des bioisostères, c'est à dire des fragments de molécules structuralement différents mais effectuant les mêmes interactions.

Summary

Evolution has led biological systems to adapt to its environment, to subsist to its needs. Proteins are an inherent part of this evolution and, in the case of enzymes, each of them has a precise catalytic function acting on a specific substrate. This mode of action takes place in several stages: the substrate get closer and goes into the protein pocket, binds to it in a non-covalent way so that, in the end, the protein can recognize it. Other molecules of low molecular weight (usually less than 800 g/mol) can bind to the target protein in order to alter its function by competing with the substrate, either by decreasing or improving its performance. However, theses processes require the involvement of many parameters, both in ligand and protein sides, thus making a comprehensive analysis very difficult to perform.

Under these conditions, it is necessary to split the problem by looking at each step independently:

- The input of the ligand or the substrate within the protein requires a cavity having the adequate geometry. Among the many cliffs, pits, holes, pockets or tunnels must be at least a cavity capable of receiving this molecular entity.
- Once the entity is within the cavity, it is necessary to interact with the target protein in a sufficiently favourable way to compete with the endogenous substrate and the solvent.

These two steps are the basis of molecular recognition and the basis of this thesis. The objective is to improve our understanding of recognition in order to refine the selection of new molecular entities that can become drug candidates.

In the first chapter, we will look at the structural and physico-chemical specificities of the binding sites and the cavities their forms. We will then discuss the non-covalent interactions that can be made between a protein binding site and low molecular weight molecules. Key interactions and unusual ones will be studied. Finally, we will analyse some of the many software and tools currently available by looking at their representations of the protein and its potential interactions. It will be underlined that it exists many ways to represent the information, either by using a direct approach, i.e. three-dimensional and focused, or indirectly, where the representation is conceptualized for comparison purposes.

The second chapter will focus on the development of a new method to describe binding site cavities. By discretising cavities by a set of points, we can assign physico-chemical properties to these points, and therefore encode local protein environment. Combined to an alignment tool, this strategy allows us to compare protein binding sites and thus to observe the similarity and the conservation, local or global, within a family or between unrelated proteins.

In a third chapter, a new representation of protein/ligand interaction patterns has been developed. Unlike existing methods that primarily focus on either the protein or the ligand site, this method takes into account the protein/ligand information at an atomic level. Associated with a graph theory algorithm to compare and align these interaction patterns, we were able to show that the similarity of interaction patterns correlates with the similarity of binding sites but not with the structural similarity of bound ligands.

This finding led us to the following hypothesis: if the interaction modes are independent of the ligand structure, are there fragments of molecules carrying the same binding mode? In this context, the last chapter will focus on the analysis of local interaction patterns of binding molecules with the objective of finding bioisosteres, i.e. structurally different fragments sharing the same binding mode.

Glossaire

Cavité	Espace formé par le site de liaison où peut s'introduire solvant, ions ou d'autres molécules telles qu'un ligand.
Cible	Macromolécule biologique, généralement protéique
CSD	Cambridge Structural Database Banque de dépôts de structures cristallographiques de composés organiques
Empreinte	Chaine de caractères caractérisant l'information structurale, physico-chimique ou d'interaction, d'une entité chimique ou macromoléculaire. Chaque caractère peut être représenté soit par un nombre binaire, un entier ou une valeur flottante.
Ligand	Molécule de faible poids moléculaire capable de se lier à une cible
PDB	Protein Data Bank Banque de dépôt de structures biologiques cristallographiques
Poche	Cas particulier d'une cavité, très ouverte et peu enfouie, facilement accessible par une molécule sans mouvement important de la part de la protéine.
Pose	Conformation et position prédite d'une molécule arrimée dans un site de liaison
sc-PDB	Screening-Protein Data Bank
Site de liaison	Ensemble des acides aminés, métaux, ions ou autres résidus d'une protéine capable d'interagir avec un ligand.
Sonde	Point représentatif d'une possible interaction non-covalente
SVM	Support Vector Machine ou Séparateur à Vaste Marge. Méthodologie permettant de trouver une fonction permettant de séparer deux ensembles distincts

Chapitre 1
Les sites de liaisons : interactions non covalentes et représentations

Le développement de nouveaux médicaments est un processus complexe dont le premier objectif est d'obtenir une molécule affine pour une cible. Dans la conception rationnelle d'un principe actif, la reconnaissance moléculaire est un critère essentiel. Il est communément admis que cette reconnaissance implique une complémentarité de formes et de propriétés physico-chimiques entre le ligand et le site de liaison. L'un des buts des chimistes médicinaux est donc de maximiser cette complémentarité afin d'obtenir des ligands affins et sélectifs. Malgré tout, cette condition est nécessaire mais pas suffisante. En effet, il est important de prendre en considération de nombreux paramètres liés à l'administration, à la distribution, à la métabolisation, à l'élimination et à la toxicité.

Assurer l'innocuité d'une molécule est extrêmement difficile, si ce n'est impossible dans l'état actuel de la recherche scientifique. Bien que de nombreuses contraintes et règles ont été mises en place dans le but de modifier le mode d'administration ou de minimiser le risque d'effets secondaires et le coût, il est important de garder à l'esprit qu'un médicament est toujours un compromis entre son bénéfice thérapeutique et ses effets non désirés. Ces derniers peuvent être issus de causes externes et variées, et de nombreux facteurs sont encore incompris tels que l'impact de environnement, de l'alimentation ou de l'interaction avec d'autres médicaments.[1] De plus, certains effets indésirables peuvent être occasionnés par le manque de sélectivité de la molécule qui se lie à des cibles possédant des fonctions différentes. A cela se rajoute le contexte budgétaire restreint, impliquant ainsi que les molécules doivent être facilement synthétisables et éliminant de ce fait celles possédant un grand poids moléculaire ou de nombreux centres chiraux.

D'autre part, le mode d'administration et la localisation de la cible ont une grande importance sur la stratégie à adopter pour trouver un candidat médicament. Après analyse physico-chimique de près de 90% des composés actifs disponibles par voie orale et ayant passé la deuxième phase clinique, Lipinski a montré en 2001 que ces molécules possèdent des caractéristiques relativement spécifiques : un poids moléculaire inférieur à 500 g/mol, un coefficient de partage octanol/eau inférieur à 5, un nombre de donneurs de liaison hydrogène inférieur à 5 et d'accepteurs de liaison hydrogène inférieur à 10.[2,3] Dans le cas du passage d'un médicament à travers la barrière hémato-encéphalique, le poids moléculaire de la molécule doit être de préférence inférieur à 400 g/mol, strictement inférieur à 600 g/mol et le nombre

maximum de donneurs et d'accepteurs de liaison hydrogène inférieur à 8.[4] De ce fait, obtenir la molécule optimale nécessite une véritable adaptation par rapport à sa cible. La difficulté est d'autant plus grande qu'il faut trouver parmi des milliards de molécules potentielles celle qui apportera la meilleure balance entre efficacité et risque. Toutefois, cela peut être simplifié par les méthodes informatiques.

Tester expérimentalement des milliers, voir des millions de composés engendre un coût, tant humain que financier ou écologique et malheureusement, nombreux sont ceux qui, une fois sélectionnés, échouent en phase clinique.[5] Il est donc d'une importance capitale d'orienter la recherche en pré-sélectionnant avec précision les molécules. Dans cette optique, la résolution atomique de structures cristallines de complexes protéine/ligand a apporté, depuis le début des années 1970, une information inestimable qui, couplée à l'essor de l'informatique, a ouvert la perspective de tester *in-silico* si une molécule peut se lier à une cible. Grâce à cela, le développement de banques de dépôts de structures ont vu le jour au cours de ces dernières décennies : la Protein Data Bank (PDB)[6] contient ainsi plus de 90000 structures cristallographiques, dont près de 80000 sont des complexes de protéine avec ou sans ligand(s). La screening-PDB (sc-PDB),[7] sous-ensemble de la PDB contient près 8000 complexes protéine/ligand dont les sites de liaison sont considérés comme drug-like, c'est-à-dire aptes à accueillir un substrat. Cependant, avec l'apport de cette information nous est parvenu la complexité inhérente des structures tridimensionnelles, qu'il nous faut traiter, analyser et très souvent simplifier.

Une protéine est une macromolécule complexe avec des caractéristiques intrinsèques : son repliement tridimensionnel, sa flexibilité, les propriétés physico-chimiques de ses acides aminés, sa fonction. Dans ces conditions, il est nécessaire de condenser ces informations suivant nos besoins. Dans le cadre de la conception de nouveaux médicaments, seul le site de liaison de la protéine est généralement considéré. Comme précédemment énoncé, la complémentarité de forme et de propriétés physico-chimiques entre une protéine et un ligand est un concept simple mais essentiel. L'ensemble forme un environnement favorable, impliquant des contacts tels que des liaisons hydrogène, des contacts apolaires, des interactions aromatiques, ioniques et bien d'autres. Puisque la majorité des médicaments actuellement mis sur le marché interagissent de façon non-covalente, l'analyse et la représentation des sites de liaisons,

tant au niveau de leurs formes que des interactions qu'ils peuvent effectuer avec leurs ligands ouvrent la perspective d'une meilleure compréhension des interactions protéine/ligand et par conséquent, nous orientent vers une meilleure sélection de molécules potentiellement actives. Afin d'être appliqué *in-silico*, ces nombreuses définitions doivent être codées d'une façon simple, rapide et efficace.

Ce chapitre commencera par décrire l'architecture générale d'un site de liaison, tant au niveau de sa configuration que de sa constitution, permettant ainsi de définir un socle dans le développement d'algorithmes. En second lieu, nous nous concentrerons sur les possibilités qu'offrent ce site pour arrimer une molécule de faible poids moléculaire, en regardant les interactions non covalentes, fortes comme faibles. Enfin, nous nous attarderons à observer les modes de représentations développés au cours de ces dernières décennies, en analysant les forces et les faiblesses de chacun. Les domaines de recherche observés au cours de ce chapitre sont très larges et divers et ne peuvent donc pas être exhaustifs. Cependant, une forte considération a été portée pour englober la majorité des concepts.

Afin de clairement définir le contexte de ce chapitre (**Figure 1**), on considèrera un ligand (en orange) comme une molécule de faible poids moléculaire interagissant de façon non covalente avec le site de liaison de la protéine (en vert et cyan). La surface de ce site (en gris) délimite les zones accessibles au ligand, c'est à dire la cavité du site (en orange), de la protéine.

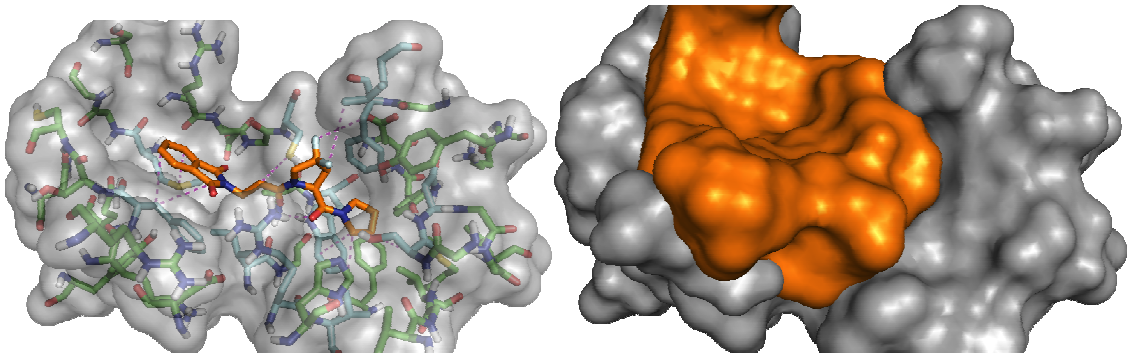


Figure 1 - Prolyle oligopeptidase en complexe avec l'inhibiteur R-Pro-(decarboxy-Pro)-Type (PDB :3EQ7). A gauche : Le ligand est colorié en orange. Les interactions de l'inhibiteur avec la protéine sont dessinées en violet (toutes ne sont pas représentées). Les résidus en interaction sont coloriés en cyan, tandis que le site de liaison contient l'ensemble des résidus. A droite : représentation de la cavité du site de liaison (en orange) et de la surface du site de liaison en gris.

1. PARTICULARITE DES SITES DE LIAISONS

L'obtention de structures cristallines de protéines, liées ou non à un ligand, a ouvert la perspective d'une analyse tridimensionnelle et atomique de phénomènes biologiques. Malgré le fait que ces structures soient statiques, elles nous apportent de nombreuses informations sur les relations protéines/ligand, leurs modes d'interaction, les caractéristiques des sites de liaison et nous montrent l'étendue de leur complexité.

1.1 ANATOMIE

L'anatomie d'une cavité, telle que sa forme, sa polarité, son accessibilité au solvant, influe directement sur les propriétés physico-chimiques du ligand à développer. Puisque les protéines sont capables d'interagir avec une grande variété de molécules, les cavités sont aussi très différentes. Un exemple typique proposé par Nisius[8] concerne deux protéines : une ribonucléase et une endonucléase, présentées sur la **Figure 2A** et la **Figure 2B**, respectivement. Toutes deux clivent des acides nucléiques, l'un un ARN et l'autre un ADN, mais leurs cavités sont cependant très différentes. Ainsi la cavité principale de l'endonucléase est sphérique et enfouie tandis que celle de la ribonucléase est allongée, grande et exposée au solvant. Cet exemple nous montre bien que même deux protéines aux fonctions similaires mais aux substrats différents peuvent avoir des sites de liaison structuralement opposés et par conséquent accueillir des ligands différents.

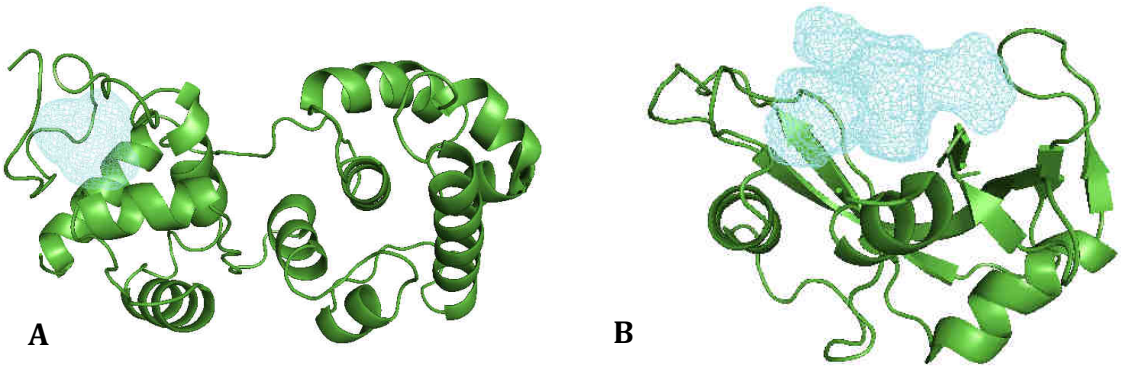


Figure 2 - Cavité (en cyan) de deux enzymes (en vert). A : Cas de l'endonucléase (Code PDB : 2ABK), où la cavité est petite et enfouie. B : Cas de la ribonucléase, (Code PDB : 1ROB) où la cavité et large et accessible.

La caractéristique primaire pour distinguer un site de liaison d'une simple concavité formée par un ensemble d'acides aminés réside dans le volume de la cavité générée par celui-ci. Il est facile de conclure dans le cas des enzymes puisque la cavité la plus volumineuse correspond généralement au site de liaison.[9] Faute de réel standard sur les définitions et les limites d'une cavité, le volume moyen reste très différent suivant les publications : 930 Å^3 pour Nayal[10] et 610 Å^3 pour An.[11] En dépit de ces différences, certaines études ont recherché des corrélations existantes avec le volume : Liang a observé en 1998, une relation entre le volume de la protéine et le nombre de cavités/poches mais pas avec le volume de ces dernières.[12] Ainsi, une augmentation du volume de la protéine de 1000 Å^3 entrainerait la création d'une nouvelle cavité ou poche. De plus, le volume de la cavité est généralement trois fois plus grand que celui du ligand.[10,13] Les caractéristiques géométriques ne sont pas les seuls critères à prendre en compte lors de l'analyse de sites de liaison. Les propriétés physico-chimiques de ces derniers sont tout aussi importantes pour trouver des ligands affins.

En 1994, Young et Covell ont observé la lipophilie des sites et ont attribué à chaque site une valeur d'hydrophobie en fonction des propriétés de ses résidus.[14] Ils conclurent que dans 25 des 38 cas étudiés, la localisation du ligand peut être correctement prédite en ne cherchant que le site le plus hydrophobe. Pérot confirme en 2010 cette proposition à travers l'analyse de 56 complexes protéine/ligand de haute résolution, montrant ainsi que la proportion d'atomes O, N, S par rapport à O, N, S et C n'est que d'environ 35%.[15] La forme et les caractéristiques physico-chimiques suivent par conséquent une certaine tendance et permettent de sélectionner le site de liaison le plus à même d'interagir avec un ligand.

En 2005, près de 60% des projets pharmaceutiques échouaient à cause d'un site non-droguable.[5] Bien que ce pourcentage ait changé au cours des années, la vaste combinatoire existante au sein des sites de liaison a amené la nécessité de présélectionner les plus propices à accueillir une petite molécule. Lorsque la recherche a permis d'obtenir pour une cible quelques molécules ayant passé au moins une phase clinique ou une mise sur le marché, il est relativement facile d'en conclure quant à « *l'accessibilité* » du site de liaison pour un médicament. Cependant, cela reste

occasionnel et, pour les autres cas, il est important de pouvoir distinguer divers sites de liaison afin de sélectionner celui qui pourrait lier des ligands avec une haute affinité.

De nombreuses méthodes de prédiction de ce concept, généralement appelé *droguabilité* ou *ligandabilité*, ont été développées au cours de ces dernières années. Celles-ci se basent en général sur des propriétés structurales et physico-chimiques des sites de liaison : la surface totale du site, le nombre de contacts polaires et apolaires et les principaux moments d'inerties sont les caractéristiques importantes trouvées par Hajduk en 2005 après une régression linéaire sur 28 sites de liaison.[16] Nayal et Honig confirment les résultats précédant à travers l'analyse de 99 complexes protéine/ligand. [10] Sur les 408 descripteurs initiaux, seuls 18 sont considérés comme importants, incluant en plus de ceux précédemment mentionnés, le nombre de résidus, d'atomes, l'enfouissement moyen, l'ouverture de la poche, ainsi que la proportion de proline et l'hydrophobie. L'influence de ce dernier critère est fortement remarquée par divers auteurs et est conforté par le pourcentage d'hydrophobie dans les sites de liaison (65%). Schmidtke considère cependant que la polarité des sites ne doit pas être négligée ni considérée comme totalement négative.[17]

Inférer une valeur de droguabilité à un site de liaison est une tâche difficile car la prédiction est très dépendante du jeu de données initial. En effet, il est nécessaire de posséder un large éventail de structures de protéines différentes pour lesquelles des projets de chimie médicinale ont échoué. Rares sont cependant les entreprises ou même les universités qui publient ou mettent sur le domaine public ce qu'ils considèrent comme des échecs. Ainsi, de nombreuses cibles sont connues pour être droguables mais trop peu sont celles considérées comme non droguables. De plus, le seuil de séparation des deux classes est souvent subjectif : Cheng considère la thrombine comme difficile,[18] tandis que Halgren la considère comme droguable.[19]

Comme annoncé au début de cette partie, les structures cristallographiques sont caractérisées par des coordonnées statiques. Cependant, un complexe protéine/ligand est une entité dotée de mouvements, locaux ou globaux, qu'il est nécessaire de prendre en considération.

1.2 FLEXIBILITE

Une protéine n'est pas un ensemble rigide et de ce fait, le mouvement des chaînes latérales des acides aminés, de boucles ou de structures secondaires influence la structure du site de liaison. Malheureusement, les structures tridimensionnelles ne donnent pas accès à toute l'étendue de la flexibilité des protéines.

Un ensemble de conformations générées par l'utilisation de la dynamique moléculaire permet de pallier à ce problème, mais augmente aussi la complexité et le temps d'analyse. L'un des premiers inhibiteurs de l'HIV-intégrase, le raltegravir,[20] a été découvert grâce à une dynamique de 2 ns montrant l'existence d'une petite cavité adjacente, difficilement visible à partir de la structure initiale à cause d'une boucle désordonnée.[21] En effet, on peut observer dans la **Figure 3A** que cette boucle, composée des résidus 141 à 144, n'est pas résolue : il n'existe aucun acide aminé liant l'hélice α en bas à droite et le début de la boucle en bas au centre. De plus, le site de liaison n'est pas enfoui, ce qui implique une certaine flexibilité conformationnelle. Dans ces conditions, l'utilisation de dynamique moléculaire pour observer les différentes conformations et prédire l'orientation de la boucle permet de compenser les problèmes liés à la résolution des structures de protéine. Le raltégravir, présenté dans la **Figure 3B**, profite ainsi de la formation de la cavité créée par la boucle précédemment manquante.

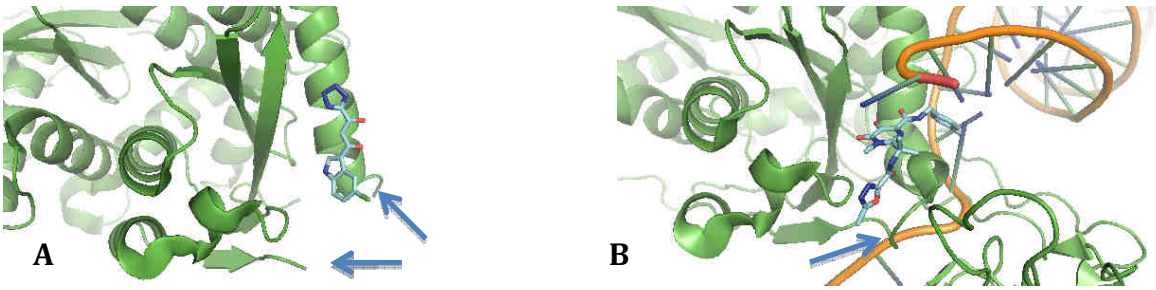


Figure 3 - A : HIV-Integrase (en vert) en complexe avec 1-(5-chloroindol-3-yl)-3-hydroxy-3-(2h-tetrazol- 5-yl)-propenone (en cyan) ; Code PDB : 1QS4. **B** : *Prototype Foamy Virus*-Integrase (en vert) en complexe avec le raltégravir (en cyan) et un acide nucléique (en orange). Code PDB : 3OYA. Les flèches bleue correspondent à la boucle flexible et non résolue dans la figure A et visible dans la figure B.

Luque et Freire[22] ont analysé les constantes de stabilité des résidus de 16 protéines de familles différentes et sont parvenus à la conclusion que la distribution de l'énergie de stabilisation n'est pas arbitraire au sein d'un site de liaison.[23] Ils observent ainsi que les parties protéiques en bordure du site sont généralement moins stables, permettant au ligand d'accéder à la cavité, d'interagir avec le site et de maximiser les contacts. D'un autre coté, les résidus assurant la fonction de la protéine ont une énergie de stabilisation beaucoup plus élevée que les autres résidus de la protéine. Ces résultats nous laissent entendre qu'il existe une forte stabilité globale des sites de liaisons, même si les chaînes latérales des résidus restent mobiles.

Les protéines présentes dans la PDB sont généralement en complexe avec une molécule de faible poids moléculaire ou avec son ligand endogène. Par conséquent, on peut librement supposer que la conformation adoptée par la protéine possède une énergie potentielle faible et qui maximise les contacts avec le ligand. De ce fait, l'utilisation de structure résolue aux rayons X pour établir des hypothèses pour la conception de médicaments reste une approximation raisonnable.

La forme du site de liaison est une condition primaire d'une liaison avec une molécule. Comme nous avons pu l'observer, la forme de la cavité créée par le site est très variable, bien plus volumineuse que le ligand qu'il accueille. Cependant, elle constitue un socle solide pour la recherche de molécules actives. L'hypothèse selon laquelle l'utilisation de structures cristallographiques est suffisante pour trouver de nouveaux médicaments est valide en raison d'une forte énergie de stabilisation au sein du site de liaison. Un ligand doit ainsi être moins volumineux que la cavité pour pouvoir rentrer dans celle-ci et être relativement hydrophobe pour s'accommoder de ses caractéristiques physico-chimiques. Cependant, pour rester au sein de la cavité, il lui est nécessaire d'interagir par le biais d'interactions non-covalentes avec le site de liaison.

2. INTERACTIONS NON-COVALENTES

Dans un état initial, la protéine et le ligand sont dissociés et entourés par le solvant. Lors de l'étape d'association, chacun des protagonistes doit être désolvaté avant de pouvoir interagir l'un avec l'autre. La complémentarité de forme est importante car elle permet de laisser entrer le ligand mais elle ne suffit pas pour qu'il reste au sein de la protéine. Pour cela, des interactions non-covalentes doivent être créées. La balance des interactions électrostatiques, de van der Waals, de liaison hydrogène, ou hydrophobe doit tendre vers une compensation énergétique favorable par rapport au coût de la désolvatation. L'énergie globale d'une complexation protéine/ligand est définie par l'énergie libre ΔG, somme de l'enthalpie ΔH et du terme entropique $-T\Delta S$. Une compétition existe donc entre les interactions non-covalentes associées à l'enthalpie et les effets de désolvatation et de réorganisation du solvant associés à la composante entropique. Afin de maximiser le gain enthalpique, il est donc nécessaire de maximiser les interactions entre la protéine et le ligand. L'énergie d'interaction peut alors être estimée par diverses méthodes, tel qu'un champ de force, un score empirique ou une fonction statistique. Bien qu'elles soient méthodologiquement différentes, elles se basent toutes sur des données expérimentales, dont les erreurs tendent à être additives.

De nombreux types d'interactions non-covalentes ont été mis en évidence dans les complexes protéine/ligand de la PDB : liaison hydrogène, ionique, aromatique, métal/accepteur, halogène et π-cation. Chacune possède des caractéristiques précises, voire directionnelles mais surtout dépendantes du contexte environnemental direct des atomes de protéine et du ligand impliqués dans l'interaction.

2.1 INTERACTIONS CLES

2.1.1 LIAISON HYDROGENE

La liaison hydrogène met en jeu un atome d'hydrogène, lié covalemment à un atome donneur de liaison hydrogène, que l'on nommera donneurH et un atome accepteur de liaison hydrogène, que l'on appellera accepteurH. Le donneurH est un atome plus électronégatif que l'hydrogène, tel qu'un oxygène, un azote ou dans une moindre mesure un carbone, tandis que l'accepteurH est un hétéroatome polarisable. La **Figure 4A** montre un exemple de liaison Hydrogène.

Les liaisons hydrogène sont importantes pour le repliement protéique, les changements conformationnels et la reconnaissance protéine/ligand. Côté ligand, elles affectent les propriétés physico-chimiques des molécules, telles que la solubilité et la perméabilité membranaire qui sont des éléments cruciaux dans le développement d'un médicament. Bien que la résolution des structures protéine/ligand par la diffraction des rayons X ne permette d'obtenir la position des hydrogènes qu'à très haute résolution, le nombre important de ces structures nous autorise une évaluation statistique de qualité de la topologie entre l'accepteur et le donneur de liaison hydrogène.

La géométrie d'une telle interaction répond à des critères précis mais dépend des atomes impliqués et de leurs environnements directs. L'utilisation, par exemple, de la CSD[24] dans le cas de molécules organiques fournit un vaste échantillon à analyser. L'équipe de Verdonk a rassemblé l'ensemble des structures issues de la CSD, de la PDB et de calculs d'énergie théorique pour former IsoStar,[25] une base de données contenant l'analyse des interactions non-covalentes de près de 250 groupements chimiques. Ils ont observé que l'orientation des hydrogènes ne suit pas toujours la paire d'électrons libres de l'accepteurH et varie grandement en fonction de la nature de celui-ci et du donneurH.

Pour mieux comprendre ce phénomène, Taylor et Kennard[26] ont regardé sur plus de 1000 échantillons, l'effet des charges de l'accepteurH et/ou du donneurH sur la distance entre les deux protagonistes, dans le cadre de la liaison N-H --- O=C. Leurs résultats dénotent une diminution de la distance lorsque l'accepteurH ou le donneurH est chargé par rapport à une forme neutre (**Tableau 1)**. Lorsque les deux atomes sont chargés, la liaison hydrogène est couplée à une interaction ionique, diminuant encore un peu plus la distance entre les deux atomes lourds.[27]

		AccepteurH		
	Distance (Å)	RC(=**O**)OH	R_2C=**O**	RC(=**O**)O^-
DonneurH	>N-H	2,002 ± 0,012	1,970 ± 0,022	1,928 ± 0,019
	NH_4^+	1,916 ± 0,041	1,995 ± 0,110	1,886 ± 0,018
	$R_2NH_2^+$	1,887 ± 0,014	1,966 ± 0,178	1,796 ± 0,014

Tableau 1 - Distance moyenne, en Angstroems, entre l'hydrogène et l'AccepteurH pour 3 donneurH et 3 accepteurH. L'oxygène accepteurH est mis en gras. Données issues de [26]

La distance n'est pas la seule composante géométrique à être influencée par la composition atomique : l'angle θ réalisé entre le donneurH, l'hydrogène et l'accepteurH est aussi impacté. Une liaison forte est caractérisée par un angle θ de 180°, cependant, plus la distance augmente, plus celui-ci tend vers 90°.[28] Murray-Rust et Glusker[29] confirme l'effet de l'environnement dans le cadre d'une étude de la liaison hydrogène formée par les oxygènes dans la CSD.

D'un point de vue énergétique, la force des liaisons hydrogène dépend ici encore des atomes impliquées. Le **Tableau 2** montre quelques valeurs d'énergie de dissociation pour des liaisons covalentes et des liaisons hydrogène. Par rapport à une liaison covalente tel que C-C ou C-N, les liaisons hydrogène sont énergétiquement beaucoup plus faibles puisque variant entre 10 et 40 kJ/mol, permettant ainsi d'être créées et rompues plus facilement qu'une liaison covalente. Le fluorure d'hydrogène et l'ion fluorure réalisent une liaison hydrogène plus stable énergétiquement qu'une liaison covalente N-N.

Liaison	***C-C***	***C-N***	**F-H--F^-**	***N-N***	**HO-H--OH^-**
Energie de dissociation (kJ/mol)	348	308	212	170	20

Tableau 2 - Energie de dissociation de diverses liaisons covalentes (en italique) ou non covalentes

Dans certains cas, le fluor peut aussi être un accepteurH et ainsi créer une liaison hydrogène avec un hydrogène lié à un donneurH. Carosati[30] a recherché de telles interactions afin d'en inférer la géométrie. Sur 105 contacts trouvés parmi 23000 complexes protéine/ligand, la distance d_{F-H} est en moyenne de 2,3 Å. Cependant, cette valeur varie en fonction des atomes avoisinants : lorsque le fluor est lié à un aliphatique, d_{F-H} est de 2,1 Å tandis qu'avec un aromatique, la distance tend vers 2,7 Å. L'angle θ_4 est compris entre 120 et 180° alors que l'angle θ_3 est compris entre 90° et 180°, montrant par conséquent une faible directionnalité de ce type d'interaction. La **Figure 4B** montre un exemple de liaison hydrogène entre un azote donneurH et le fluor accepteurH.

Dunitz et Taylor se sont concentrés sur l'analyse des interactions du fluor.[31] Sur les 5947 liaisons C-F trouvés au sein de la CSD, seul 37 réalisent des liaisons hydrogène, dont certaines sont issues de complexes organométalliques. Ils avancent alors une trop grande différence d'électronégativité entre le fluor et les donneurs de liaison hydrogène.

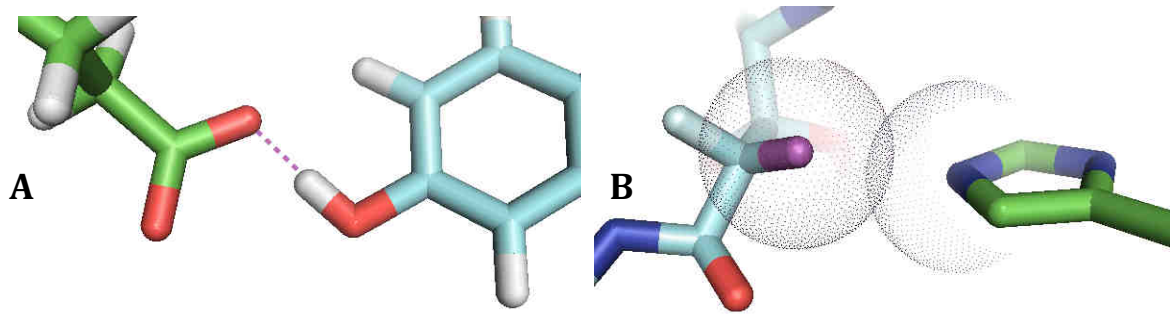


Figure 4 - Exemples de liaison hydrogène. **A** : 4-Hydroxytamoxifen (en cyan) en complexe avec le récepteur d'œstrogène alpha (PDB : 3ERT). Liaison hydrogène entre l'oxygène du groupement hydroxyle et l'oxygène du carboxylate du glutamate 47 (en vert). **B** : FPA (en cyan) en complexe avec l'élastase pancréatique porcine (en vert). Liaison hydrogène entre le fluor (en violet) de FPA et l'histidine 57.

2.1.2 INTERACTION IONIQUE

Les interactions ioniques résultent d'une force d'attraction électrostatique entre deux atomes de charges opposées, tels qu'un groupement guanidinium et un carboxylate (**Figure 5**). Contrairement à une liaison hydrogène, ce type d'interaction n'est pas directionnel et ne nécessite pas de précision supplémentaire quant à la géométrie. La nature et la force de ces interactions sont ici aussi très dépendantes du contexte atomique des protagonistes et de leur distance. Dans le cadre d'une interaction entre deux charges, l'énergie associée à cette interaction est inversement proportionnelle à la distance qui les sépare. Par conséquent, l'énergie d'interaction décroit lentement lorsque la distance augmente, impliquant une composante énergétique favorable même à longue distance et par conséquent une reconnaissance protéine/ligand renforcée dans des gammes de distance plus lointaines que d'autres interactions non covalentes. La fonction représentant ce lien est définie par la loi de Coulomb:

$$\frac{q_1 q_2}{D r}$$

où q_1 et q_2 représentent les charges des deux atomes en interaction, D la constante diélectrique et r la distance entre les deux atomes. Les interactions ioniques sont des interactions fortes pouvant aller de 40 à 500 kJ/mol.

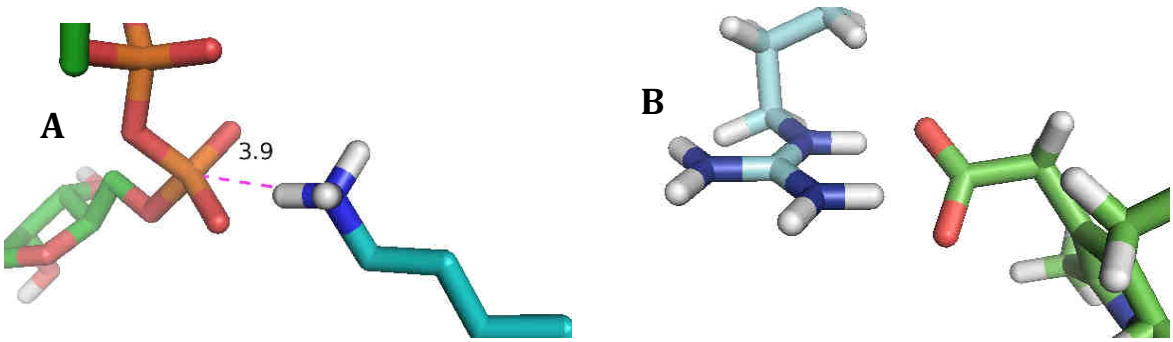


Figure 5 - Exemple d'interaction ionique dans un complexe protéine (en cyan) ligand (en vert). A) Cas d'une interaction entre un groupement phosphate et une amine quaternaire. (code PDB : 2y27) B) Cas d'une interaction entre un groupement guanidinium et un acide carboxylique (code PDB : 3EQ1).

2.1.3 COORDINATION DE METAUX

Les métaux sont souvent présents dans les sites de liaison des protéines et peuvent jouer un rôle crucial dans le processus catalytique. Dans le cas de la version 2012 de la sc-PDB ne contenant que des sites considérés comme droguables, 16% contiennent au moins l'un des métaux suivant : Zinc, Fer, Magnésium, Manganèse, Cobalt et Calcium. Ces métaux sont enfouis dans le site de liaison mais restent suffisamment accessibles pour interagir avec des ligands.

Une étude de la coordination métal/ligand a été réalisée par Harding en 2000.[32] Elle a observé tous les atomes autres que le carbone, le phosphore ou l'hydrogène à moins de 3,6 Å d'un atome de métal de complexes protéine/ligand tirés de la PDB version 1999. Les distances moyennes entre un métal et des atomes de protéine oscillent entre 1,88 et 2,36 Å. La gamme de distance est globalement en accord avec les données tirées de la CSD : les déviations sont en général autour de 0,1 Å pour des entrées possédant une résolution inférieure à 2,2 Å et de 0,2-0,3 Å pour des résolutions supérieures. Seule la distance Zn-$O_{carboxylate}$ ne respecte pas ces normes, en raison de la possibilité du carboxylate de se coordiner avec soit un de ses oxygènes soit les deux, rendant l'analyse plus difficile.

Les métaux peuvent coordiner un nombre variable d'atomes de protéine ou de ligand. L'atome de zinc par exemple se coordine avec un nombre d'atomes assez variés, allant de moins de 4 jusqu'à 6 atomes, tandis que l'atome de manganèse forme essentiellement des complexes hexadentates (**Figure 6**). Ce nombre variable de coordination dépend alors de l'encombrement autour du métal et du rayon de van der Waals de ce dernier.

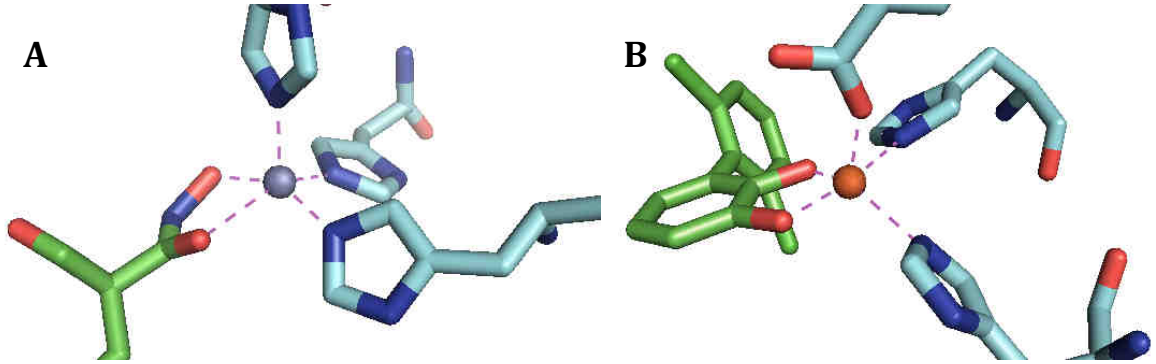


Figure 6 - Exemple de coordination entre un métal (sphère centrale), la protéine (en cyan) et le ligand (en vert). **A :** MMP8 en complexe avec un inhibiteur à base d'asparagine et d'acide malonique. L'atome de zinc réalise 5 contacts dont 3 avec des atomes d'azote d'histidines et 2 atomes d'oxygène du ligand. **B :** 2',6'-dichloro-biphenyl-2,6-diol en complexe avec le biphenyl-2,3-diol 1,2-dioxygenase. L'atome de fer réalise 5 contacts dont 2 avec le ligand, 2 avec les atomes d'azotes des histidines et un avec l'atome oxygène de l'acide carboxylique de l'acide glutamique.

2.1.4 INTERACTION AROMATIQUE

Les interactions aromatiques sont des exemples typiques d'interactions entre des groupements hydrophobes. Bien que nous discuterons plus tard de l'existence d'interactions entre des cycles aromatiques et des donneurs de liaison hydrogène ou des cations, celles-ci restent plus occasionnelles que les interactions entre deux aromatiques. Il existe différents types de positionnement entre deux cycles aromatiques, présentés dans la **Figure 7** : la forme en T où un cycle se positionne perpendiculairement au plan et dans l'axe du centre du second ; la position parallèle où les deux cycles sont coplanaires; le parallèle déplacé, équivalent au parallèle mais avec un décalage de l'un des deux cycles ; le chevron où les plans des deux cycles ne sont ni parallèles ni perpendiculaires. L'orientation choisie par deux cycles aromatiques est dépendante du contexte structural du ligand comme de la protéine, impliquant un effet conformationnel des deux parties et de leurs substituants.

Les cycles aromatiques sont plans avec des électrons π, qui lors de la superposition, maximise les contacts de van der Waals. Imai et Yamamoto[33] ont analysé plus de 25000 structures de complexes protéine/ligand à la recherche d'interactions entre résidus aromatiques. Ils ont ainsi montré que les histidines s'agglomeraient préférentiellement selon la forme parallèle, bien que la forme en T et Chevron soit aussi présente. Dans le cas de la phénylalanine et du tryptophane, tous les types d'empilements existent, mais la forme en T reste la plus fréquente. Enfin, dans le cas de la tyrosine, l'influence de l'hydroxyle entraine de façon majoritaire l'adoption de la forme en T.

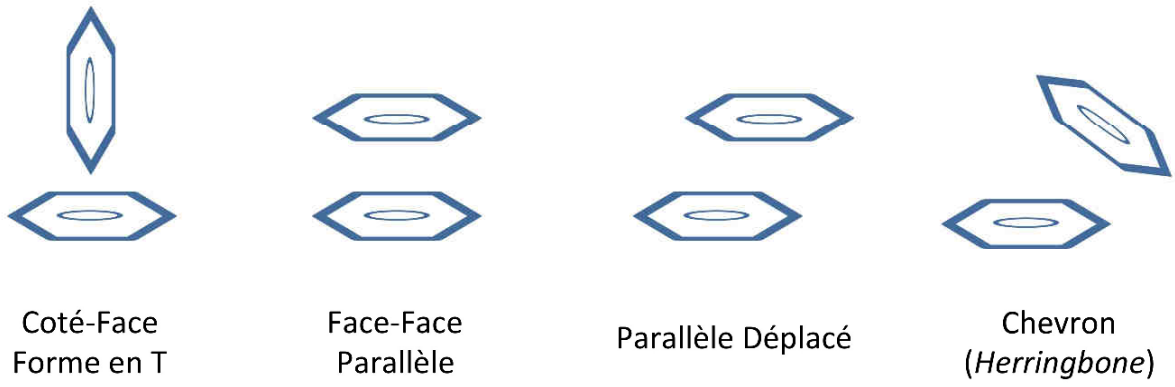


Figure 7 - Liste des superpositions existantes entre deux cycles aromatiques

Certaines publications ne différentient pas la superposition parallèle du parallèle déplacé puisque seule la distance entre le centre des cycles est généralement considérée et non le déplacement entre ces derniers. Cela implique cependant une grande différence puisque le premier est énergétiquement défavorable tandis que le second est favorable. La **Figure 8** montre les interactions électrostatiques en fonction de la distance et de l'angle entre les cycles aromatiques. On peut ainsi clairement voir que la forme parallèle est une forme répulsive contrairement à la forme en T ou parallèle déplacée. Il est par conséquent difficile de conclure quant à l'orientation préférentielle des cycles aromatiques des différents acides aminés. De nombreuses revues ont traité ce sujet[34,35] et s'accordent à dire qu'en milieu aqueux, l'empilement parallèle déplacé et coté-face sont les plus fréquents, la forme en T étant la forme la plus énergétiquement favorable.

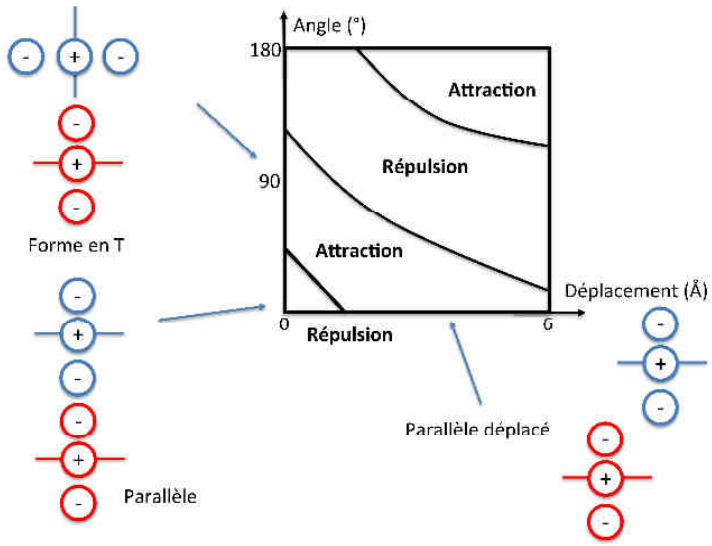


Figure 8 - Interaction électrostatiques pour la distribution des électrons π par rapport à l'angle (°) et la distance (Å) entre les cycles aromatiques. L'image est tirée de [34]

2.2 AUTRES INTERACTIONS NON-COVALENTES

De nombreux autres types d'interactions non-covalentes ont émergé au cours de ces dernières années grâce l'augmentation exponentielle de structures cristallographiques et d'études de relations structure/activité. Bien qu'elles aient pu être omises par un manque d'information ou une faible contribution énergétique, elles ne doivent cependant pas être ignorées.

2.2.1 INTERACTION Π/CATION

Ce type d'interaction implique un cycle aromatique et un atome chargé positivement. Prenons le cas du groupement guanidinium de l'arginine : sur 27000 interactions non-covalentes réalisées par l'arginine, près de 5% concernent des interactions π-cation.[33] La **Figure 9** présente 2 configurations adoptées par cet acide aminé : Parallèle déplacée (A) et Chevron (B).

L'amine d'une lysine, protonée à pH physiologique, a aussi tendance à réaliser ce type d'interaction : près de 15% des contacts du résidu sont des interactions π-cation. Des calculs théoriques sur un jeu de données de 68 complexes a amené Gallivan et Dougherty à la conclusion que ce sont des interactions fortes, en phase gazeuse comme aqueuse.[36] Bien que restreinte aux interactions intra-moléculaires, ils ont montré que l'arginine est plus propice que la lysine pour réaliser ce type interaction et tend à opter préférentiellement pour un arrangement parallèle. L'argument proposé est celui d'une meilleure solvatation de la lysine par rapport à l'arginine permettant ainsi à ce dernier de réaliser de meilleurs contacts de Van der Waals. L'ensemble des interactions π-cation représente cependant moins d'1 % des contacts protéine/ligand existants parmi 6091 structures cristallographiques de la PDB.[33]

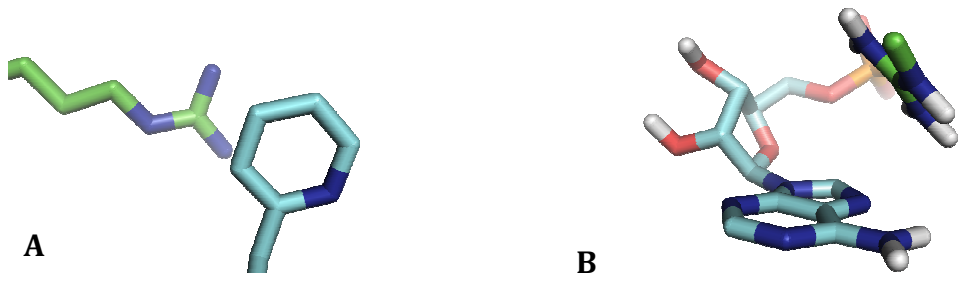


Figure 9 - Cas d'interactions non-covalentes de type π-cation entre le groupement guanidinium de l'arginine et un cycle aromatique. A : HIV-1 protéase en complexe avec l'inhibiteur MSA367 (PDB : 1EC3) B : Adenylate kinase (en vert) en complexe avec l'AMP (en cyan). (PDB : 1ANK)

2.2.2 INTERACTION C-H/Π

Les interactions C-H/π, bien que plus faibles que les liaisons hydrogène, jouent cependant un rôle de stabilisateur structural en biologie.[37] Des calculs ab-initio entre des hydrocarbures et le benzène ont montré une d'énergie d'interaction d'au mieux -12 kJ/mol, soit bien moins qu'une liaison hydrogène.[38] Cela est causé par un effet compensatoire des composantes énergétiques : d'un côté, la partie électrostatique répulsive et d'un autre, une délocalisation énergétiquement favorable des électrons π.

La géométrie de ce système est assez caractéristique et exemplifiée dans la **Figure 10A**. Brandl et Weiss[37] ont défini la géométrie de tels systèmes à travers 1154 chaînes protéiques et ont montré l'existence d'une distance optimale d_{C-X} de 3,7 Å, avec une variation d'environ de ±0,7Å. L'angle entre C-H et le centre du cycle varie de 135° à 156° (**Figure 10B**). Ces distributions sont assez indépendantes de la résolution de la structure puisque les variations sont inférieures à 0,1 Å. Ces auteurs mettent en parallèle l'effet hydrophobe et l'interaction C-H/π : ils soulignent le fait que cette interaction ne peut pas être « opportuniste » et est liée probablement à une réduction de l'entropie vu l'existence d'une directionnalité et d'un recouvrement des sphères de van der Waals entre le carbone donneur et le centre du cycle aromatique (**Figure 10A**).

Au final, l'analyse fournit par Imai indique que ces interactions représentent près de 3% de toutes les interactions non-covalentes protéine/ligand parmi les 6091 entrées de la PDB et sont majoritairement réalisé avec la phénylalanine.[33]

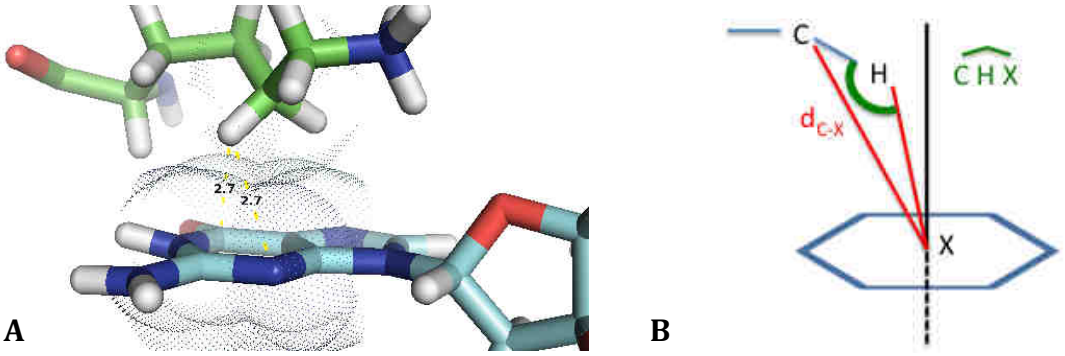


Figure 10 - **A :** Interaction CH-π entre p21-H-ras (en vert) et GNP (en cyan). La sphère de van der Waals du carbone de la lysine 117 intersecte avec celles des atomes du cycle aromatique de GNP. **B** : Représentation schématique d'une interaction CH-π. Le centre de masse du cycle est indiqué par X. Les paramètres géométriques sont aussi décrits : la distance entre le Carbone et le centre de masse d_{C-X} et l'angle formé par l'atome de carbone, l'hydrogène et le centre de masse

2.2.3 LIAISON HALOGENE

Les halogènes sont très souvent utilisés afin de modifier les propriétés pharmacocinétiques des molécules en modulant la réaction métabolique. Grâce leur forte électronégativité, ils peuvent être impliqués dans différentes interactions.

Une liaison halogène correspond à une liaison hydrogène où l'hydrogène est remplacé par un halogène (**Figure 11A**). Cependant, les deux possèdent des charges partielles très différentes : négative pour les halogènes et positive pour l'hydrogène. Afin d'expliquer pourquoi les halogènes sont capables de réaliser de telles liaisons, Politzer a réalisé des calculs quantiques sur de petites molécules halogénées. Il a mis en évidence une zone positive sur le potentiel électrostatique des halogènes, dans le plan C-X où X=Br ou I. Cette observation est reproduite lorsque des substituants électro-attracteurs sont ajoutés à la molécule dans le cas du chlore, du brome ou de l'iode.[39] C'est ce caractère localisé et positif qui permet de réaliser des liaisons halogène. Il a déduit dans le cas de ces molécules que l'énergie d'interaction varie entre 2 et 33 kJ/mol.

La géométrie de telles liaisons a été étudiée par Auffinger en 2004 sur 226 structures protéique/ligand possédant des liaisons halogènes.[40] La distance moyenne $d_{O\text{-}X}$ est dépendante de l'électronégativité de l'halogène : 3,24 Å, 3,15 Å et 3,06 Å pour respectivement l'iode, le brome et le chlore. Celle-ci est en général inférieure à la somme des rayons de van der Waals des deux atomes. Les angles θ_1 et θ_2 ont des valeurs moyennes de 165° et 120° respectivement.

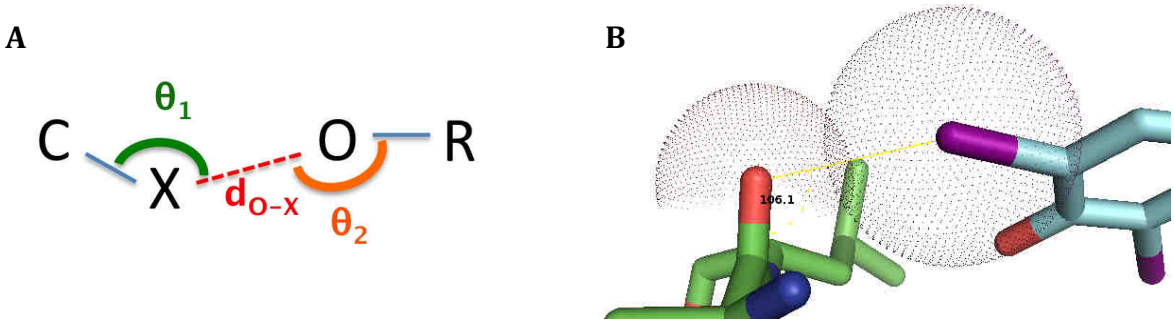


Figure 11 - **A :** Représentation schématique d'une liaison halogène. Cas où le carbone joue le rôle de donneur avec X=F,Cl,Br,I et R=C,P,S. **B:** Leucine 110 de la Transthyretine (en vert) en complexe avec T44 (en cyan). L'iode interagit avec le carbonyle de la leucine en formant un angle de 106.1° avec celui-ci. La sphère de van der Waals de chaque atome est représentée afin de montrer le recouvrement de ces dernières.

Dans de rares cas, des halogènes polarisés peuvent réaliser des interactions avec les électrons π du groupement carbonyle, en adoptant alors une configuration en T, le phénomène restant encore peu étudié. La **Figure 11B** présente une interaction entre un carbonyle et l'iode du ligand. L'angle I--O=C est très proche de 90° montrant ainsi l'interaction avec les électrons π du carbonyle.

Enfin, il existe quelques cas d'interactions halogène/aromatique[41] où les atomes de chlore et de fluor ont tendance à se positionner dans le plan des cycles aromatiques. Cependant, ces études montrent la très faible proportion de ces interactions et fautes de données suffisantes, celles-ci restent encore incomprises.

L'éventail d'interactions qu'offre la nature est large et difficile à étudier, en particulier dans des cas peu observés. Toutes les interactions précédemment décrites possèdent une certaine directionnalité et spécificité. Cependant, dans le contexte des interactions protéine/ligand et même si ce sont des interactions fortes, elles restent occasionnelles et peuvent être considérées comme négligeables en comparaison des liaisons hydrogène, ioniques ou aromatiques.

2.2.4 LIAISON HYDROGENE FAIBLE

Comme précédemment décrit dans la partie **2.1.1**, une liaison hydrogène existe lorsqu'un atome d'hydrogène, lié à un atome plus électronégatif que lui, interagit avec un atome polarisable. Une liaison hydrogène est considérée comme faible si l'hydrogène possède une faible polarisabilité, tel qu'observable dans les liaisons aliphatiques. La distance entre le donneurH et l'accepteurH est alors plus importante que pour des liaisons hydrogène fortes, entre 3 et 4 Å contre 2.2 - 2.5 Å, alors que l'angle tend plus facilement vers 90° que 180°. Ces interactions sont cependant énergétiquement faibles, de l'ordre de 1 kJ/mol contre 20 à 40 kJ/mol pour une liaison forte.[42]

L'analyse réalisée par Desiraju sur 251 structures cristallographiques a permis d'observer la proportion de liaison hydrogène forte et faible dans les interactions protéine/ligand. Cette proportion est ainsi dépendante de la position du donneurH : 34% d'interactions fortes et 66% pour les interactions faibles lorsque le ligand est donneurH contre respectivement 54% et 46% dans le cas d'un atome de ligand accepteurH. Cette proportion majoritaire d'interactions fortes dans le cas de ligand

accepteurH est rassurante puisque les ligands sont composés de bien plus d'accepteurH que de donneurH.[28] De plus, elles sont très souvent couplées entre elles et accentuent le phénomène de liaisons multiples.

Cette coopérativité est souvent présente au sein des liaisons hydrogène. Si l'on considère par exemple un groupement hydroxyle (**Figure 12**), celui-ci peut réaliser jusqu'à deux liaisons hydrogène. Une amine peut effectuer jusqu'à 3 liaisons hydrogène. Ainsi, la coopérativité se conçoit à travers la multiplicité de liaison hydrogène au sein d'un même atome ou groupement d'atomes.

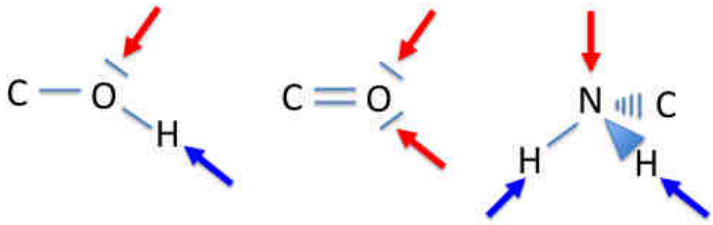


Figure 12 - Zone d'interaction de liaison hydrogène. De gauche à droite : cas d'un hydroxyle et d'un carbonyle où 2 liaisons hydrogène peuvent être réalisées. Cas d'une amine primaire pouvant effectuer 3 liaisons hydrogène. En rouge les potentiels zones d'accepteurs de liaison hydrogène et en bleu les potentiels donneurs de liaison hydrogène.

Sarkhel et Desiraju ont analysé les tendances de chaque groupement à réaliser des liaisons hydrogène multiples.[42] Ils ont ainsi observé parmi 28 structures protéine/ligand que l'oxygène du carbonyle réalise des interactions bidentates et plus rarement tetradentates. Les ethers et les carboxylates sont quant à eux souvent bidentates, avec l'exception du carboxylate qui peut être heptadentate. Lorsque l'on regarde les types de donneurH dans le cas d'accepteurH bi- et tri-dentate, on s'apercoit d'une forte proportion de C-H---O par rapport à O-H---O ou N-H---O. Les interactions C-H---O ayant tendance à être plus faible d'un point de vue énergétique, la multiplicité de telles interactions permet de compenser cette problématique. Cependant, de multiples liaisons hydrogènes sur un même atome peuvent avoir un effet répulsif. Cet effet se traduit par une interaction faible qui tendra à augmenter la distance donneurH-accepteurH d'une liaison hydrogène forte réalisée avec le même atome.

2.2.5 PHENOMENES DE DESOLVATATION

A l'état non complexé, le site de liaison ainsi que le ligand sont tous les deux solvatés (**Figure 13** à gauche). Les atomes polaires réalisent alors des liaisons hydrogène avec le solvant (A et C) tandis que les groupements apolaires interagissent de façon hydrophobe avec celui-ci (B et D). Lorsque le ligand rentre dans la cavité du site de liaison (E), un processus de désolvatation du site et du ligand rentre alors en jeu (F et G). Les interactions réalisées par les molécules d'eau sont ainsi rompues au profit de nouvelles interactions protéine/ligand. Une fois le complexe protéine/ligand formé, le solvant se réorganise. Cette dernière étape prend en compte la création de nouvelles liaisons hydrogènes entre les molécules d'eau et une adaptation de la couche de solvatation autour du ligand et de la protéine. Cette réorganisation modifie l'énergie globale du système, tant d'un point de vue entropique qu'enthalpique. Ces ruptures et formations possèdent un coût énergétique qu'il est nécessaire de compenser pour que la complexation soit possible.

Les phénomènes de solvatation/désolvatation peuvent grandement influencer les l'énergie libre du complexe protéine/ligand. Par exemple, dans le cas de la protéase HIV, l'ajout de groupements propyle sur les azotes du cycle de la (4R,5S,6S,7R)-4,7-dibenzyl-5,6-dihydroxy-1,3-diazepan-2-one permet de gagner 2 unités logarithmiques en terme de constante d'inhibition (**Figure 14**).[43] L'une des explications proposées serait le déplacement de molécules d'eau en contact avec les atomes d'azote du ligand au profit de nombreuses interactions apolaires entre les groupements propyles et des résidus d'isoleucine et de valine. Cet exemple montre bien comment la moindre molécule d'eau peut influencer énormément la capacité d'inhibition des molécules actives.

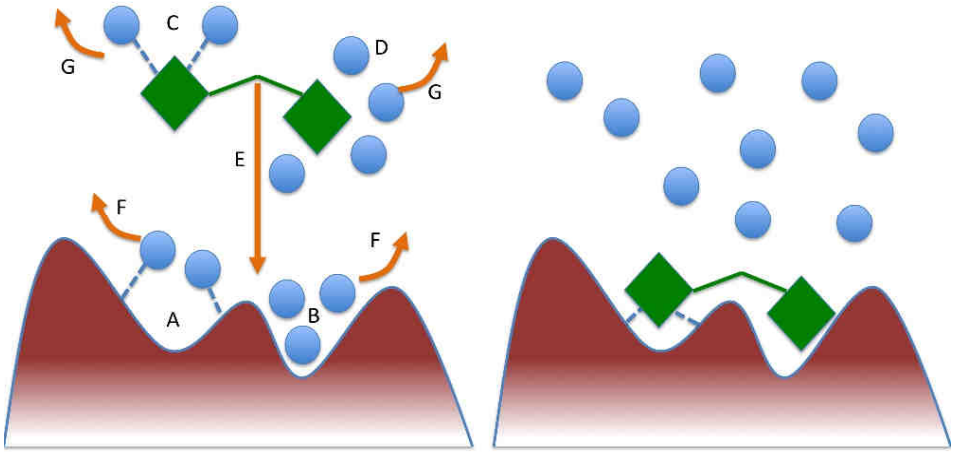
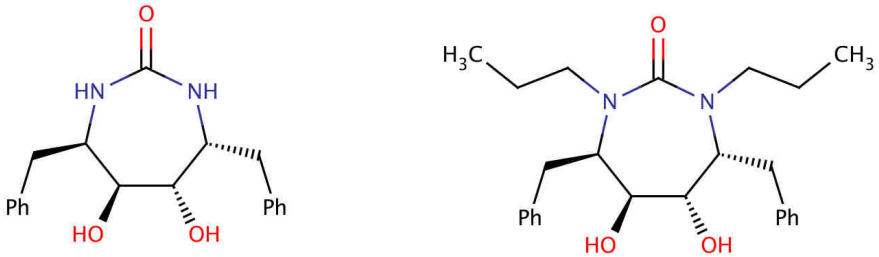


Figure 13 - Processus de complexation protéine/ligand. A l'état initial, les molécules réalisent des liaisons hydrogène ou des contacts apolaires avec la protéine et le ligand (**A,B,C,D**). Lors du processus de complexation (**E**), les molécules d'eau sortent de la cavité du site de liaison (**F**) et le ligand est désolvaté (**G**), rompant ainsi toutes interactions. Une fois le ligand en contact avec la protéine, de nouvelles interactions et/ou contacts apolaires sont formés.



A: K_i=0.2 µM

B: K_i= 8 nM

Figure 14 - A : Structure de la (4R,5S,6S,7R)-4,7-dibenzyl-5,6-dihydroxy-1,3-diazepan-2-one. B : Structure de la (4R,5S,6S,7R)-4,7-dibenzyl-5,6-dihydroxy-1,3-dipropyl-1,3-diazepan-2-one. Les valeurs correspondent aux constantes d'inhibitions des deux composés avec la protéase HIV.

2.3 CONCLUSION

La reconnaissance protéine/ligand est un concept difficile à appréhender tant les paramètres mis en jeux sont nombreux et variés.

D'un côté, il est nécessaire de comprendre la logique structurale du site de liaison d'une protéine. La composition et la géométrie de ces sites sont implicitement liées à sa fonction catalytique mais l'évolution a entrainé des modifications de certains acides aminés à des fins d'adaptation. Malgré cela, de nombreux paramètres sont relativement constants et nous orientent dans la conception de nouveaux médicaments : un site doit être préférentiellement enfoui, relativement apolaire sans pour autant omettre des points d'accroches polaires et le volume de sa cavité doit être relativement important pour laisser un ligand rentrer. Bien que la flexibilité soit un critère permettant de générer des modèles plus réalistes, il est possible de la négliger, tout du moins dans le cadre de criblage virtuel.

D'un autre côté, les interactions non-covalentes offrent un large éventail de possibilités dans le cadre de cette reconnaissance. Certaines d'entres elles sont majoritairement présentes comme les liaisons hydrogène fortes, les liaisons ioniques ou les contacts hydrophobes. Ces interactions, de part leur nombre et leur force, doivent obligatoirement être prises en compte dans la recherche de nouvelles molécules actives. Les liaisons halogène, π/cation ou C-H/π, sont plus occasionnelles et peuvent ainsi être ignorées. Le cas des liaisons hydrogène faibles est assez particulier puisqu'elles sont tout aussi nombreuses que les liaisons hydrogène fortes. De ce fait, ce type d'interaction risque d'induire du bruit qui, se rajoutant à leur contribution énergétique faible, nous pousse ainsi à les ignorer.

De cette première partie découle une conclusion simple : il est nécessaire de simplifier l'information de la reconnaissance protéine/ligand, ou tout du moins tenter d'opter pour une représentation alternative de celle-ci. Dans ce contexte, la deuxième partie se penche sur ces représentations.

3. REPRESENTATION DES INTERACTIONS PROTEINE/LIGAND

L'utilisation de méthodes informatiques a grandement facilité la découverte de nouveaux candidats médicament.[44] En effet, celles-ci permettent de tester si des millions de molécules sont capables d'interagir avec le site de liaison d'une protéine. Cependant, trouver une molécule complémentaire en terme de forme au site de liaison et pouvant réaliser des nombreuses interactions non covalentes avec celui-ci implique une première considération des interactions potentielles que peut fournir le site. La méthode la plus exhaustive possible serait de le considérer tel qu'il est, c'est à dire un ensemble d'atomes spatialement ordonnés où chaque atome est capable de réaliser une à plusieurs interactions. Nous avons vu dans la partie précédente que les sites de liaisons ont un faible pourcentage d'atomes polaires par rapport aux atomes apolaires. En terme d'interactions, cela implique aussi un faible ratio d'interactions potentiellement orientées par rapport aux interactions non spécifiques. Sachant que l'objectif initial est de trouver le meilleur mode d'interaction possible pour une molécule dans une cible afin de prédire son affinité, il est crucial de sélectionner les atomes de protéine les plus favorables pour réaliser des interactions.

3.1 PRINCIPALES METHODES EXISTANTES

Trois grandes étapes caractérisent toute méthodologie de prédiction du mode de liaison d'une molécule à une cible :

1) La préparation de la protéine, c'est à dire la sélection des atomes, groupes d'atomes ou résidus pouvant interagir ou interagissant avec le ligand ;
2) L'optimisation du recouvrement entre le ligand et ces points potentiels d'interaction ;
3) L'attribution d'un score au ligand et à son mode d'interaction.

Nous ne nous focaliserons ici que sur la préparation de la protéine en considérant la tautomérie, les états d'hybridation et la protonation des atomes comme effectués. Ainsi, nous analyserons la représentation des interactions, potentielles ou existantes, réalisées entre une protéine et un ligand.

Deux techniques de représentation du site de liaison sont possibles : simplifier l'information en 3 dimensions, ou la convertir en 1 dimension (**Tableau 3**). Dans le premier cas, les méthodes positionnent des sondes dans la cavité du site de liaison, qui représentent des interactions potentielles et/ou connues. Celles-ci peuvent être attribuées soit à travers l'utilisation d'une grille tridimensionnelle où une recherche exhaustive est réalisée, soit à travers l'analyse atomique du site de liaison, en recherchant les zones de l'espace pour lesquelles une géométrie favorable existe pour une interaction donnée. Dans la seconde, ces sondes sont positionnées sur les atomes de la protéine, du ligand, ou des résidus, afin de coder l'information concernant les interactions sous la forme d'empreinte.

3.1.1 ARRIMAGE MOLECULAIRE

L'une des méthodologies les plus utilisées à l'heure actuelle est l'arrimage moléculaire (*molecular docking*), c'est à dire le positionnement automatique d'une molécule de faible poids moléculaire au sein du site de liaison.[45] La procédure générale de ces outils est divisible en 5 étapes: la représentation de la protéine, celle du ligand, l'échantillonnage conformationnel de la protéine et du ligand, et enfin l'attribution d'un score à chaque pose à travers une fonction de score. La sélection des molécules est finalement réalisée en choisissant celles possédant les meilleurs scores. D'un point de vue algorithmique, ces méthodes essayent de réaliser la recherche conformationnelle et géométrique (translation/rotation) la plus exhaustive. Les contacts apolaires et les liaisons hydrogène sont les deux principales interactions prises en compte afin d'orienter cette recherche exhaustive. Il existe de nombreuses revues se concentrant sur divers aspects de l'arrimage moléculaire tels que l'échantillonnage du ligand ou les fonctions de score.[45–47]

3.1.2 LES PHARMACOPHORES

Les pharmacophores représentent une alternative simple et efficace à l'arrimage moléculaire.[48,49] Selon la définition de l'IUPAC (1998), un pharmacophore est un « ensemble de propriétés stériques et électroniques défini à partir d'une interaction entre deux entités moléculaires et nécessaire pour induire la réponse biologique souhaitée ».[50] Autrement dit, les pharmacophores se basent sur les propriétés physico-

chimiques et les interactions non covalentes potentielles ou existantes des molécules et non sur leurs structures.

Il existe deux familles distinctes de pharmacophores: la première se base seulement sur les ligands et tente d'observer, pour un ensemble de ligands, le meilleur recouvrement possible de propriétés physico-chimiques. La seconde famille requiert une analyse complète de la complémentarité physico-chimique protéine/ligand et de l'environnement spatial, converti ensuite sous la forme de sondes positionnées sur le ligand. Les logiciels permettant de réaliser des criblages pharmacophoriques considèrent en général 6 types de propriétés : DonneurH, AccepteurH, hydrophobe, charge positive, charge négative et aromatique.

3.1.3 LES EMPREINTES

Les empreintes ne codent pas en trois dimensions l'information des interactions protéine/ligand comme peuvent le faire les méthodes précédemment mentionnées. En effet, ces dernières détectent les interactions et/ou les contacts protéine/ligand afin de les coder sous la forme d'une chaîne de caractères. Leur comparaison permet d'évaluer la similarité des modes d'interaction de molécules prédits par l'arrimage moléculaire, d'observer la conservation des modes d'interactions au sein d'une famille de protéine, ou de faire ressortir les caractéristiques des liaisons communes à une série de molécules actives.

Tableau 3 - Liste non exhaustive des méthodes existantes représentant l'information des interactions protéine/ligand

Application	Méthode	Auteurs	Reference
Sondes fictives			
Grille tridimensionnelle			
	GRID	Goodford, Wade, Clark	51–53
Arrimage	**GLIDE**	Friesner, Shenkin	54
	DOCK	Oshiro, Kuntz , Dixon	55
	Hammerhead	Welch,Ruppet,Jain	56
	Surflex	Jain	57
Arrimage/ Pharmacophore	**FLAP**	Baroni, Cruciani, Mason	58
	FLAPPharm	Cross,Baroni,Cruciani	59,60
Sonde géométrique			
Design De novo	**LUDI**	Böhm	61,62
Arrimage	**FRED**	McGann, Brown	63
	GOLD	Verdonk, Taylor	64
	FlexX	Rarey, Klebe	65
Pharmacophore	**LigandScout**	Wolber, Langer	66
	CATALYST	Kurogi, Güner	67
	PHASE	Dixon, Friesner	68
Représentation monodimensionnelle			
Basée sur les ligands			
Post-Traitement d'arrimage	**IF-FP**	Tan, Lounkine,Bajorath	69,70
	AIF	Tan, Bajorath	71
	IASF	Crisman, Sisay, Bajorath	72
Basée sur la protéine			
Post-Traitement d'arrimage	**SIFT/pSIFT/wSIFT**	Deng, Chuaqui, Singh	73–76
	IFP	Marcou, Rognan	77
	IBAC	Kroemer, Stouten	78
	Expanded IFP	Kelly, Mancera	79
	CIF/CHIF/HIF	Mpamhanga, Willet	80
Basée sur la protéine et le ligand			
Post-Traitement	**APIF**	Pérez-Nueno, Teixido	81
Phamacophore	**Pharm-IF**	Sato, Yokohama	82

3.2 SONDES FICTIVES

Dans toute cette partie, on considèrera une sonde fictive comme un point dans l'espace représentant la position possible d'un atome ou d'un groupe d'atomes d'une molécule pouvant réaliser une interaction non covalente avec la protéine. Représenter la protéine sous la forme de ses points d'ancrage potentiels avec le ligand permet de limiter l'information protéique aux seules interactions non covalentes.

On peut distinguer deux méthodologies différentes pour définir ces sondes : la première consiste à discrétiser la cavité du site de liaison sous la forme d'une grille tridimensionnelle afin d'analyser l'environnement protéique local à chaque cube. La seconde méthode se focalise sur la pré-sélection des atomes protéiques d'intérêt, tels que les donneurs et les accepteurs de liaison hydrogène, afin de positionner des sondes réalisant une géométrie raisonnable pour chaque type d'interaction.

3.2.1 GRILLE TRIDIMENSIONNELLE

La discrétisation d'un objet complexe sous la forme d'une grille est une approche facilement réalisable qui permet d'analyser de façon exhaustive des effets locaux. Le but est alors de trouver la position et la propriété physico-chimique des atomes de ligand nécessaires pour réaliser les interactions non covalentes les plus énergétiquement favorables avec la protéine.

GRID[51] est la première méthode recherchant en tout point de l'espace la contribution énergétique des atomes de protéine avoisinants. Pour cela, elle génère une grille de telle sorte que toutes ses arêtes soient à l'extérieur de la protéine. La grille est ensuite discrétisée en un ensemble de points, espacés de façon régulière (0.5 Å par défaut). Chacun de ces points est alors remplacé par une ou plusieurs sondes, chacune représentant un type d'interaction avec un groupement chimique particulier: amino NH_3^+, Oxygène d'un carboxylate =O et O^-, hydroxyle OH, méthyle CH_3 et l'eau H_2O (34 sondes sont actuellement disponibles dans MOE[83]). Pour chaque point, le potentiel énergétique de chaque sonde est calculé en fonction de l'environnement local (**Figure 15.2**), afin la somme de toutes les sondes est effectuée afin d'obtenir le potentiel énergétique d'un point selon la formule de l'**Équation 1**.

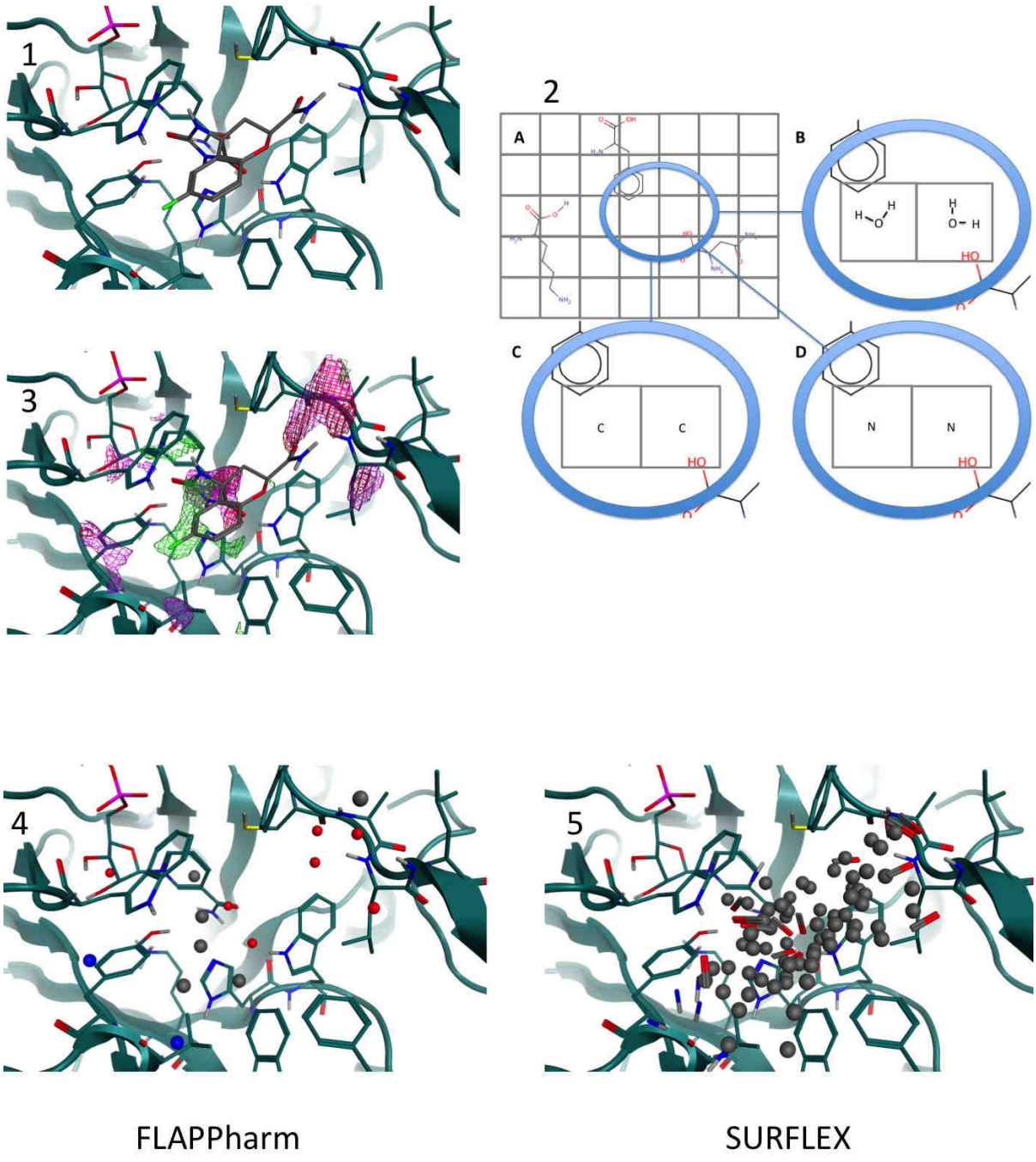


Figure 15 - Aldose réductase humaine en complexe avec Fidarestat (Code PDB :1PWM). Le site de liaison est détecté autour du ligand (1). Une grille tridimensionnelle est générée (2A). Pour chaque cube diverses sondes sont positionnées et orientées suivant l'environnement protéique (2B : sonde H2O; 2C : sonde CH4 ; 2D : sonde N). Un champ d'interaction moléculaire est alors déduit de ces sondes (3) avec une énergie définie à -6, -4.4, -2.6 kcal/mol pour les sondes H2O (en violet), N (en rouge) et CH4 (en vert), respectivement. Ces potentiels sont ensuite convertis avec FLAPPharm sous la forme de points (4). Bien que Surflex n'utilise pas le champ d'interaction moléculaire, les procédures 1 et 2 sont identiques, donnant ainsi le protomol (5) : en noir les sonde CH4 (hydrogène non représenté), les sondes C=O (en rouge) et N-H (en bleu)

$$E = \sum E_{ij} + \sum E_{el} + \sum E_{hb}$$

Équation 1 - Formule du calcul du potentiel énergétique selon GRID. E représente l'énergie totale, E_{ij} le potentiel de Lennard Jones, E_{el} la composante électrostatique et Ehb la fonction décrivant l'énergie d'une liaison hydrogène.

La composante E_{hb}, qui mesure le potentiel en fonction de la géométrie propre à la liaison hydrogène, a fait l'objet de différentes améliorations afin de prendre en compte la multiplicité de ce type de liaisons pour un même atome.[30,52,53] Une fois l'énergie calculée en chaque point, ceux correspondant à des énergies favorables sont regroupés afin de générer un champ d'interaction moléculaire (MIF), c'est à dire un ensemble de potentiels énergétiques (**Figure 15.3**).

Diverses applications sont dérivées, directement ou indirectement, de la méthode GRID. Dans le cadre de l'arrimage moléculaire, GLIDE,[54] DOCK,[55] Hammerhead[56] et SURFLEX[57] utilisent ce principe. Tous discrétisent la cavité du site de liaison sous la forme d'un ensemble de points. Pour DOCK, 3 potentiels sont utilisés : le potentiel attractif et dispersif de van der Waals, ainsi que le potentiel électrostatique Coulombien. Pour Hammerhead, ce sont les sondes H, C=O et N-H qui sont employées. Enfin, SURFLEX altère légèrement ce principe en ne regardant que les cubes en contact direct avec le site de liaison. Pour chacun des cubes sélectionnés, il positionne d'abord des sondes hydrophobes CH_4 et des sondes polaires C=O et N-H dans 36 directions de l'espace, puis calcule les énergies d'interactions de chaque orientation afin de ne garder que celles possédant un score énergétiquement favorable. Les positions et les caractéristiques de chaque sonde sont enregistrées sous la forme d'un *protomol* (**Figure 15.4**), prêtent à être utilisées pour l'arrimage moléculaire.

Le domaine méthodologique des pharmacophores utilise aussi les grilles tridimensionnelles. FLAP (Fingerprint for Ligand And Proteins)[58] génère pour le site de liaison le champ d'interaction moléculaire qu'il condense sous la forme de points pharmacophoriques (**Figure 15.5**). Cette procédure, initialement prévue pour les sites de liaison, peut être appliquée sur les ligands en orientant les sondes par rapport aux atomes de ligands afin de dériver un champ d'interactions pharmacophoriques (PIF). La superposition des champs MIF et PIF permet une comparaison et l'obtention des points pharmacophoriques en commun. Ces points sont alors convertis sous la forme de d'empreintes en générant l'ensemble des combinaisons de 4 points pharmacophoriques (**Figure 16**). Les FLAP sont alors utilisables pour comparer les modes d'interactions entre différentes protéines et arrimer des molécules dans les sites de liaisons.

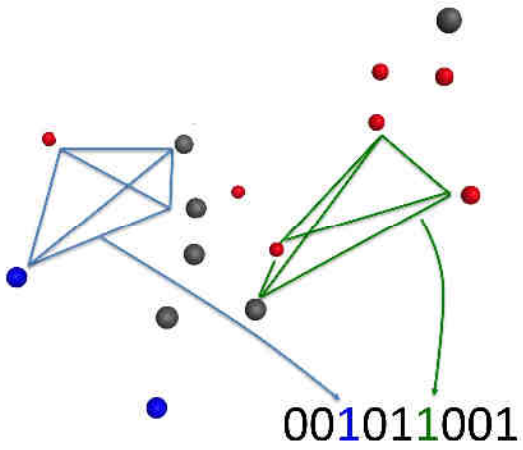


Figure 16 - Représentation d'un champ moléculaire d'interaction. Toutes les combinaisons de 4 points sont générées (en vert et en bleu) et associées à une position dans la chaîne d'entiers.

FLAPpharm[59] est une récente amélioration de FLAP. Développé en 2012 par Cross, celle-ci utilise les champs d'interactions pharmacophoriques pour trouver les caractéristiques communes d'un ensemble de ligands actifs pour une même cible. Toutes les conformations des molécules sont générées et leurs empreintes FLAP déduites. Une mesure de similarité est alors calculée entre tous les empreintes afin de trouver les 5 meilleurs groupes de conformations, c'est à dire les 5 groupes possédant les meilleures valeurs de similarité globale. Une fonction de score basée essentiellement sur la forme et sur le poids de chaque sonde sélectionne les meilleurs ensembles de conformères. Bien que le site de liaison soit ignoré durant la procédure, le modèle recouvrant correctement la forme du site est toujours trouvé parmi les 5 premiers. Cela a été testé sur un jeu de données de 81 cibles comprenant 960 molécules. Parmi ces cibles, FLAPpharm est capable de retrouver près de 67% des conformations bioactives.[60]

L'utilisation de grilles tridimensionnelles présente l'avantage d'évaluer des effets locaux afin de conclure sur le potentiel de divers types d'interaction dans la cavité du site de liaison. Ainsi, GRID considère, pour un endroit donné de l'espace, non pas une mais un ensemble d'interactions possibles, plus ou moins favorables selon le contexte protéique. Il est important de souligner que l'utilisation des champs d'interactions moléculaires est très intuitive et permet d'orienter le chimiste médicinal dans les modifications à apporter à la structure du ligand afin de maximiser les interactions avec le site de liaison. Cependant, le nombre de sondes peut être très important et la quantité d'information fournie par les MIF bruts trop grande pour être facilement utilisable en l'état.

3.2.2 SONDES GEOMETRIQUES

La discrétisation du site de liaison permet de réaliser une recherche exhaustive de toutes les interactions potentielles entre un site de liaison et une molécule. Cependant, cela implique une grande quantité d'information à traiter qui joue énormément tant sur les temps de calculs que sur la précision. Orienter la sélection en regardant les groupes fonctionnels connus pour réaliser des interactions non-covalentes fortes permet alors de s'affranchir de cette recherche.

LUDI[61,62] est un programme développé en 1992 par Böhm, dans le cadre du *de-novo design*, c'est à dire de la génération automatique de nouvelles molécules s'adaptant à la forme et aux propriétés physico-chimiques du site de liaison. Ce programme utilise des sondes, appelées points d'interactions, qu'il considère comme étant toute position dans l'espace non occupé par la protéine, où un atome ou un groupe fonctionnel d'une molécule peut réaliser des interactions favorables avec le site de liaison. Quatre types de sondes sont prises en compte : AccepteurH, DonneurH, Aromatique et Aliphatique.

Le choix du positionnement des sondes est laissé à l'utilisateur : l'utilisation de règles géométriques simples, définies après analyse statistique des interactions non covalentes de la CSD, ou alors en utilisant le champ d'interaction moléculaire de GRID. Dans le premier cas, la distance hydrogène-AccepteurH et les angles hydrogène--AccepteurH-Atome lié et DonneurH-hydrogène--AccepteurH sont respectivement définis à 1.9 Å, 120° et 180° (**Figure 17**). Pour les atomes hydrophobes, un ensemble de points est disposé de façon régulière dans l'espace sur une sphère de 4 Å centrée sur l'atome. Pour les aromatiques, deux points, en dessous et au dessus du cycle, sont ajoutés à 6 Å du centre du cycle. L'environnement des sondes est vérifié afin de s'assurer qu'aucune gêne stérique n'existe avec la protéine.

La seconde étape consiste alors à maximiser la superposition entre les sondes et un ensemble de fragments. Une fois ces derniers sélectionnés et alignés sur les sondes, un ensemble de petits fragments (« linkers ») les associe pour fournir une liste de molécules.

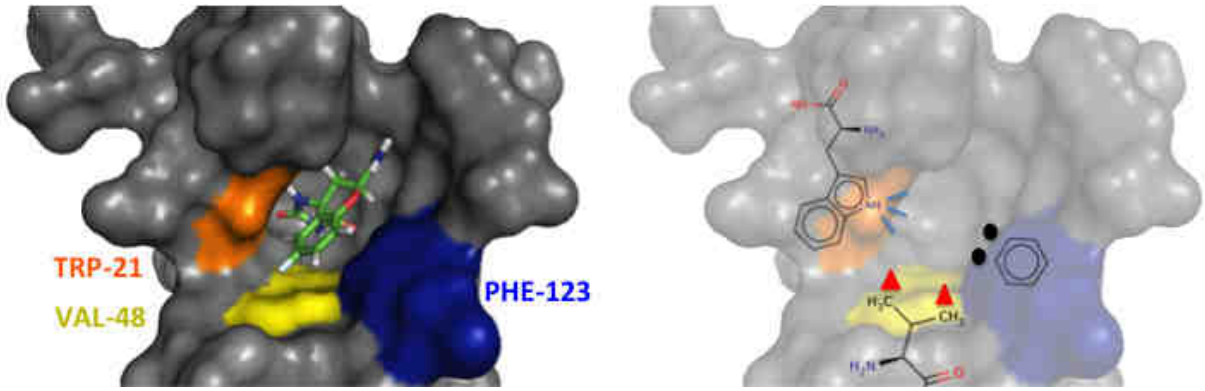


Figure 17 - Exemple de placement des sondes selon LUDI. Les résidus du site de liaison sont d'abord détectés. Pour chaque atome, différentes sondes sont positionnées dans l'espace en fonction des atomes ou groupes d'atomes de la protéine. (Trait bleu : donneur de liaison hydrogène ; Rond noir : Aromatique ; Triangle rouge : Contact apolaire).

Équation 2 - Equation de score LUDI pour les liaisons hydrogène. ΔR correspond à la déviation par rapport à la longueur idéale de 1.9 Å des liaisons hydrogène de H--O/N. $\Delta\alpha$ est la déviation de l'angle D-H--A par rapport à sa valeur idéale de 180°. NCONTACT représente la surface de contact lipophile entre la protéine et le ligand en Å^2.

L'une des spécificités de LUDI est la création d'une fonction de score pour quantifier la géométrie des liaisons hydrogène, telle que définie dans l'**Équation 2**. Cette dernière implique que plus on s'éloigne de la distance optimale de 1.9 Å et d'un l'angle D-H-A de 180°, plus le score diminue. Similaire à la composante E_{hb} de GRID, l'avantage d'une telle définition est de pouvoir distinguer des liaisons fortes de liaisons faibles.

LUDI génère l'ensemble des points d'interaction potentiels, mais ne réalise aucune sélection pour trouver les meilleurs candidats. Dans cette optique, Barillari utilise une approche basée sur une discrimination des atomes de protéine.[84] Ainsi, l'ensemble des atomes des sites de liaisons de près de 3500 entrées de la sc-PDB[7] a été analysé en codant leurs propriétés, les interactions qu'ils effectuent avec le ligand et leurs environnements locaux. Une machine d'apprentissage a été employée afin de discriminer des atomes connus pour réaliser des interactions avec le ligand de ceux n'interagissant pas. La combinaison de cette information avec les points de LUDI permet de créer un pharmacophore d'interaction potentiel plus restreint et donc plus précis.

Cependant, cette sélection reste biaisée par la faible diversité des interactions (liaison Hydrogène et contact hydrophobe), limitant ainsi la qualité des modèles pharmacophoriques.

Rarey a dérivé le concept de Böhm afin de réaliser un algorithme d'arrimage moléculaire: FlexX.[65] Il part alors du principe que toutes les interactions intermoléculaires peuvent être classées selon la force de leurs contraintes géométriques. Le logiciel considère les mêmes types d'interactions que LUDI à la différence qu'il utilise des cônes afin de représenter les zones géométriquement favorables. Lors de l'étape de positionnement du ligand, les cônes des atomes de la protéine comme du ligand sont observés afin de conclure à un recouvrement des cônes par les atomes de l'autre entité moléculaire (**Figure 18**). Analytiquement parlant, ces cônes sont très facilement définissables via des calculs de distances et d'angles et ne sont pas plus complexes que d'autres stratégies de détection de liaisons hydrogène.

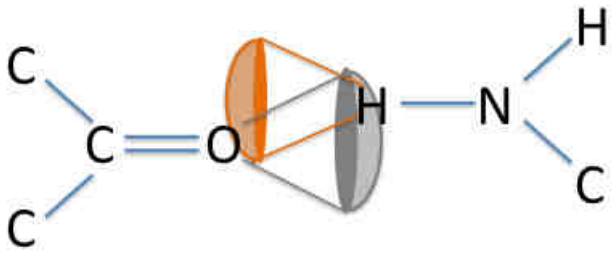


Figure 18 - Représentation d'une liaison hydrogène dans FlexX. Le cône correspondant aux positions favorables pour qu'un hydrogène puisse interagir avec l'oxygène du carbonyle est représenté en gris. Dans le cas des positions de l'accepteur de liaison hydrogène, le cône est représenté en orange. Une interaction existe si l'hydrogène se positionne dans le cône gris et l'oxygène dans le cône orange.

GOLD[64,85] détecte quant à lui un ensemble de potentiels accepteurH et donneurH qu'il introduit comme chromosome d'un algorithme génétique permettant d'associer une éventuelle liaison Hydrogène protéine/ligand.

Les outils de pharmacophores se basent essentiellement sur ce concept de sondes géométriques. Bien que la nomenclature les définisse comme étant un ensemble de propriétés, un point pharmacophorique se caractérise par une position donnée de l'espace possédant une propriété physico-chimique et ressemble ainsi fortement à notre définition de sonde. LigandScout,[66] Catalyst[67] et PHASE[68] sont trois programmes commerciaux développés respectivement par Inte:Ligand, Accelrys et Schrödinger qui permettent de réaliser des criblages virtuels basés sur des modèles pharmacophoriques.

Catalyst et LigandScout dérivent leurs modèles pharmacophoriques selon l'approche de Greene[86] tandis que PHASE a développé sa propre définition. Pour ces trois méthodes, nous nous focaliserons seulement sur leurs représentations des complexes protéine/ligand et non sur les ligands seuls. Contrairement à LUDI et FlexX, les sondes sont positionnées sur les atomes ou les groupes d'atomes du ligand et codent chacune une interaction protéine/ligand existante. Une sonde représente alors une interaction, dont la directionnalité peut être explicitement déterminée. La différence entre LigandScout et les autres logiciels résident dans la possibilité d'affecter à une même sonde des propriétés multiples et ainsi d'affiner la représentation. Par conséquent, un groupement guanidinium pourra être considéré comme un positif ionisable mais aussi un donneur de liaison hydrogène. Les pharmacophores (**Figure 19**) se résument ainsi en un ensemble de sondes, appelées sphères, colorées par la nature des interactions non-covalentes.

Les pharmacophores sont utilisés dans le cadre du criblage virtuel. Ainsi, les modèles pharmacophoriques de chaque molécule sont comparés et alignés aux modèles de référence. Dans une récente étude comparant divers outils de pharmacophore sur des criblages virtuels de 4 différentes cibles,[87] LigandScout et Catalyst donnent des résultats relativement similaires, avec une augmentation des facteurs d'enrichissement pour LigandScout.

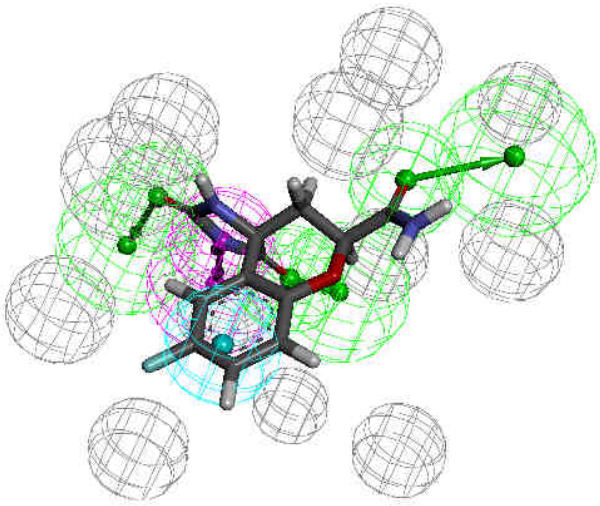


Figure 19 - Pharmacophore généré avec Catalyst du Fidarestat (en noir), en complexe avec l'aldose réductase humaine. Sphères grises: sphères d'exclusion ; sphères bleues: hydrophobes ; sphères magenta: donneurs de liaison H ; sphères vertes: accepteurs de liaison H. La directionnalité de la liaison hydrogène est codée par un vecteur reliant donneurH et accepteurH.

Cependant, en combinant les outils deux par deux, l'utilisation de LigandScout en préfiltre d'une autre méthode montre un enrichissement bien supérieur. Pour pouvoir correctement comparer les méthodes, les auteurs prennent en compte la totalité du concept, c'est à dire la définition des pharmacophores, l'algorithme d'alignement et la fonction de score. Catalyst et LigandScout considèrent tous deux le rayon des sondes afin de quantifier la qualité des alignements des sondes et la seule différence notable réside dans leurs fonctions de scores : Catalyst évalue la distance entre la propriété moléculaire et la sonde associée en respectant le rayon des sondes, tandis que LigandScout utilise un critère géométrique simple de recouvrement des propriétés physico-chimiques. Bien que les auteurs ne se soient concentrés que sur l'aspect coût en terme de temps de calculs, ils ne fournissent pas de réelles explications sur le sujet. On peut alors se demander si la différence de résultats de cette étude est due aux fonctions de scores utilisées ou à la prise en compte de propriétés multiples par sonde.

L'avantage de ce type de représentations réside dans sa simplicité : c'est un ensemble de points caractéristiques des interactions protéine/ligand. De plus, des sphères d'exclusion, généralement associées aux atomes tapissant le site de liaison de la protéine, permettent de coder la forme du site de liaison et ainsi d'améliorer la définition. Cependant, tous les points d'ancrage de la protéine ne réalisant pas forcement des interactions non covalentes avec le ligand, certains sont ignorés, résultant dans une perte d'information. Généralement, l'utilisation de plusieurs ligands corrige cette approximation, mais à condition de posséder des molécules aux modes d'interactions différents.

Une analogie peut être facilement observée entre les pharmacophores et HS-Pharm.[84] En effet, les pharmacophores vont regarder l'occurrence d'une interaction à une position de l'espace parmi une série de ligands tandis qu'HS-Pharm va prédire la position dont l'environnement protéique est le plus similaire à des positions connues pour réaliser des interactions. En somme, on tente dans les deux cas de ne considérer dans la sélection et le score que les interactions qui contribuent vraiment à la liaison protéine/ligand, améliorant ainsi la prédiction de nouvelles molécules.

Le concept de sondes géométriques, bien que normalement concret, peut aussi être réalisé à travers un certain niveau d'abstraction. Le logiciel FRED,[63] développé par OpenEye, représente ainsi une molécule par un ensemble de Gaussiennes.[88–90] Ces dernières, qui ont l'avantage d'être facilement manipulables mathématiquement, permettent de maximiser un recouvrement entre 2 ensembles, c'est à dire dans le cas présent, une molécule et la cavité du site de liaison. En plus de la caractérisation géométrique des objets, FRED introduit la notion de propriétés physico-chimiques associées aux fonctions gaussiennes, permettant alors de maximiser la reconnaissance moléculaire protéine/ligand. Cette méthode permet ainsi de ne plus s'intéresser à la position atomique en soit mais à un ensemble de formes et de propriétés.

3.2.3 Conclusion

L'utilisation de sondes fictives est un moyen efficace pour représenter l'information. Elle permet une définition suffisamment fine et précise des interactions non-covalentes pour pouvoir positionner et sélectionner des molécules de faible poids moléculaire. Un risque dans ce type de méthodologies est le surplus d'information. En effet, des méthodes telles que GRID fournissent des potentiels énergétiques, difficilement analysables et par conséquent pouvant induire en erreur lors de la transformation en un ensemble de points. A l'opposé, les méthodes basées sur les pharmacophores codent à la fois les interactions mais aussi les contraintes spatiales imposées par le site de liaison (sphère d'exclusion). Cependant, ces dernières sont limitées par les modes d'interactions existants pour une cible et ne permettent pas d'extrapoler à d'autres interactions potentiellement favorables. Dans le cadre de l'arrimage moléculaire, l'une des grandes problématiques actuelles réside dans le processus d'attribution de score à chaque molécule arrimée. En effet, les fonctions de scores inhérentes au logiciel d'arrimage n'arrivent pas à prédire correctement l'affinité des molécules et donc à discriminer des molécules potentiellement actives de molécules inactives.

3.3 REPRESENTATION MONODIMENSIONNELLE

Contrairement aux méthodes précédentes utilisant les sondes positionnées dans l'espace pour représenter les interactions protéine/ligand, ce type de représentation transforme les interactions en empreintes. Ces dernières, représentées sous la forme d'une chaine de caractères, ont l'avantage d'être facilement comparables aux moyen de mesure de similarité.[91] Cependant, ils sont difficilement interprétables et ne permettent pas de revenir à une dimension supérieure. Il existe trois moyens de coder les interactions, en fonction du référentiel : côté protéine à travers les atomes ou les résidus, côté ligand à travers les atomes ou les groupements chimiques et à travers les interactions elles-mêmes. L'utilisation d'empreintes pour coder l'information structurale ou physico-chimique est un domaine à part entier. Les clés structurales tels que les clés MACCS[92] ou les empreintes ECFP[93] décrivent la composition chimique et les motifs structuraux sous la forme de chaîne d'entiers. Chaque position de cette chaîne décrit alors un atome, un groupement chimique, une propriété physico-chimique ou une combinaison de ces derniers. Dans le cas présent, chaque position décrit une interaction, un ensemble d'interactions, un atome ou un groupe d'atomes réalisant des interactions.

3.3.1 BASEE SUR LE LIGAND

L'information des interactions protéine/ligand peut être implicitement codée dans le graphe moléculaire du ligand. Une telle représentation permet de réaliser du criblage virtuel à haut débit, en recherchant des molécules possédant des atomes ou groupements d'atomes connus pour réaliser certaines interactions avec la protéine.

Tan, Batista et Bajorath ont largement contribué à ce type de représentations. L'idée sous-jacente de leur méthode consiste à dire que tous les atomes d'un ligand ne réalisent pas d'interaction non-covalente avec la protéine et de ce fait, seuls ceux interagissant devraient être considérés lors de la génération de clés structurales. Les IF-FP (*Interacting Fragment-Fingerprint*)[69] sont définis en détectant en premier lieu à partir des complexes protéine/ligand, les interactions protéine/ligand à travers de simples règles de distances : 3.8 Å pour les liaisons hydrogène et 4.5 Å pour les interactions ioniques et les contacts apolaires. Dans un second temps, seuls les atomes en interaction sont conservés puis convertis en chaîne binaire au moyen des clés MACCS.

Tan a ensuite évolué les IF-FP en IF-TFP (*Interaction Fragments-Transfert Fingerprint*)[70] afin de transférer l'information structurale d'un ligand d'une cible A sur d'autres ligands dont la cible B n'est pas cristallisée. Ils utilisent alors une union des clés MACCS originelles des ligands de la cible B avec les IF-FP de la cible A en supposant une similarité de mode d'interactions pour les deux cibles.

L'une des modifications les plus intéressantes concerne les 3D-IFS-FP.[94] Ces derniers calculent des poids à partir de l'occurrence de chaque fragment parmi tous les ligands actifs d'une cible. Cette modification permet de distinguer des fragments souvent présents parmi les molécules et ainsi d'orienter la sélection de molécules lors d'un criblage virtuel. Sur les dix cibles testées, les 3D-IFS-FP ont largement dépassés les IF-FP et les clés MACCS d'un facteur 0.1 à 10.

Les AIF[71] constituent une légère modification de IF-FP puisqu'ils remplacent les clés MACCS par des ECFP4. En comparaison des IF-FP, les AIF apportent une augmentation minime des facteurs d'enrichissement. Enfin, un dérivé des AIF, l'IASF (*Interaction Annotated Structural Features*) a été développé par Crisman afin d'associer à chaque fragment une contribution énergétique.[72] L'idée sous-jacente est d'aller au-delà de l'occurrence des fragments en incorporant directement l'information des interactions. Chaque fragment dérivant des IF-FP se voit associer une valeur, correspondant à la somme des contributions énergétiques de chacun des atomes. Pour cela, la fonction de score de FlexX est utilisée en prenant en compte les liaisons hydrogène, les interactions ioniques et les contacts apolaires. On notera des résultats ambivalents et souvent équivalents à l'arrimage moléculaire

Toutes ces méthodes montrent l'importance de la sélection et de la représentation des interactions. La problématique inhérente à ces méthodes provient de l'utilisation de chaînes d'entiers basés sur la structure (ECFP/MACCS) et non les propriétés physico-chimiques. Sachant qu'une même interaction peut être réalisée par différents groupements chimiques, la recherche de nouveaux châssis est rendue difficile par cette approche. Elle permet cependant de clairement distinguer des groupements chimiques importants et impliqués dans des interactions non-covalentes avec la protéine. Le fait d'associer des poids est d'autant plus crucial qu'il affine la discrimination des groupements importants par rapport au bruit.

3.3.2 BASEE SUR LA PROTEINE

Positionner l'information des modes d'interaction sur la protéine permet d'induire une constante à travers les atomes et résidus du site de liaison. En effet, dans le cas des empreintes basées sur les ligands, l'information est toujours dépendante de la structure initiale. Dans le cas de multiples ligands actifs pour une même cible, la combinatoire peut très vite devenir très importante. Se positionner sur la protéine permet ainsi, pour une même cible, d'avoir un référentiel fixe et d'être indépendant de la structure des ligands. Dans ces conditions, plusieurs référentiels sont possibles : les atomes et les résidus.

Deng, Chuaqui et Singh ont été les tout premiers à convertir les interactions protéine/ligand sous la forme d'une chaîne binaire. Leur méthode SIFt[73] consiste à convertir chaque résidu du site de liaison sous la forme de 7 nombres binaires, décrivant ainsi des spécificités d'interaction non covalentes (**Figure 20**):

1. Le résidu est en contact avec le ligand
2. L'interaction implique un atome de la chaîne principale
3. L'interaction implique un atome de la chaîne latérale
4. L'interaction est polaire
5. L'interaction est apolaire
6. L'accepteur de liaison hydrogène est localisé sur la protéine
7. L'accepteur de liaison hydrogène est localisé sur le ligand

Cette représentation a été utilisée avec succès dans le cadre de la sélection de poses d'arrimage moléculaire et la classification de modes d'interaction dans le cas de multiples molécules actives pour une cible. De nombreuses améliorations de SIFt ont vu le jour. Parmi elles, r-SIFt[74] réalise une analyse et une agglomération des modes d'interaction d'une molécule dont les substituants sont modifiés afin d'aider à la génération de librairies focalisées. w-SIFt[75] étend le concept des SIFt afin de prendre en compte l'importance relative de chaque interaction. Très similaire aux 3D-IFS-FP, les w-SIFt ont montré, dans un jeu de données de 89 molécules, leur capacité à ordonner des molécules selon leur activité et ainsi aider à la sélection de nouvelles molécules. Enfin, p-SIFT[76] dérive des profils d'interactions permettant d'inférer la probabilité de l'existence d'une interaction à une position donnée du site de liaison.

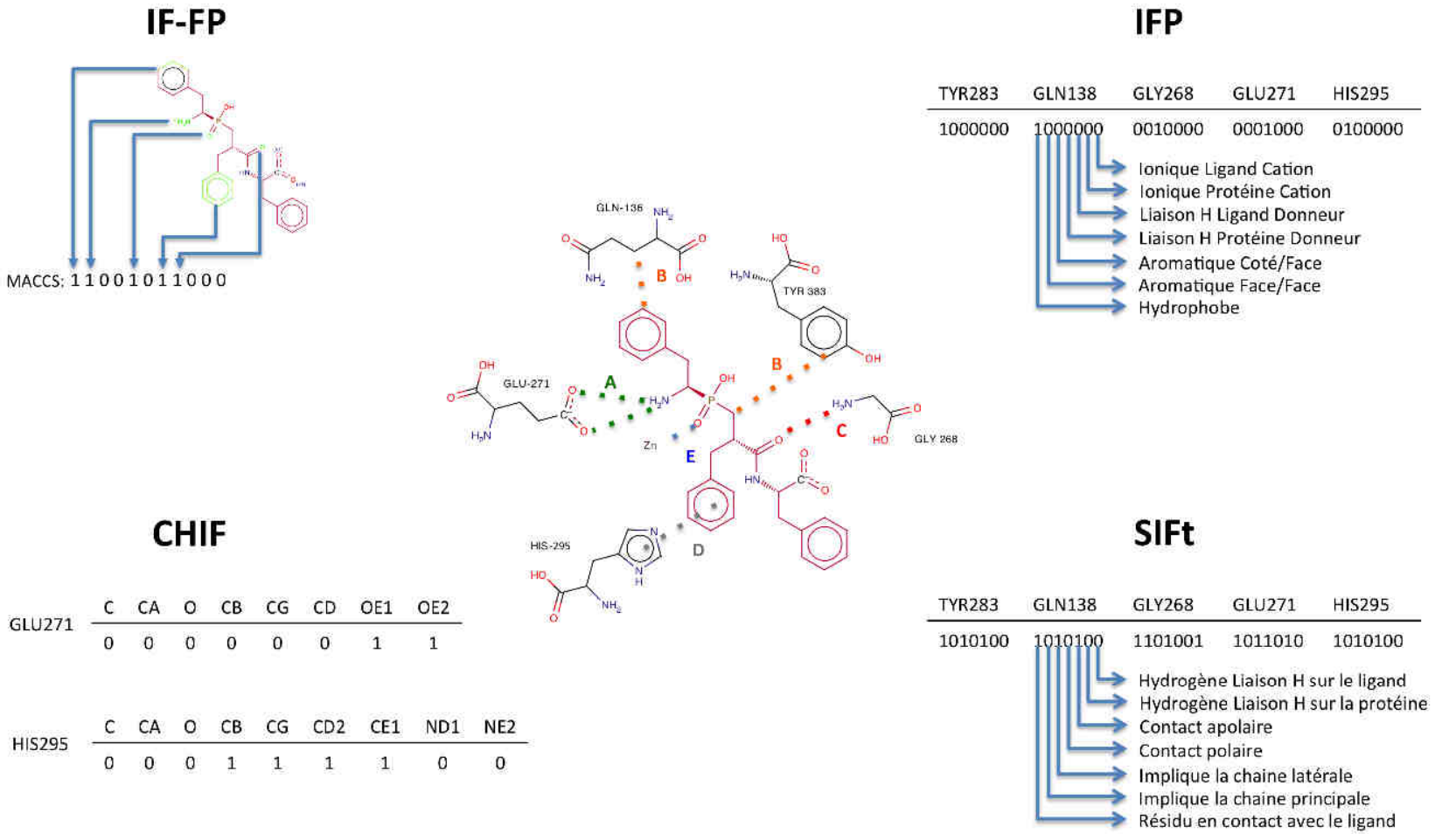


Figure 20 - Représentation des modes d'interaction de la leukotriène A4 hydrolase en complexe avec l'inhibiteur RB3041, selon 4 méthodes. En haut à gauche : Les IF-FP ne gardent que les atomes du ligand en interaction et convertissent les fragments atomiques sous la forme de clés MACCS. En haut à droite : les IFP caractérisent chaque résidu du site de liaison par les interactions qu'il peut effectuer avec le ligand. En bas à gauche : Les CHIF affectent une valeur de 1 à chaque atome de chaque résidu du site de liaison s'il réalise une interaction avec le ligand. En bas à droite : Les SIFt décrivent le type d'interaction et le positionnement de celle-ci par rapport à la chaîne principale ou secondaire. Note : afin d'être lisible, chaque description ne représente qu'une partie de la chaîne d'entiers ou de la définition du site de liaison.

Les IFP de Marcou[77] décrivent chaque résidu du site de liaison par un ensemble de cases représentant les interactions non-covalentes existantes : liaisons hydrogène forte et faible, interaction aromatique, ionique, métal/accepteur, contact apolaire et π-cation. Lorsqu'une interaction est détectée entre le ligand et le résidu, la case correspondant au type d'interaction est alors définie à 1, ou 0 dans le cas contraire. La longueur de la chaîne d'entiers est donc dépendante du nombre de résidus et souffre par conséquent d'un manque de transférabilité vers des sites de taille variable. Cependant, ce principe a été utilisé avec succès dans le cas de post-traitements des poses d'arrimage moléculaire. Les modes d'interactions des poses sont comparés aux modes d'interactions des ligands de référence afin de ne sélectionner que ceux mimant le mieux les modes d'interactions des ligands de référence. Une approche alternative des IFP dans le cas du post-traitement a été implémentée par Kroemer : IBAC (Interactions-Based Accuracy Classification).[78] Afin d'inférer la qualité d'une pose d'arrimage moléculaire, Kroemer considère qu'une pose est correctement arrimée si elle reproduit au moins 75% des interactions par rapport au ligand de référence. Cette méthode simple représente alors une alternative au calcul de déviation des coordonnées atomiques (rmsd) pouvant facilement induire en erreur.

Alternativement aux chaînes basées sur les résidus, Kelly et Mancera[79] ont mis en place un nouveau type de chaînes d'entiers, basé sur les atomes de protéine : Extended-FP. La longueur de la chaîne dépend du nombre d'atomes du site de liaison pouvant réaliser des liaisons hydrogène avec le ligand. Trois approches sont alors développées : assigner une valeur de 1 à chaque case lorsque l'atome de protéine réalise une liaison hydrogène, définir une valeur en fonction de l'accessibilité au solvant et de la force de l'interaction, ou utiliser la fonction de score de Böhm, précédemment décrite. L'originalité de cette méthode réside dans le regroupement de plusieurs liaisons hydrogène se situant dans la même zone du site actif. Lors de la mesure de similarité, les interactions au sein d'un même groupe sont ainsi favorisées afin de prendre en compte des changements locaux dans les interactions (**Figure 21**). Bien que les différences soient minimes, dans le cadre du criblage virtuel la méthode d'association est aussi efficace que celle basée sur l'accessibilité au solvant ou celle utilisant la fonction de score de Böhm. Cela prouve ainsi qu'une simple méthode géométrique est tout aussi efficace que l'utilisation de fonction empirique, plus coûteuse en temps de calcul.

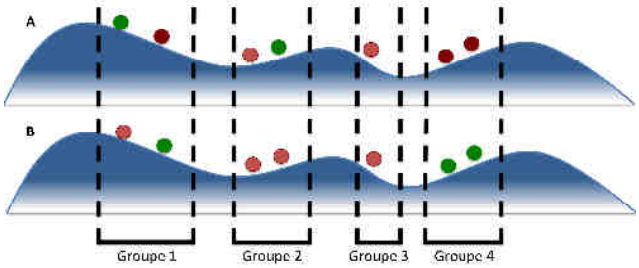


Figure 21 - Représentation schématique de l'association selon Kelly et Mancera. Les liaisons hydrogène, représentées par des cercles, qui sont spatialement proches sont groupées ensembles. En vert, les liaisons hydrogène réalisées avec un ligand. En rouge, les liaisons hydrogène potentielles. **A** et **B** représentent les modes d'interactions de deux molécules. Lors de la mesure de similarité, le groupe 1 sera considéré comme identique dans A&B, bien que les interactions ne soient pas réalisées avec les mêmes atomes de protéine. Le groupe 2 sera partiellement différent, 3 identique et 4 complètement différent.

Afin d'observer l'apport de contacts apolaires par rapport aux polaires, 3 chaînes binaires ont été développées par Mpamhanga,[80] dont la longueur correspond au nombre d'atomes lourds dans le site de liaison. 1) CIF encode des contacts en considérant toute distance inter-atomique protéine/ligand inférieure à la somme des rayons de van der Waals entre deux atomes. 2) HIF définit l'ensemble des liaisons hydrogène, avec pour seuil de distance donneurH/accepteurH une valeur de 3 Å. 3) CHIF est une combinaison de CIF et HIF. Il a testé ces méthodes sur une analyse de poses d'arrimage moléculaire réalisée avec le logiciel GOLD sur un jeu de molécules composé de 490 leurres et 10 molécules de haute affinité pour le récepteur alpha des œstrogènes. En premier lieu, les trois méthodes récupèrent plus de molécules actives que la fonction de score native du logiciel GOLD. Cette cible possédant un site de liaison assez hydrophobe, le nombre d'actifs récupérés est plus faible dans le cas de HIF que de CIF ou de CHIF. A la vue de ces derniers, il est clair qu'il est important de prendre en compte tant les contacts apolaires que les interactions polaires afin de pouvoir gérer l'ensemble des combinaisons d'interactions au site de liaison existant.

Les méthodes positionnant les interactions sont sensiblement identiques. En effet, que ce soit à travers l'utilisation des atomes ou des résidus, toutes codent plus ou moins de la même façon les interactions. Ainsi, mis à part les IFP, ce sont essentiellement les liaisons hydrogène et les contacts apolaires qui sont pris en compte. Les informations concernant les atomes en interactions sont définies de façon plus ou moins floues afin d'induire une certaine flexibilité dans les définitions. Leurs applications sont ainsi diverses et variées, allant de l'analyse des poses d'arrimage moléculaire, au profilage pharmacologique de ligands, en passant par de l'analyse de relations structure/activité.

3.3.3 BASEE SUR LA PROTEINE ET LE LIGAND

APIF[81] et Pharm-IF[82] sont les seules méthodes ne comparant pas des modes d'interactions par rapport aux atomes du ligand ou de la protéine mais par rapport aux distances entre ces interactions. Bien que Pharm-IF positionne l'information au niveau des atomes de ligand, cette méthode a été placée dans la partie complexe protéine/ligand car la transformation des interactions en chaîne d'entiers diffère grandement de ce qui a été précédemment décrit. Toutes deux utilisent les fonctions intégrées à MOE pour détecter les interactions non-covalentes. Cependant, APIF se limite aux interactions hydrophobes et liaisons hydrogène tandis que Pharm-IF intègre aussi les interactions ioniques. Les combinaisons de paires d'interactions sont ensuite analysées et la distance est calculée entre les deux. Dans le cas d'APIF, 6 combinaisons sont possibles (**Figure 22**)

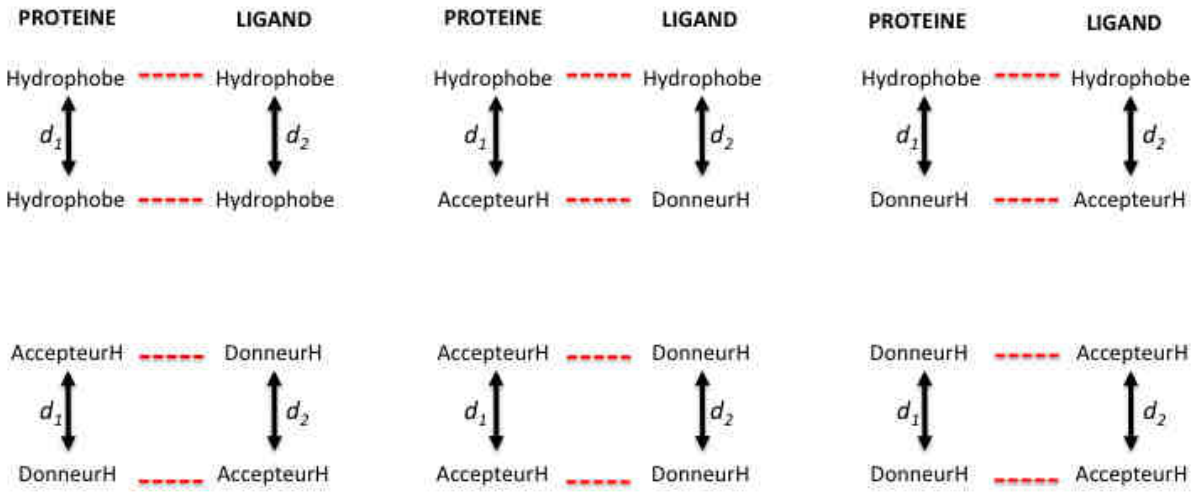


Figure 22 - Représentation des 6 combinaisons de paires d'interaction prises en compte dans APIF. Trois types d'interactions sont détectés : Hydrophobe, donneur de liaison hydrogène (donneurH) et accepteur de liaison hydrogène (accepteurH). Les traits rouges correspondent aux interactions non covalentes. Les distances d1 et d2 correspondent respectivement, pour une paire donnée, à la distance entre les deux atomes de protéines et les deux atomes de ligand

Chaque paire et leurs distances associées sont ensuite converties sous la forme d'une chaîne binaire de 294 caractères. Dans le cadre de Pharm-IF, les distances associées aux paires d'interactions sont discrétisées sous la forme de nombres flottants :

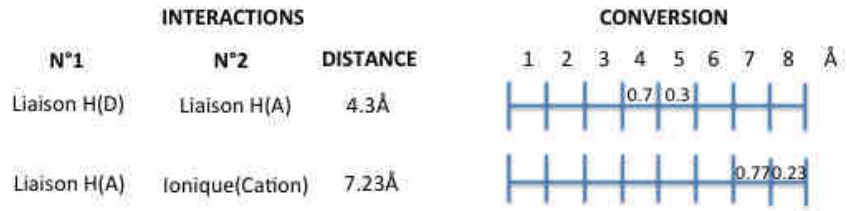


Figure 23 - Représentation schématique de la conversion de deux paires d'interaction sous la forme de distances discrétisées.

Comme nous pouvons le voir **Figure 23**, deux liaisons hydrogène sont séparées par une distance de 4.3 Å. La case dans la chaîne de nombres flottants correspondant à la paire d'interactions et à la distance de 4 Å sera incrémentée d'une valeur de 0.7, tandis que celle définie pour une distance de 5 Å sera incrémentée d'une valeur de 0.3. Cette représentation des distances permet de faire face aux problèmes d'effets de bords dus à la discrétisation de l'information. Pharm-IF et APIF ont montré des gains en termes de facteurs d'enrichissement sur diverses cibles par rapport à respectivement PLIF et CHIF.

Il est important de souligner que dans ces deux programmes, les interactions ne sont pas comparées par rapport à leurs positions absolues au sein du site actif ou du ligand mais par rapport à leurs positions relatives. En analysant l'importance de chaque paire, Sato et Yokohama se sont aperçus que les interactions proches en terme de distance sont plus importantes que des interactions plus éloignées. A titre d'exemple, la paire de liaisons hydrogène avec dans un cas un donneurH et dans un autre un accepteurH à moins de 3 Å revient très souvent dans le cadre de l'anhydrase carbonique 2, la tyrosine kinase SRC et la Cathepsine K. Ce type de paires peut, dans le cas d'une géométrie convenable, réaliser des interactions intra-moléculaires et ainsi être une composante non négligeable des effets de désolvatation.

4. CONCLUSION

De nombreuses interactions non-covalentes existent entre une protéine et un ligand et sont la base de la reconnaissance moléculaire. Le panel d'interactions que nous offre la nature est large et complexe. Certaines interactions, telles que les liaisons hydrogène, les interactions ioniques ou aromatiques sont bien décrites, avec une géométrie bien précise et suffisamment importantes pour ne pas être négligées. Les contacts apolaires plus difficiles à évaluer, particulièrement à un niveau énergétique en partie à cause des effets de désolvatation. Enfin, d'autres interactions sont encore mal caractérisées, peu présentes, car peu étudiées dans les systèmes biologiques. Rarey et al. a proposé une hiérarchie simple des modèles d'interaction.[95] Plus on augmente le niveau, plus les contraintes sont importantes et plus l'universalité diminue alors au profit de la spécificité. Cela résume bien la problématique des interactions non-covalentes où il est nécessaire de faire un compromis entre simplicité, interprétation et précision.

Le panel des outils codant ces interactions est impressionnant et couvre un large éventail méthodologique. De l'utilisation de descripteurs tridimensionnels géométriques ou énergétiques à la conversion en empreintes, chacune de ces méthodes propose une vision différente de ces interactions. Divers points ressortent cependant de cette analyse. En premier lieu, les interactions occasionnelles sont rarement prises en considération. En second lieu, l'utilisation de poids pour discriminer ces interactions semble être d'une extrême importance afin de compenser la forte présence de contacts apolaires par rapport à des interactions polaires, tel qu'on a pu l'observer avec IASF. Enfin, l'utilisation de propriétés potentielles multiples semble s'être développée au cours des dernières années, augmentant la combinatoire mais aussi la possibilité d'affiner la sélection de molécules actives.

5. BIBLIOGRAPHIE

(1) Scripture, C. D.; Figg, W. D. *Nat. Rev. Cancer* **2006**, *6*, 546–58.

(2) Lipinski, C. a.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. *Adv. Drug Deliv. Rev.* **2001**, *46*, 3–26.

(3) Lipinski, C. a. *Drug Discov. Today: Technologies* **2004**, *1*, 337–341.

(4) Pardridge, W. M. *Adv. Drug Deliv. Rev.* **1995**, *15*, 5–36.

(5) Brown, D.; Superti-Furga, G. *Drug Discov. Today* **2003**, *8*, 1067–77.

(6) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucl. Acids Res.* **2000**, *28*, 235–42.

(7) Meslamani, J.; Rognan, D.; Kellenberger, E. *Bioinformatics* **2011**, *27*, 1324–6.

(8) Nisius, B.; Sha, F.; Gohlke, H. *J. Biotech.* **2012**, *159*, 123–34.

(9) Laskowski, R. a; Luscombe, N. M.; Swindells, M. B.; Thornton, J. M. *Prot. Sci.* **1996**, *5*, 2438–52.

(10) Nayal, M.; Honig, B. *Proteins* **2006**, *63*, 892–906.

(11) An, J.; Totrov, M.; Abagyan, R. *Mol. Cell. Proteomics: MCP* **2005**, *4*, 752–61.

(12) Liang, J.; Edelsbrunner, H.; Woodward, C. *Prot. Sci.* **1998**, *7*, 1884–97.

(13) Kahraman, A.; Morris, R. J.; Laskowski, R. a; Thornton, J. M. *J. Mol. Biol.* **2007**, *368*, 283–301.

(14) Young, L.; Jernigan, R. L.; Covell, D. G. *Prot. Sci.* **1994**, *3*, 717–29.

(15) Pérot, S.; Sperandio, O.; Miteva, M. a.; Camproux, A.-C.; Villoutreix, B. O. *Drug Discov. Today* **2010**, *15*, 656–67.

(16) Hajduk, P. J.; Huth, J. R.; Fesik, S. W. *J. Med. Chem.* **2005**, *48*, 2518–25.

(17) Schmidtke, P.; Barril, X. *J. Med. Chem.* **2010**, *53*, 5858–67.

(18) Cheng, A. C.; Coleman, R. G.; Smyth, K. T.; Cao, Q.; Soulard, P.; Caffrey, D. R.; Salzberg, A. C.; Huang, E. S. *Nat. Biotechnol.* **2007**, *25*, 71–5.

(19) Halgren, T. a *J. Chem. Inf. Model.* **2009**, *49*, 377–89.

(20) Summa, V.; Petrocchi, A.; Bonelli, F.; Crescenzi, B.; Donghi, M.; Ferrara, M.; Fiore, F.; Gardelli, C.; Gonzalez Paz, O.; Hazuda, D. J.; Jones, P.; Kinzel, O.; Laufer, R.; Monteagudo, E.; Muraglia, E.; Nizi, E.; Orvieto, F.; Pace, P.; Pescatore, G.; Scarpelli, R.; Stillmock, K.; Witmer, M. V; Rowley, M. *J. Med. Chem.* **2008**, *51*, 5843–55.

(21) Schames, J. R.; Henchman, R. H.; Siegel, J. S.; Sotriffer, C. a; Ni, H.; McCammon, J. A. *J. Med. Chem.* **2004**, *47*, 1879–81.

(22) Luque, I.; Freire, E. *Proteins* **2000**, *Suppl 4*, 63–71.

(23) Hilser, V. J.; Freire, E. *J. Mol. Biol.* **1996**, *262*, 756–72.

(24) Allen, F. H. *Acta Cryst. B, Structural science* **2002**, *58*, 380–8.

(25) Bruno, I. J.; Cole, J. C.; Lommerse, J. P. M.; Rowland, R. S.; Taylor, R.; Verdonk, M. L. *J. Comput. Aided Mol.* **1997**, *11*, 525–37.

(26) Taylor, R.; Kennard, O. *Acc. Chem. Res.* **1984**, *17*, 320–326.

(27) Taylor, R.; Kennard, O.; Versichel, W. *Acta Cryst. B Structural Science* **1984**, *40*, 280–288.

(28) Panigrahi, S. K.; Desiraju, G. R. *Proteins* **2007**, *67*, 128–141.

(29) Murray-Rust, P.; Glusker, J. P. *JACS* **1984**, *106*, 1018–1025.

(30) Carosati, E.; Sciabola, S.; Cruciani, G. *J. Med. Chem.* **2004**, *47*, 5114–25.

(31) Dunitz, J. D.; Taylor, R. *Chem. Eur. J.***1997**, *3*, 89–98.

(32) Harding, M. M. *Acta Cryst.. D, Bio. Cryst* **2001**, *57*, 401–11.

(33) Imai, Y. N.; Inoue, Y.; Yamamoto, Y. *J. Med. Chem.* **2007**, *50*, 1189–96.

(34) Hunter, C. a.; Lawson, K. R.; Perkins, J.; Urch, C. J. *Perkin Trans. 2* **2001**, 651–669.

(35) Meyer, E. A.; Castellano, R. K.; Diederich, F. *Angew. Chem. Int. Ed.* **2003**, *42*, 1210–50.

(36) Gallivan, J. P.; Dougherty, D. A. *P.N.A.S.* **1999**, *96*, 9459–9464.

(37) Brandl, M.; Weiss, M. S.; Jabs, A.; Sühnel, J.; Hilgenfeld, R. *J. Mol. Biol.* **2001**, *307*, 357–77.

(38) Tsuzuki, S.; Honda, K.; Uchimaru, T.; Mikami, M.; Tanabe, K. *JACS* **2000**, *122*, 3746–3753.

(39) Politzer, P.; Lane, P.; Concha, M. C.; Ma, Y.; Murray, J. S. *J. Mol. Model.* **2007**, *13*, 305–11.

(40) Auffinger, P.; Hays, F. A.; Westhof, E.; Ho, P. S. *P.N.A.S.* **2004**, *101*, 16789–94.

(41) Bissantz, C.; Kuhn, B.; Stahl, M. *J. Med. Chem.* **2010**, *53*, 5061–84.

(42) Sarkhel, S.; Desiraju, G. R. *Proteins* **2004**, *54*, 247–59.

(43) Jadhav, P. K.; Woerner, F. J.; Lam, P. Y.; Hodge, C. N.; Eyermann, C. J.; Man, H. W.; Daneker, W. F.; Bacheler, L. T.; Rayner, M. M.; Meek, J. L.; Erickson-Viitanen, S.; Jackson, D. a; Calabrese, J. C.; Schadt, M.; Chang, C. H. *J. Med. Chem.* **1998**, *41*, 1446–55.

(44) Kubinyi, H. *Comp. Appl. Pharm. Res. Dev.* **2006**, 377–424. ISBN 0 471737798

(45) Taylor, R. D.; Jewsbury, P. J.; Essex, J. W. *J. Comput. Aided Mol.* **2002**, *16*, 151–66.

(46) Halperin, I.; Ma, B.; Wolfson, H.; Nussinov, R. *Proteins* **2002**, *47*, 409–43.

(47) Cole, J. C.; Murray, C. W.; Nissink, J. W. M.; Taylor, R. D.; Taylor, R. *Proteins* **2005**, *60*, 325–32.

(48) Leach, A. R.; Gillet, V. J.; Lewis, R. a; Taylor, R. *J. Med. Chem.* **2010**, *53*, 539–58.

(49) Yang, S.-Y. *Drug Discov. Today* **2010**, *15*, 444–50.

(50) Wermuth, C.; Ganellin, C.; Lindberg, P.; Mitscher, L. *Pure Appl. Chem.* **1998**, *70*, 1129–1143.

(51) Goodford, P. J. *J. Med. Chem.* **1985**, *28*, 849–857.

(52) Wade, R. C.; Clark, K. J.; Goodford, P. J. *J. Med. Chem.* **1993**, *36*, 140–7.

(53) Wade, R. C.; Goodford, P. J. *J. Med. Chem.* **1993**, *36*, 148–56.

(54) Friesner, R. a; Banks, J. L.; Murphy, R. B.; Halgren, T. a; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. *J. Med. Chem.* **2004**, *47*, 1739–49.

(55) Oshiro, C. M.; Kuntz, I. D.; Dixon, J. S. *J. Comput. Aided Mol.* **1995**, *9*, 113–30.

(56) Welch, W.; Ruppert, J.; Jain, A. N. *Chemistry and Biology* **1996**, *3*, 449–462.

(57) Jain, A. N. *J. Med. Chem.* **2003**, *46*, 499–511.

(58) Baroni, M.; Cruciani, G.; Sciabola, S.; Perruccio, F.; Mason, J. S. *J. Chem. Inf. Model.* **2007**, *47*, 279–94.

(59) Cross, S.; Baroni, M.; Goracci, L.; Cruciani, G. *J. Chem. Inf. Model.* **2012**, *52*, 2587–98.

(60) Cross, S.; Ortuso, F.; Baroni, M.; Costa, G.; Distinto, S.; Moraca, F.; Alcaro, S.; Cruciani, G. *J. Chem. Inf. Model.* **2012**, *52*, 2599–608.

(61) Böhm, H.-J. *J. Comput. Aided Mol.* **1992**, *6*, 61–78.

(62) Böhm, H.-J. *J. Comput. Aided Mol.* **1992**, *6*, 593–606.

(63) McGann, M. R.; Almond, H. R.; Nicholls, A.; Grant, J. A.; Brown, F. K. *Biopolymers* **2003**, *68*, 76–90.

(64) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. *Proteins* **2003**, *52*, 609–23.

(65) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. *J. Mol. Biol.* **1996**, *261*, 470–89.

(66) Wolber, G.; Langer, T. *J. Chem. Inf. Model.* **2005**, *45*, 160–9.

(67) Kurogi, Y.; Guner, O. *Curr. Med. Chem.* **2001**, *8*, 1035–1055.

(68) Dixon, S. L.; Smondyrev, A. M.; Knoll, E. H.; Rao, S. N.; Shaw, D. E.; Friesner, R. a *J. Comput. Aided Mol.* **2006**, *20*, 647–71.

(69) Tan, L.; Lounkine, E.; Bajorath, J. *J. Chem. Inf. Model.* **2008**, *48*, 2308–12.

(70) Tan, L.; Bajorath, J. *Chem. Biol. Drug Des.* **2009**, *74*, 25–32.

(71) Batista, J.; Tan, L.; Bajorath, J. *J. Chem. Inf. Model.* **2010**, *50*, 79–86.

(72) Crisman, T. J.; Sisay, M. T.; Bajorath, J. *J. Chem. Inf. Model.* **2008**, *48*, 1955–64.

(73) Deng, Z.; Chuaqui, C.; Singh, J. *J. Med. Chem.* **2004**, *47*, 337–44.

(74) Deng, Z.; Chuaqui, C.; Singh, J. *J. Med. Chem.* **2006**, *49*, 490–500.

(75) Nandigam, R. K.; Kim, S.; Singh, J.; Chuaqui, C. *J. Chem. Inf. Model.* **2009**, *49*, 1185–92.

(76) Chuaqui, C.; Deng, Z.; Singh, J. *J. Med. Chem.* **2005**, *48*, 121–33.

(77) Marcou, G.; Rognan, D. *J. Chem. Inf. Model.* **2007**, *47*, 195–207.

(78) Kroemer, R. T.; Vulpetti, A.; McDonald, J. J.; Rohrer, D. C.; Trosset, J.-Y.; Giordanetto, F.; Cotesta, S.; McMartin, C.; Kihlén, M.; Stouten, P. F. W. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 871–81.

(79) Kelly, M. D.; Mancera, R. L. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1942–51.

(80) Mpamhanga, C. P.; Chen, B.; McLay, I. M.; Willett, P. *J. Chem. Inf. Model.* **2006**, *46*, 686–98.

(81) Pérez-Nueno, V. I.; Rabal, O.; Borrell, J. I.; Teixidó, J. *J. Chem. Inf. Model.* **2009**, *49*, 1245–60.

(82) Sato, T.; Honma, T.; Yokoyama, S. *J. Chem. Inf. Model.* **2010**, *50*, 170–85.

(83) Chemical Computing Group MOE **2011**.

(84) Barillari, C.; Marcou, G.; Rognan, D. *J. Chem. Inf. Model.* **2008**, *48*, 1396–410.

(85) Jones, G.; Willett, P.; Glen, R.; Leach, A. R.; Taylor, R. *J. Mol. Biol.* **1997**, *267*, 727–48.

(86) Greene, J.; Kahn, S.; Savoj, H.; Sprague, P.; Teig, S. *J. Chem. Inf. Model.* **1994**, *34*, 1297–1308.

(87) Sanders, M. P. a; Barbosa, A. J. M.; Zarzycka, B.; Nicolaes, G. a F.; Klomp, J. P. G.; de Vlieg, J.; Del Rio, A. *J. Chem. Inf. Model.* **2012**, *52*, 1607–20.

(88) Grant, J. A.; Pickup, B. T. *J. Phys. Chem.* **1995**, *99*, 3503–3510.

(89) Grant, J. A.; Gallardo, M. a.; Pickup, B. T. *J. Comput. Chem.* **1996**, *17*, 1653–1666.

(90) Nicholls, A.; Grant, J. A. *J. Comput. Aided Mol.* **2005**, *19*, 661–86.

(91) Willett, P.; Barnard, J. M.; Downs, G. M. *J. Chem. Inf. Model.* **1998**, *38*, 983–996.

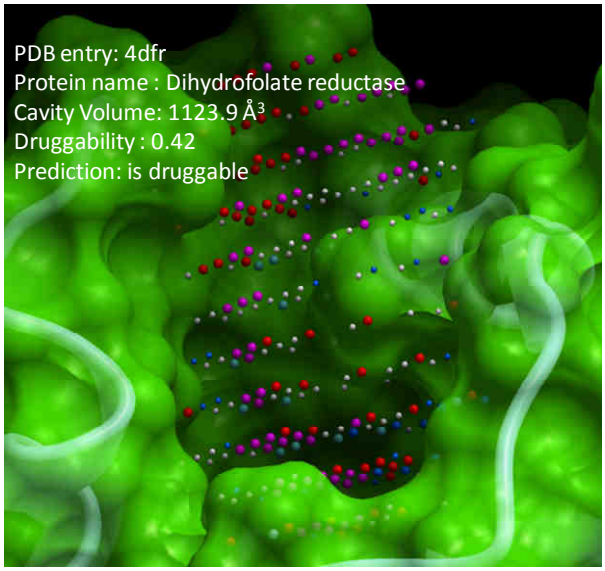
(92) MDL Information Systems, I. MACCS Keys.

(93) Rogers, D.; Hahn, M. *J. Chem. Inf. Model.* **2010**, *50*, 742–54.

(94) Tan, L.; Vogt, M.; Bajorath, J. *Chem. Biol. Drug Des.* **2009**, *74*, 449–56.

(95) Rarey, M.; Kramer, B.; Lengauer, T. *Bioinformatics* **1999**, *15*, 243–50.

Chapitre 2
Description et comparaison de cavités protéiques



Ce chapitre a fait l'objet d'une publication :

Comparison and Druggability Prediction of Protein-Ligand Binding Sites from Pharmacophore-Annotated Cavity Shapes

Jérémy DESAPHY, Karima AZDIMOUSA, Esther KELLENBERGER, Didier ROGNAN

Journal of Chemical Information and Modelling, 2012,52 (8), pp 2287-2299

1. CONTEXTE

L'agencement et la composition en acides aminés d'un site de liaison sont caractéristiques de la fonction de la protéine et permettent l'accueil de ligands. Détecter et caractériser ce site est par conséquent un enjeu majeur dans la recherche de nouvelles entités moléculaires. La comparaison de ces sites permet d'attribuer une fonction à une protéine orpheline, et d'observer les conservations locales ou globales de la structure et des propriétés physico-chimiques au sein d'une famille de protéines ou entre familles différentes. De plus, la caractérisation de ces sites est cruciale pour mieux comprendre les phénomènes de liaison entre une protéine et un ligand, puisque le mode d'interaction de ce dernier est intrinsèquement lié aux possibilités d'ancrage dans le site. Par conséquent, le ligand doit avoir la forme adéquate pour accéder et entrer dans la poche, et doit montrer une capacité à interagir de façon favorable avec le site aux travers d'interactions non-covalentes.

Aucune règle n'est à l'heure actuelle clairement définie pour délimiter les bordures d'une poche car les caractéristiques géométriques et physico-chimiques de celles-ci sont très dépendantes du contexte fonctionnel et topologique de la protéine. Ainsi, certaines poches seront sphériques et enfouies tandis que d'autres seront plus allongées, grandes, et accessibles au solvant.[1] Les sites de liaisons répondent à ces critères, et l'analyse de ces derniers a par ailleurs été longuement étudiée. Dans le cas des enzymes par exemple, la poche du ligand est généralement la plus grande des poches de la protéine.[2] Cependant, pour les ribonucléases, celle-ci est sphérique, relativement petite, et très enfouie. Par ailleurs, toutes sont en général majoritairement hydrophobes, comme l'a prouvé dernièrement Villoutreix et son équipe à travers l'analyse de 564 complexes protéine/ligand,[3] cette conclusion confortant de nombreuses autres études sur le sujet.[4–6]

En dépit de ce large éventail possible de formes, il existe une conservation de certains acides aminés,[7] tel que la triade catalytique des protéases à sérine,[8] et qui sont les garants de la fonction. Cela implique une conservation structurale, plus ou moins locale nous permettant ainsi de les différencier et de caractériser la fonction d'une

protéine à partir de son seul site de liaison. Dans ces conditions, comparer deux cavités permet ainsi d'observer si une famille se dégage par sa similarité et d'attribuer une fonction.

De très nombreuses méthodes existent pour détecter toutes les poches et cavités d'une protéine. D'excellentes revues récapitulent et analysent l'ensemble de ces méthodes.[3,9,10] Dans l'état actuel, elles réussissent à trouver avec près de 70% de rappel la cavité du ligand connu. Cependant, peu d'algorithmes leur attribuent une valeur de droguabilité, c'est à dire la probabilité qu'un ligand droguable puisse s'y lier.[5,11,12] Bien que cette définition soit très large et ne reflète pas toute sa complexité, elle reste cependant essentielle pour réduire l'attrition de molécules passant les phases cliniques.

L'objectif de ce chapitre n'est pas de développer un nouvel outil de détection des cavités et poches d'une protéine mais de caractériser celles connues pour accueillir une molécule de faible poids moléculaire. VolSite est une méthode d'analyse et de caractérisation des cavités d'une protéine, basée sur la discrétisation de la cavité sous la forme d'une grille tri-dimentionnelle. Chaque point se voit alors attribué une propriété physico-chimique complémentaire à l'atome de protéine le plus proche. L'ensemble de ces points et de leurs propriétés permet d'obtenir une valeur de droguabilité, calculé à partir d'un modèle de classification basé sur un jeu d'entrainement de 76 entrées. La comparaison de deux cavités est réalisée à travers Shaper, un logiciel reposant sur l'utilisation de Gaussiennes pour aligner et calculer un score.

2. INTRODUCTION

Although the pace of protein structure determination is by far inferior to that of protein sequencing, outstanding efforts of structural genomics consortia[13,14] and methodological advances in structural biology[14–16] contribute to considerably change our understanding of the structural proteome. For example, the most important family of drug targets (G Protein-coupled receptors) has long been described by a single representative[17] in the Protein Data Bank (PDB),[18] but has been supplemented by 14 novel receptors and 30 new receptor-ligand complexes in the last 4 years.[19,20] A comprehensive coverage of UniProt targets by the PDB is therefore anticipated in ca. 15 years.[21] Among the applications that will benefit from a better structural coverage of biological space is the prediction of off-targets and resulting side effects for known drug candidates.[22,23] Hence, if one assumes that similar protein pockets accommodate similar ligands,[24] it is possible to predict ligand cross-reactivity to targets sharing similar ligand-binding sites.[25–28] To achieve this goal, a pocket detection algorithm must first be precise enough to focus on druggable binding sites only, quantitative comparison should then remain fuzzy enough to accommodate ligand-induced structural changes.[29] A key issue is the ability to detect a three-dimensional (3-D) similarity for binding sites of unrelated proteins in absence of fold conservation. In most cases, cavity comparison tools rely on the prior 3-D alignment of a set of representative protein atoms.[29] These methods are relatively slow (1-10 comparisons/min) and highly dependent on atomic coordinates but are easy to interpret and to couple with other computational methods like ligand docking.[22] A wrong 3-D alignment of two binding sites, whatever the reason, will however underestimate their true similarity.[30] Alignment-independent methods,[30–34] amenable to a very high throughput (up to 1 000 comparisons/s) have thus been recently described to quantify pocket similarities but still suffers from a lack of interpretability since 3-D information is often loss upon converting cavity properties into simple 1-D fingerprints.

We herewith introduce a novel cavity description and comparison method combining the advantages of alignment-dependent methods with the speed of alignment-independent methods. In contrast to many existing tools, the protein-ligand

binding pocket is not represented by either protein atoms or surface points but by regularly-spaced pharmacophoric grid points defining its inverse image from a protein-ligand interaction point of view. We next used a shape-based alignment method[35] using a smooth Gaussian function approximating molecular volumes, to align cavities by optimizing the volume overlap of their pharmacophore-annotated shapes. The method is simple, fast, and particularly efficient in detecting binding site similarity in absence of sequence and fold conservation. It can be used for three main applications: (i) infer the function of a protein by measuring the similarity of its known or potential ligand-binding pockets to a collection of functionally annotated binding sites, (ii) classify targets according to the similarity of their binding sites, (iii) predict the structural druggability (ligandability) of a binding site from the properties of its pharmacophoric shape.

3. METHODS

3.1 PHARMACOPHORIC ANNOTATION OF CAVITY GRID POINTS (VOLSITE)

Starting from atomic coordinates of a protein-ligand complex, a three-dimensional cube of 20 Å-edge is centered on the center of mass of the bound ligand and filled with a 1.5 Å-resolution grid defining 2 370 cells of 3.375 $Å^3$ volume each (**Figure 1**). To each cell is associated a grid point and a property at its center. If the corresponding cell comprises a protein atom or if its center is less than 2.5 Å away from any protein atom, the grid point is given the 'IN' property. Any other point is then checked for buriedness by generating, from its coordinates, a set of 120 regularly-spaced rays of 8 Å length. If the number of rays intersecting a 'IN' cell (N_{ri}) is smaller than a user-defined threshold (by default 40), the corresponding point is considered to be outside the enclosing cavity and is assigned the 'OUT' property. Remaining points are considered to encompass the cavity and checked for direct neighborhood with other cavity points. If isolated (less than 3 neighbors in adjacent cells), cavity points are deleted. Site points closer than 4.0 Å to a protein atom are assigned one of seven possible pharmacophoric properties (H-bond acceptor, H-bond donor, H-bond acceptor and donor, negative ionizable, positive ionizable, hydrophobic, aromatic) by complementary to that of the closest protein atom using standard interaction rules[36] (**Table 1**). Points with no protein atoms within 4 Å are assigned a null property. The pharmacophoric properties of protein atoms are detected on-the-fly from their names (PDB input) or atom types (MOL2 input) thus enabling in the latter case to consider additional molecules (ions, cofactors, water, prosthetic groups, nucleic acid) as part of the protein. Once every point has been assigned a pharmacophoric property, 5 sets of cavity points are defined with respect to their largest distance (4 Å, 6 Å, 8 Å, 12 Å, any) to any protein-bound ligand heavy atom. From here on, we will refer these binding sites of increasing sizes to cavities truncated at 4, 6, 8 and 12 Å, respectively. The full cavity is defined when no truncation is applied.

Since every cell has a fixed and unique volume (3.375 $Å^3$), the total number of pharmacophore-annotated cells approximates the global cavity volume. In absence of co-crystallized ligands (apo-proteins), any set of atomic coordinates (e.g. center of a

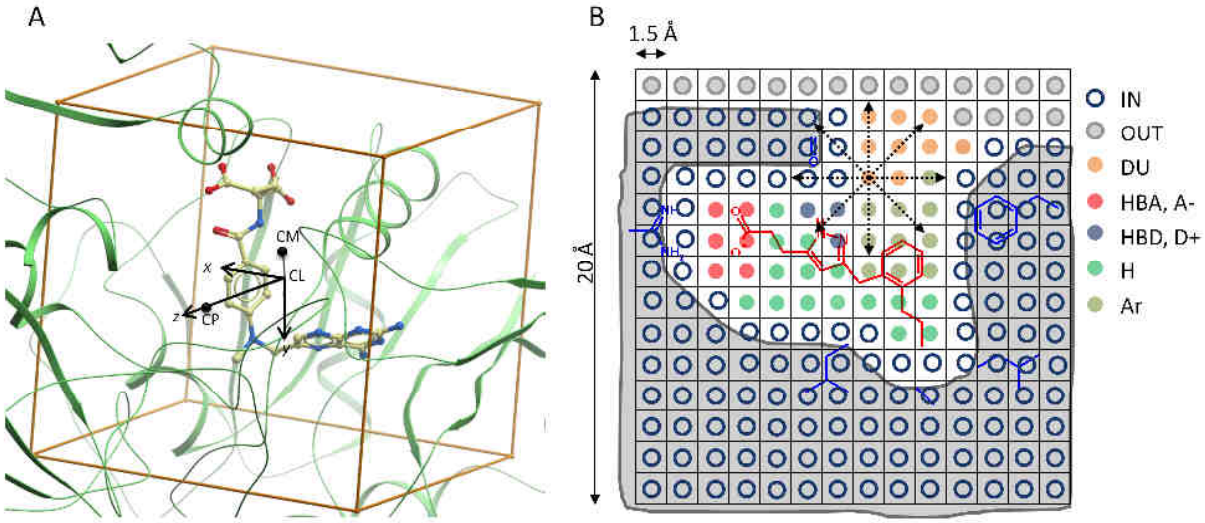


Figure 1 - Orientation and definition of the grid lattice from atomic coordinates of a protein-ligand complex. A) From the center of mass of the ligand (CL, grid center), and the center of mass of the protein (CP), coordinates of the reference inertia point CM are computed — N_l : number of ligand heavy atoms, w_i : mass of atom i, a_i : coordinates of atom i). The three main axes defining the grid lattice are deduced computing first the vector normal to and (x axis), then the vector normal to and (y axis), and last the vector normal to and (z axis). B) A 3-D lattice of size 20 Å and resolution 1.5 Å is centered on the ligand (red sticks) center of mass. A site point is placed at the center of each of the 2 730 cells and assigned a property according to its location with respect to the protein (gray solid surface) active site residues (blue sticks) : IN (cell intersecting any protein atom or site point closer than 2.5 Å to any protein atom), OUT (site point outside the cavity), DU (site point inside the cavity but farther than 4.0 Å from any protein atom), HBA (site point inside the cavity close to a protein H-bond donor), HBD (site point inside the cavity close to a protein H-bond acceptor), D+ (site point inside the cavity close to a negative ionizable protein atom), A- (site point inside the cavity close to a positive ionizable protein atom), HYD (site point inside the cavity close to a protein aliphatic apolar atom), AR (site point inside the cavity close to a protein aromatic atom). Assigning site points inside or outside the cavity is decided with respect to the proportion of 8 Å-long rays (dotted arrows) projected from every point intersecting a 'IN' cell.

Table 1 - Cavity point properties and pharmacophore matching rules

Property	Name	Residue	Closest protein atom
Hydrophobic (HYD)	CA	Gly	Hydrophobic
Aromatic (AR)	CZ	Phe	Aromatic
Acceptor (HBA)	O	Ala	Donor
Donor (HBD)	N	Ala	Acceptor
Acceptor/Donor (HBAD)	OG	Ser	Acceptor/Donor
Positive ionizable (D+)	NZ	Lys	Negative ionizable
Negative ionizable (A-)	OD1	Asp	Positive ionizable
Null (DU)	DU	Cub	none

cavity detected by a third-party software) can be given as input to define the 3-D grid and generate cavity points. To enable their visualization by any software, standard protein atom names with corresponding pharmacophoric properties are given to each cavity point (**Table 1**).

3.2 DRUGGABILITY PREDICTION

The recently-described non redundant set of druggable and less druggable binding sites (NRDLD)[11] describing 113 cavities (71 druggable, 42 undruggable) was utilized for assessing the suitability of Volsite attributes to predict the structural druggability (or ligandability) from protein x-ray structures. The corresponding protein-ligand complexes were retrieved from the sc-PDB[37] or the Protein Data Bank;[18] cavity points were generated using standard parameters and no truncation to consider the entire cavity. The dataset was split, as originally proposed[11] into a training set of 76 entries (48 druggable, 28 undruggable) and a test set of 37 entries (23 druggable, 14 undruggable).

For each entry, 73 VolSite descriptors (**Supplementary Table 1**) were read as input values for a binary classification model using a support vector machine algorithm (SVM), as implemented in SVM^{light}.[38]

These descriptors encode the volume of the cavity as the total number of cavity points (Descriptor #1), the proportion (expressed in percent) of points having each of the eight pharmacophoric types (in other words, hydrophobocity, aromaticity and polarity; Descriptors #2-9) and the accessibility of every site point (Descriptors #10-73) expressed for each of the 8 pharmacophoric types, as the number of 'IN' cell-intersecting rays (N_{ri}) within 8 ranges (40-50, 50-60, 60-70, 80-90, 90-100, 100-110, 110-120).

Optimal values for gamma and c parameters were found after systematic variation of both parameters in a 5-fold crossvalidation procedure, using the rbf kernel, applied to the entire training set. We first iterate gamma from 0 to 1 using a 0.1 increment and c from 0 to 100 with a step of 1. The best average F-measure of the 5 folds gave gamma and c values around which a novel systematic variation was repeated using a tighter range and increment (10% of the previous ones) until no gain in the F-measure was observed. The model leading to the best F-measure (gamma=4.10^{-6} and c=100) was finally selected for external predictions.

The predicted druggability of all cavities in the external test set was compared to that reported for three state-of-the-art methods, DrugPred,[11] Fpocket[12] and SiteMap.[5] DrugPred and Fpocket values were directly taken from the literature.[11] In SiteMap v.2.2,[39] protein and ligand files were first extracted from the Protein DataBank and transformed in mol2 format in SybylX1.3.[40] After adding hydrogen atoms and manually optimizing intermolecular hydrogen bonds, protein and ligand coordinates were converted in mae file format using Maestro v8.5.[39] If the cavity contains a metal ion, the 'Protein preparation wizard' in Maestro is used to verify and manually correct whenever necessary atom types. The bound ligand is used as constraint to detect the cavity boundaries within 6 Å of any ligand atom.

3.3 BINDING SITE ALIGNMENT (SHAPER)

3.3.1 METHODS

The alignment tool (Shaper) relies on OEChem and OEShape toolkit.[41] The main advantage of these toolkits is the possibility to describe molecular shapes by a smooth Gaussian function and to align two molecules by optimizing the overlap of their corresponding volumes.[35,42,43] During the alignment, a reference set of cavity points is kept rigid while the set of cavity points to fit (fit object) undergoes rigid body rotations and translations. To speed-up calculations, the 'Grid' volume overlap method was chosen to represent the volume of the target molecule and all atom radii were set to that of carbon (1.7 Å). Once the best shape alignment has been achieved, it is scored by a 'Color Force Field' (a color being a pharmacophoric feature) similar to that used by the ligand matching tool ROCS[41] to account for pharmacophoric properties matching. The force field (**Supplementary Table 2**) consists in SMARTS patterns for 6 pharmacophoric properties (H, Ar, HBA, HBD, A^-, D^+) and 6 pattern matching rules (H to H, Ar to Ar, HBA to HBA, HBD to HBD, A^- to A^-, D^+ to D^+) to score the shape-based alignment by pharmacophoric similarity. Two pharmacophoric properties (HBAD, DU) were not considered for the color alignment since the first one is implicitly taken into account by either acceptor or donor SMART patterns, and because the latter was not found relevant in preliminary trials. Color matches were considered for cavity points up to 1.5 Å apart with a single weight for all matching rules (**Supplementary Table 2**).

3.3.2 SIMILARITY METRICS AND STATISTICAL EVALUATIONS

The similarity $S_{A,B}$ between cavities A (reference) and B (fit) was calculated by a Tversky index as follows:

$$S_{A,B} = \frac{O_{A,B}}{\alpha I_A + \beta I_B + O_{A,B}}$$

where $O_{A,B}$ is the overlap between colors of cavities A and B, and I non-overlapped colors of each entity A and B. By contrast to a Tanimoto index (α=β=1), the Tversky index gives more importance to either the reference or the fit object by assigning different weights (α≠β, α+β=1) to the self-color non-overlap I_A and I_B values. The metric is asymmetric and varies between 0 and 1. Preliminary trials indicated that the peak performance was reached with α =0.95 and β =0.05 (from hereon *RefTversky metric*). Classification models based on pairwise similarity values were assessed by computing the area under the receiver operating characteristic (ROC) curve,[44] the F-measure, the accuracy and the Matthew's correlation coefficient (MCC) as follows:

$$\text{Recall} = \frac{TP}{TP+FN} \qquad \text{Precision} = \frac{TP}{TP+FP} \qquad F - \text{Measure} = \frac{2*(\text{recall})(\text{precision})}{\text{precision}+\text{recall}}$$

$$\textbf{Accuracy} = \frac{\mathbf{TP+TN}}{\mathbf{TP+FN+FP+TN}} \qquad \textbf{MCC} = \frac{(\mathbf{TP}\times\mathbf{TN}) - (\mathbf{FP}\times\mathbf{FN})}{\sqrt{(\mathbf{TP+FP})(\mathbf{TP+FN})(\mathbf{TN+FP})(\mathbf{TN+FN})}}$$

where TP are true positives, FN false negatives, FP false positives and FN false negatives. The best similarity threshold is found by the maximum of the F-measure curve when the threshold was varied from 0 to 1 with an increment of 0.01.

3.3.3 PARAMETERS SELECTION AND OPTIMIZATION

To select the most appropriate parameters and their values, a previously reported dataset of 1 538 pairs of ligand-binding sites[30] was used for benchmarking. The dataset comprises 769 pairs of known similar sites (similar protein binding sites co-crystallized with two different ligands) and 769 pairs of sites randomly chosen among known dissimilar sites.[30] The systematic variation of 31 VolSite and Shaper parameters **(Supplementary Table 3**) yielded 16 384 different parameter sets, and as many lists of 1 538 pairs sorted by decreasing similarity score. The binary classification of pairs (similar/dissimilar) allowed the calculation of the area under the ROC curve for all lists. The standard Shaper parameters were thus defined as those maximizing the value of the area under the ROC curve.

3.3.4 VIRTUAL SCREENING OF THE SC-PDB DATABASE

The similarity of 5 952 sc-PDB (v.2009) binding sites to the inhibitor-binding site in bovine tryspin (PDB entry 1aq7) was computed from the corresponding cavity points with standard VolSite parameters. sc-PDB entries were classified by fold and substrate cleavage specificity in 5 groups according to the CATH protein structure classification[45] and the CutDB proteolytic event database.[46] The first group is composed of 271 serine endopeptidase entries sharing a trypsin-like fold and trypsin substrate cleavage specificity. The second group is composed of 17 other serine endopeptidase entries presenting a trypsin-like fold but substrate cleavage specificity different from that of trypsin. The third group is composed of 11 serine endopeptidase entries with a subtilisin-like fold. The fourth group is composed of 13 entries with a α/β hydrolase fold. The last class is composed by the 5 640 remaining scPDB entries. All entries were ranked by decreasing RefTversky similarity score and the rank list used to compute the area under the ROC curve for a binary classification model considering iteratively each of the group as positive instances.

A second virtual screening for similarity to a structurally different ligand-binding site (ATP-competitive inhibitor-binding site in Pim-1 kinase) was undertaken, using three reference sites of the same enzyme co-crystallized by three inhibitors of different sizes (PDB codes 1hys, 3cy3, 1yi4). In this second screen, 9 877 sc-PDB entries (v.2011) were classified in 4 groups according to their E.C. number (protein kinases, other kinases, other ATP/ADP-binding sites, other sc-PDB entries) as previously described.[30] All entries were ranked by decreasing RefTversky similarity score to each of the three references and the rank lists used to compute the area under the ROC curve for a binary classification model considering iteratively each of the group as positive instances.

3.3.5 DATASET OF PROMISCUOUS PROTEIN-LIGAND COMPLEXES

A dataset of promiscuous ligands was set-up by parsing the sc-PDB database (v.2010) for ligands co-crystallized with at least 2 different proteins, according to their sc-PDB name.[37] The sc-PDB name is derived from the UniProt recommended name, without indications for source organism, cellular location or maturation state. Remaining ligands were then manually filtered to remove oligomeric compounds (nucleic acids, peptides, oligosaccharides) and lipids. 247 promiscuous pharmacological ligands were finally identified, bound to 401 different proteins in 689 unique sc-PDB entries.

The corresponding protein sequences in fasta format were downloaded from the RCSB PDB.[18] The all- against-all comparison of sequences was performed for all the targets of each promiscuous ligand using default parameters of the Needle routine for global sequence alignment in the EMBOSS package.[47] Only protein chains involved in the ligand binding site were considered. If several comparisons were made for a given pair of proteins, only the highest sequence identity value was retained. A sequence identity above 30% is a good indicator of protein homology.[48] In the present analysis, we consider that an evolutionary link exists between two proteins aligned over more than 100 residues with a sequence identity above 25%.

The structures of complexes in PDB format were downloaded from the RCSB PDB. The all-against-all comparison of structures was performed for all the targets of each promiscuous ligand using default parameters of the CE program.[49] Only protein chains involved in the ligand binding site were considered. If several comparisons were made for a given pair of proteins, only the result with the highest Z-score was retained. A Z-score value lower than 3.7 indicates that the similarity is of low significance. A Z-score value higher than 4.5 denotes conservation of the overall fold. In the 3.7-4.0 range, CE Z-score values define a twilight zone. In the present analysis, we consider that two protein chains aligned with a Z-score higher than 4 share a similar fold.

The similarity of 1 070 binding site pairs sharing the same ligand was estimated with three different tools (SiteAlign,[50] FuzCav,[30] Shaper) using default parameters of each program, and previously-defined similarity thresholds[30,50] (SiteAlign: d1 <0.6 and d2 <0.2; FuzCav: score >0.16; Shaper: score >0.35).

3.3.6 ALL-AGAINST-ALL COMPARISON OF SC-PDB DRUGGABLE BINDING SITES

Cavity points were computed with standard VolSite parameters (no truncation) for the 9 877 binding sites of the current sc-PDB database (v.2011), and further compared with Shaper using a RefTversky metric to generate a full similarity matrix out of which 300 000 values were randomly chosen to select a statistically relevant sample for further analysis.

3.3.7 CLASSIFICATION OF GPCR X-RAY STRUCTURES

A set of 30 X-ray structures of G protein-coupled receptors co-crystallized with low molecular weight ligands was retrieved from the RCSB PDB. This collection comprises 14 different receptors (adenosine A2a, β1 and β2 adrenergic receptors, chemokine CXCR4, dopamine D3, histamine H1, muscarinic M2, muscarinic M3, delta opiate, kappa opiate, mu opiate, nociceptin, rhodopsin, and sphingosine-1 phosphate S1P1) sharing the same fold and co-crystallized with ligands exhibiting different functional effects (full agonists, inverse agonists, neutral antagonists; **Supplementary Table 4**). For each entry, all molecules were removed with the exception of protein and pharmacological ligand whose coordinates were separately saved in mol2 file format. Cavities around the bound ligand were computed with default VolSite settings. All pairwise cavity comparisons were done with default Shaper settings and similarities expressed by the RefTversky score. The full similarity table was filtered to remove pairs with a similarity below 0.35, and then imported into Cytoscape v2.8.2[51] to depict a network rendered with a force directed layout using similarity-dependent edge lengths.

4. RESULTS AND DISCUSSION

4.1 BINDING SITE DESCRIPTION

When applied to a typical protein-ligand cavity (e.g. dihydrofolate reductase), site points defined by VolSite encompass the bound ligand but also ligand-unexplored regions of the pocket and stops at soon as accessibility is too high (**Figure 2A**). The pharmacophoric mapping of cavity points is in overall agreement with the known inhibitor binding mode, cavity and ligand pharmacophoric features matching well (**Figure 2A**). Applying the VolSite algorithm to the entire sc-PDB dataset (9 877 entries) shows that hydrophobic points are the most frequent, whatever the ligand-binding site definition (**Figure 3**). Interestingly, the respective proportions of pharmacophoric types are independent on the binding site definition (**Figure 3A**). The distribution of entire cavity volumes, computed over all sc-PDB binding sites is centered at the value of 735 $Å^3$ (**Figure 3B),** inferior to that of 930 $Å^3$ previously reported for a much smaller subset of 99 drug-binding sites.[52]

We next confirmed that aligning cavity site points indeed leads to a correct alignment of the corresponding protein 3-D structures. We chose as prototypical example the ATP-binding site of three serine/threonine protein kinases: Pim-1, Rac-□, Akt-2. Their binding site-based alignment is problematic due to the protein flexibility (diverse conformations of the loop connecting the N- and C-terminal domains). Moreover, the presence of a few additional charged residues in one of the proteins (Akt-2) significantly modifies the pharmacophoric description of the sites, and was shown to cause the failure of alignment-free comparisons.[30] As illustrated in **Figure 2B,** Shaper proposes a very reliable alignment of the three structures with root-mean-square deviations to the Pim-1 structure below 1.5 Å (Cα atoms of the catalytic site only).

Many tools are available to detect protein cavities from 3-D atomic coordinates.[3] The herein presented approach should not be considered as a cavity detection utility since it relies on user-defined atomic coordinates (that of the bound ligand) and does not scan the entire target surface.

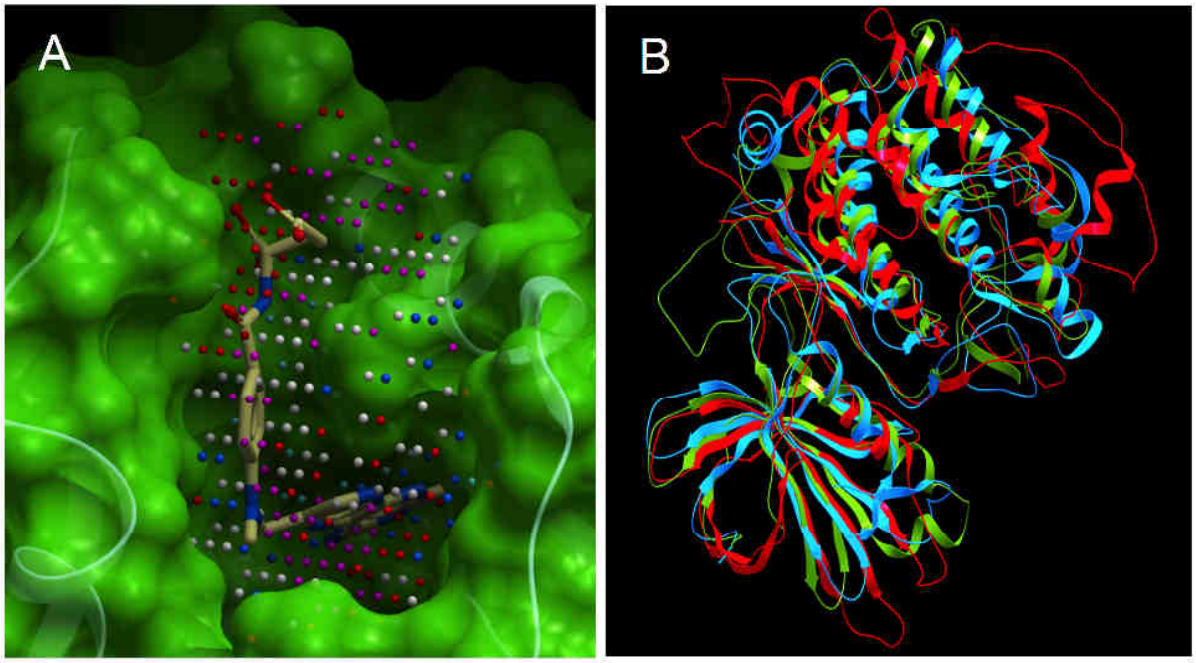


Figure 2 - Cavity description and alignment. A) Detection and pharmacophore annotation of all cavity points in the X-ray structure of *L. bacillus* dihydrofolate reductase (PDB code 4dfr). The cognate ligand (methotrexate, sticks) is shown in the binding site of the protein (green transparent surface). Cavity points are colored by pharmacophoric properties (H-bond acceptor and negative ionizable, red; H-bond donor and positive ionizable, blue; hydrophobe, white; aromatic, cyan; null, magenta). B) Site points-based alignment of three protein kinases (Pim-1, cyan, PDB entry 3cy3; Rac-β; red, PDB entry 1uv5; Akt-2, green, PDB entry 2jdr)

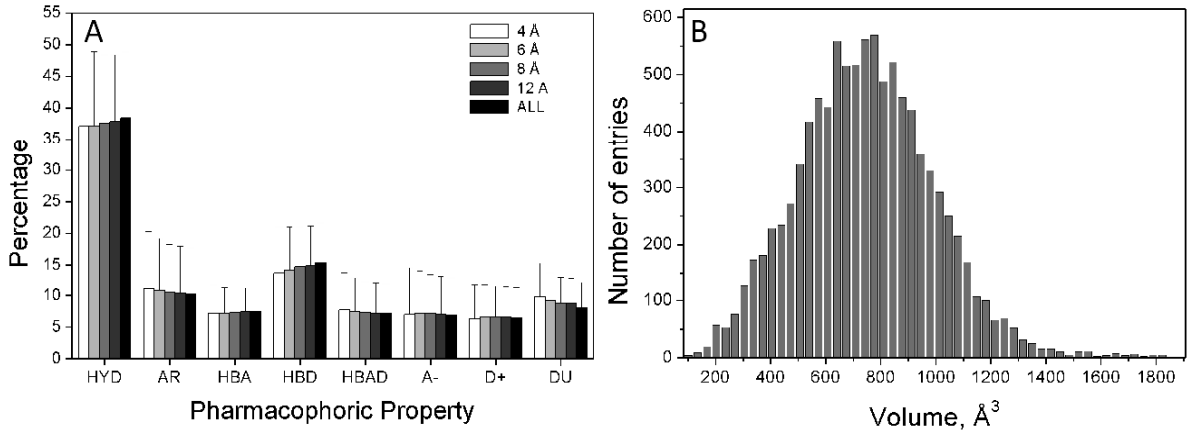


Figure 3 - Physicochemical properties of VolSite cavities. A) Distribution of pharmacophoric properties among cavity points truncated at various distances of the bound ligand, B) Distribution of the overall volume of sc-PDB cavities.

VolSite uses a standard grid-based cavity detection algorithm similar in spirit to the recently-published SiteMap method,[5] and presents the advantage of nicely delimiting pocket boundaries by projecting exit vectors from grid points and counting those intersecting protein atoms. Importantly, the computation of the cavity volume requires a reliable way of detecting its boundaries, notably at the interface to solvent. We herein apply a 'shrink-wrap' protocol by generating 120 uniformly spaced rays from accessible site points and retaining only those that are buried enough. Visual inspection of many cavities suggests a threshold of 40 for the Nri parameter (see Methods).

VolSite is thus robust in detecting numerous cavity shapes (grooves, clefts, needles). Our algorithm presents however some peculiarities with respect to existing methods. First, Shaper just focuses at the neighborhood of protein-bound ligands in the sc-PDB archive of druggable binding sites, and not at all possible binding cavities. Known pocket occupancy by druggable ligands is a restrictive but safe pocket selection filter. The method could be applied to *de novo* cavity detection but we wanted here to restrict our analysis to a well-defined repertoire of binding pockets with strong physicochemical and biological annotations. Second, site points are labeled with pharmacophoric properties complementary to that of the closest protein atom. This labeling procedure enables a chemically relevant description of the pocket as a 3-D-lattice of pharmacophoric features characteristic of potential ligands of this pocket. Since the default grid resolution (1.5 Å) is close to common bond lengths in organic compounds, pseudoatoms describing the protein cavity should feature true ligand atoms of these pockets. Therefore, typical ligand-based alignment methods may also be applied to the cavity points. Third, we store for each target 5 ligand-binding sites of increasing size. Hence, a ligand-binding site (immediate vicinity of the bound ligand) may or may not correspond to the entire protein cavity regarding their respective size and ligand buriedness. The method can thus be applied to compare binding sites (ligand-dependent object) but also to assess the structural druggability of the corresponding cavity (ligand-independent object).

4.2 PREDICTION OF STRUCTURAL DRUGGABILITY

Predicting the druggability of a given target from its three-dimensional structure is an intense field of research in order to reduce attrition rates in pharmaceutical discovery.[45] As druggability is by far more complex than the simple propensity of a particular protein cavity to accommodate high-affinity drug-like compounds, other terms like "ligandability"[53] or "bindability"[54] have recently been proposed since they better capture target property ranges (cavity volume, polarity and buriedness) known to be important for druggable targets.[4–6,12,54] Since those important properties are theoretically encoded in the herein-described cavity points, we investigated whether the present cavity descriptors might be suitable for predicting the druggability of cavities from their 3-D structures. A recently described training set (NRDLD) of 76 cavities (48 druggable, 28 undruggable) was retrieved from literature[11] and the distribution of site point properties was given as input for a support vector machine (SVM) classifier. The best 5-fold crossvalidated classification model achieves an accuracy of 0.80 and a Matthew's correlation coefficient (MCC) of 0.60. It was further challenged to predict the druggability of 37 novel cavities (23 druggable, 14 undruggable) still from the NRDLD dataset (**Table 2**). Our SVM classifier exhibits a significantly better performance than two state-of-the-art druggability prediction methods (SiteMap, Fpocket), and an accuracy similar to that of DrugPred,[11] one of the best method reported up to now.

Table 2 - Accuracy of 4 computational methods in predicting the structural druggability of 37 cavities (23 druggable, 14 undruggable) of known X-ray structure.[25]

Method	VolSite[a]	SiteMap[b]	Fpocket[c]	DrugPred[d]
Accuracy	0.89	0.65	0.73	0.89
MCC	0.77	0.24	0.39	0.77

[a] druggable if score > 0
[b] druggable if SiteScore > 0.8.[29]
[c] druggable if DG score > 0.50.[28]
[d] druggable if DrugPred score > 0.50.[25]

The current SVM model mispredicts only two druggable proteins of the dataset (DNA gyrase B, thymidine phosphorylase) as undruggable (**Table 3**). DNA gyrase is a clear false negative since our approach also failed in predicting other inhibitor-bound DNA gyrase PDB entries (e.g. 3ttz, 3g75) as druggable. We suspected that the main reason lies in the open and polar binding site of this enzyme, despite a deeply buried subsite. All methods used herein failed in predicting human thymidine phosphorylase as druggable, and this protein should probably be defined as weakly or not druggable at all, as recently suggested.[11] Conversely, only two false positives (glutamate racemase, dialkyglycine dicarboxylase) were observed out of the 14 non druggable entries of the test set (**Table 3**). This is less than Fpocket (5 false positives) and far less than SiteMap which tends to mispredict almost all nondruggable entries.

The good performance of our druggability prediction model can be explained by an extended dataset of druggable and undruggable cavities,[11] and by the use of a machine learning algorithm as predictor. Training a support vector machine on global and local pocket descriptors has also been reported to yield excellent results (90% of correct classification), although on a different training set.[55] A direct comparison of all druggability prediction methods is however difficult since they usually rely on different principles to define pocket boundaries (ligand-based method[5,11] or de novo pocket detection[55,56]), and are applied to different training and test sets. If initiatives to harmonize datasets of druggable and undruggable sites should be acknowledged,[11,12] a uniform definition of binding pockets (e.g. list of cavity lining residues including or not accessory molecules like ions, co-factors) would be desirable. Concluding, we can safely estimate that our approach is fast (ca. 10 sec for cavity detection and druggability prediction) and very competitive with the yet best available methods.[11,55]

Table 3 - Predicted druggability values for a test set of 37 entries. False predictions are underlined

PDB entry	Name	Method			
		DrugPred[a]	SiteMap[b]	Fpocket[c]	VolSite[d]
Druggable					
1e66	Acetylcholinesterase	0.81	1.14	0.75	0.80
1fk9	HIV reverse transcriptase	0.79	1.27	0.84	1.07
1kzn	DNA gyrase	0.81	1.04	0.75	-0.48
1lox	15-lipoxygenase	1.15	1.13	0.76	1.01
1oq5	Carbonic anhydrase II	0.77	1.00	0.10	0.48
1owe	Urokinase plasminogen activator	0.40	0.93	0.25	0.28
1pmn	C-Jun kinase	0.93	1.09	0.88	1.25
1pwm	Aldose reductase	0.94	0.97	0.86	0.67
1q41	Glycogen synthase kinase 3	0.55	1.09	0.46	1.15
1r55	ADAM33	0.69	0.89	0.08	0.07
1sqn	Progesterone receptor	1.11	1.28	0.95	1.99
1t46	c-Kit kinase	1.17	1.12	0.84	1.37
1unl	Cyclin-dependent kinase 5	0.56	1.06	0.12	0.47
1uou	Thymidine phosphorylase	0.28	n.d.[e]	0.40	-0.65
1xoz	Phosphodiesterase 5A	1.14	1.10	0.81	1.06
2aa2	Mineralocorticoid receptor	1.02	1.24	0.92	1.16
2cl5	Catechol-O-methyltransferase	0.82	1.19	0.70	1.48
2i1m	FMS kinase	0.74	1.10	0.82	0.70
3b68	Androgen receptor	1.13	1.29	0.95	2.09
3etr	Xanthine oxidase	0.85	1.13	0.67	1.01
3f0r	Histone deacetylase 8	0.89	1.12	0.59	0.96
3f1q	Dihydroorotate dihydrogenase	1.15	1.20	0.90	1.73
3ia4	Dihydrofolate reductase	0.79	1.07	0.65	0.63
Undruggable					
1ajs	Aspartate aminotransferase	0.49	1.14	0.60	-0.60
1b74	Glutamate racemase	0.41	1.05	0.56	0.60
1bls	Beta-lactamase	0.34	1.04	0.26	-0.01
1bmq	Interleukin-1beta-converting enzyme	0.38	0.79	0.01	-0.05
1ec9	D-glucarate dehydratase	-0.31	1.03	0.15	-1.28
1g98	Phosphoglucose isomerase	0.09	1.14	0.03	-0.57
1kc7	Pyruvate phosphate dikinase	0.01	0.86	0.01	-1.61
1m0n	Dialkylgycine decarboxylase	0.50	0.98	0.76	0.46
1mai	Phospholipase C	0.09	0.90	0.03	-1.93
1od8	Xylanase	0.06	0.79	0.05	-1.01
1px4	Beta-galactosidase	0.55	1.06	0.13	-0.03
1v16	a-keto acid dehydrogenase	0.41	1.08	0.02	-0.57
1wvc	CDP-D-glucose synthase	0.65	1.03	0.67	-0.52
3jdw	L-Arginine:glycine amidinotransferase	0.17	1.06	0.09	-0.18

[a] druggable if DrugPredScore > 0.50
[b] druggable if SiteScore > 0.8
[c] druggable if DGscore > 0.5
[d] druggable if score > 0
[e] no cavity detected

4.3 DETERMINING A ROBUST SIMILARITY THRESHOLD FOR PAIRWISE COMPARISON OF BINDING SITES

To determine a reliable metric and similarity threshold for distinguishing similar from dissimilar binding sites, we measured the pairwise similarity of 1 538 pairs of protein-ligand binding sites[30] split in two categories, 769 pairs of known similar and 769 pairs of known dissimilar sites. Among the 16 384 possible classifications obtained by systematically varying 31 VolSite and Shaper parameters (**Supplementary Table 3**), the best classification was obtained using ligand-binding sites truncated at 6 Å and discarding 'null' pharmacophoric points from matching rules. The corresponding binary classification model gives an area under the ROC curve of 0.983 and nicely segregates similar from dissimilar pairs (**Figure 4A**). This result is of course to be expected since pairs of similar/dissimilar binding sites were chosen on pupose. However, it enables the automated determination of the optimal similarity threshold, found for a RefTversky similarity value of 0.35 that gives excellent F-measure, recall and precision values of 0.932 (**Figure 4B**). Alternatively, a more conservative cut-off of 0.44 could also be chosen, enabling a perfect prediction of all true positives (100 % precision). To ascertain that the two above-mentioned thresholds are not dataset-dependent, we generated a full similarity matrix from all current sc-PDB entries and examined the distribution of similarity scores from two randomly chosen subsets of 300 000 pairs (**Figure 4C**). First, the distribution of scores was identical (according to the Kruskall-Wallis test) whatever the sample chosen, suggesting that the selected number of 300 000 pairs was statistically significant. According to the Kolmogorov-Smirnov test, it follows a generalized extreme value distribution (test statistic D = 0.03281, P-value = 0.64132, α =0.05) with a probability density function of the type:

$f(x)=\frac{1}{\sigma}\exp(-(-1+kz)^{-1/k})\,(1+kz)^{-1-k/z}$ with

k = -0.24296

σ = 0.05309 (standard deviation)

μ = 0.24827 (mean value)

$$z=\frac{x-\mu}{\sigma}$$

The significance level *p* of the detected similarity represents the probability of obtaining the same or higher similarity score Z > z by chance is:

Plotting the probability p-value against the raw similarity scores from the random distribution (**Figure 4D**) indicates that the two thresholds previously used (0.35 and 0.44) corresponds to very low probabilities of 0.05 and 0.005, respectively.

By opposition to the Tanimoto similarity index, The Tversky index used by Shaper presents the advantage to delineate similarity among ligand-binding sites of different sizes (e.g. monomer vs homodimer-lining sites, site co-crystallized with two ligands of very different sizes). In this case, only the Tversky index detects similarity between both entries when the smallest of both sites is used as a reference.

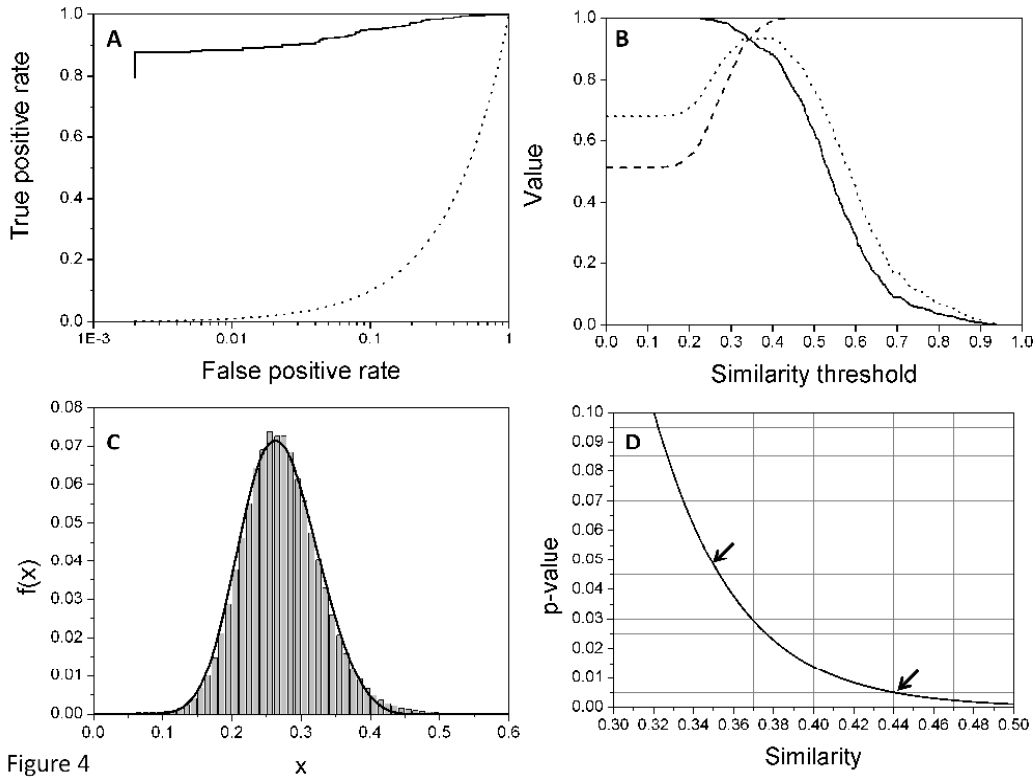


Figure 4

Figure 4 - Statistical evaluation of Shaper similarity scores. A) ROC plot (solid line) obtained by sorting 1 538 sc-PDB pairs of binding sites by decreasing RefTversky similarity values. True positives (n=769) are pairs of similar binding sites predicted similar whereas true negatives (n=769) are pairs of dissimilar sites predicted dissimilar. Accuracy of random picking is represented by a dotted line. B) Variation of statistical parameters (recall, solid line; precision, dashed line; F-measure, dotted line) of a binary classification model (similar/dissimilar) for increasing Tversky similarity score thresholds. C) Distribution of Shaper similarity scores for a randomly-chosen population of 300 000 scores retrieved from the all-against-all comparison of 9 877 sc-PDB binding sites (97 555 129 comparisons in total). The fit (bold line) to a generalized extreme value distribution represents the ideal probability density f(x) for a similarity value of x. D) Decay of the p-value (probability to get by chance a similarity score Z >x) as a function of the observed similarity value.

4.4 INFLUENCE OF VARIOUS PARAMETERS (GRID RESOLUTION AND ORIENTATION, ATOMIC COORDINATES) ON SIMILARITY MEASUREMENTS

Since the ligand-binding site is discretized at regularly-spaced grid points, we first investigated whether changing either the grid resolution or the grid center alters the description and comparison of ligand-binding sites. When applied to the classification of the above-described pairs of similar and dissimilar ligand-binding sites, obtained results slightly varies with the grid resolution (**Table 4**). As to be expected, a tighter grid spacing (1.0 Å) enables a better distinction of the two sets of binding sites (AUC = 0.991) whereas a smoother resolution (2.0 Å) deteriorates the binary classification (AUC = 0.947). Interestingly, the optimal similarity threshold (RefTversky index) and the classification accuracy (F-measure) slightly decrease when the grid spacing increases (**Table 4**). However, increasing the grid resolution significantly increases the cpu time necessary for defining the cavity points (**Table 4**). The intermediate resolution of 1.5 Å was therefore chosen as default since it yields the best compromise between speed and accuracy. This parameter is however user-tunable in Volsite.

Table 4 - Influence of the grid resolution on the correct classification of 1 538 pairs of ligand-binding sites (769 similar and 769 dissimilar pairs; see Methods for dataset description)

	Grid resolution, Å		
	1.0	**1.5**	**2.0**
AUC[a]	0.991	0.983	0.947
Threshold[b]	0.422	0.350	0.272
F-measure[c]	0.956	0.932	0.870
cpu[d]	46.29	8.18	2.36

[a] Area under the ROC curve for classifying the set of 769 similar and 769 dissimilar pairs, according to the measured similarity value (RefTversky metric)
[b] Optimal similarity threshold (RefTversky metric) for discriminating 769 similar from 769 dissimilar pairs.
[c] F-measure of the classification at the optimal similarity threshold
[d] cpu time in seconds (3.40 GHz Intel Pentium D processor with 2 Go RAM) for computing pharmacophoric site points for a ligand-binding site (PDB entry 1bji, ligand HET code: DPC) of 735 Å^3 (average volume among 9 877 sc-PDB binding sites)

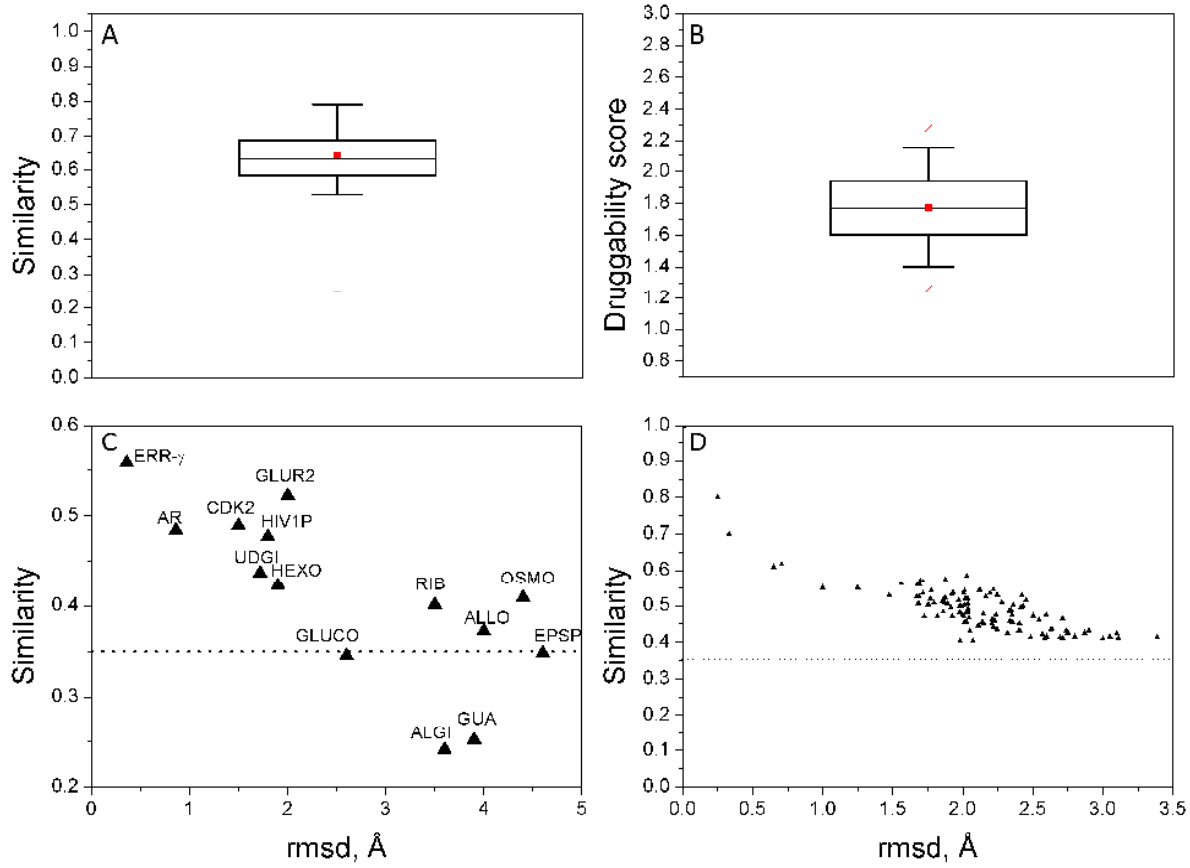


Figure 5- Sensitivity of Volsite and Shaper to input data.

A) Distribution of the pairwise similarity for an entire ligand-binding site (PDB entry 3k3i) after 30^3 -1 systematic translations (increment: 0.1 Å, range 3.0 Å) of the grid center along the three main axes. The box delimit the 25th and 75th percentiles, the whiskers delimit the 5th and 95th percentiles. The median and mean values are indicated by a horizontal line and a filled square in the box. Crosses delimit the 1% and 99th percentiles, respectively. Minimum and maximum values are indicated by a dash.

B) Distribution of the predicted druggability score for the 27 000 representations of the 3k3i ligand-binding site. The box delimit the 25th and 75th percentiles, the whiskers delimit the 5th and 95th percentiles. The median and mean values are indicated by a horizontal line and a filled square in the box. Crosses delimit the 1% and 99th percentiles, respectively. Minimum and maximum values are indicated by a dash.

C) Sensitivity to atomic coordinates: Shaper similarity versus rms deviations of the holo from the apo structure (active site only) of 14 targets: uracil DNA-glycosylase inhibitor (UDGI, 36 residues, pdb identifier 1udi vs. 1ugi), cell division protein kinase 2 (CDK2, 36 residues, 1dm2 vs. 2jgz), HIV-1 protease (HIV-1p, 52residues, 1qbs vs. 1hhp), estrogen-related receptor gamma (ERRg, 27 residues, 2zkc vs. 2zbs), aldose reductase (AR, 24 residues, 1ads vs. 2nvd), glutamate receptor subunit 2 (GLUR2, 22 residues, 1ftm vs. 1fto), DNA-β-glucosyltransferase (GLUCO, 28 residues, 1jg6 vs. 1jej), D-allose binding protein (ALLO, 21 residues, 1rpj vs. 1gud), D-ribose binding protein (RIB, 20 residues, 2dri vs. 1urp), 5-enolpyruvylshikimate-3-phosphate synthase (EPSP, 33 residues, 1rf4 vs. 1rf5), Osmo-protection protein (OSMO, 19 residues, 1sw2 vs. 1sw5), Guanylate kinase (GUA, 24 residues, 1ex7 vs. 1ex6), Hexokinase (HEX, 25 residues, 2e2o vs. 2e2n), and Alginate-binding protein (ALGI, 19 residues, 1y3n vs. 1y3q). The active site was defined from the holostructure by considering any amino acid with at least one heavy atom present in a 6.5 Å-radius sphere centered on the center of mass of the bound ligand.

D) Sensitivity to atomic coordinates: Shaper similarity versus rms deviations of active site heavy atoms to the X-ray structure for a 1-ns molecular dynamics (MD) simulation of the water solvated cyclin-dependent kinase type 2 (PDB entry 1dm2). Rmsd for 100 MD snapshots are displayed. The horizontal dotted line represents the similarity threshold (0.35) used throughout this study to discriminate similar from dissimilar protein-ligand binding sites.

In a second computational experiment, the center of the grid encompassing the ligand-binding site was systematically translated from the origin (center of mass of the bound ligand) by 0.1 Å increments up to 3.0 Å (1.5 Å in each direction from the origin) along the three main axis, therefore leading to 26 999 (30^3 -1) alternative grid lattices for a unique binding site. The pairwise similarity of all these representations to the native one is far above the previously-defined similarity threshold of 0.350 (**Figure 5A**), therefore demonstrating that both VolSite and Shaper are insensitive to overall translations of the grid lattice up to 1.5 Å (the default grid resolution). It should be noted that the druggability score predicted by the herein presented SVM model is also relatively insensitive to grid translations (**Figure 5B**).

In a last control experiment, we investigated how much the Volsite druggability and Shaper similarity scores vary with moderate variations in atomic coordinates of the cavity under investigation. A dataset of 14 proteins for which ligand-free and ligand-bound X-ray structures are available was investigated for Shaper pairwise comparisons (**Figure 5C**). In 12 out of 14 cases, despite significant changes occurring at the binding site (up to 4.6 Å rmsd on binding site heavy atoms), the computed similarity was above similarity threshold of 0.35. Likewise, molecular dynamics snapshots of a typical druggable cavity (ATP-binding site of cyclin dependent kinase type 2) were still considered similar enough (RefTversky similarity > 0.35) to the native crystal structure (**Figure 5D**). Importantly, the VolSite druggability score (1.15 ± 0.21) did not vary along either the MD trajectory. Both tools are therefore relatively insensitive to moderate variation in protein coordinates up to 3.0-3.5 Å rmsd.

4.5 VIRTUAL SCREENING FOR SIMILARITY TO A KNOWN CAVITY

Predicting the function of a protein from the similarity of its cavities to functionally annotated binding sites may help the annotation of novel genomic structures. To address this issue, we measured the pairwise similarity between the inhibitor-binding site in bovin trypsin (PDB entry 1aq7) and 5 952 sc-PDB binding sites. The canonical inhibitor binding site in trypsin was chosen here as a template for two main reasons: (i) trypsin belongs to the family of serine endopeptidases which presents the interest to share the same catalytic activity, a common catalytic triad, but different 3-D folds (trypsin, subtilisin, αβ hydrolase); (ii) the same query has already been

conducted by us with two other binding site similarity search programs (SiteAlign[50] and FuzCav[30]) thus enabling to compare the herein presented approach to two state-of-the-art methods.

When comparing the close proximity of bound ligands (cavities truncated at 4 Å), Shaper clearly discriminate endopeptidases from all other entries (**Table 5**). As expected, the highest ROC score is observed for the group of entries (group 1) sharing the fold and substrate specificity with the trypsin query. However, the second (trypsin fold, other cleavage specificity), third (subtilisin fold) and fourth (αβ hydrolase fold) groups are also statistically enriched in binding sites found similar to that of bovine trypsin. Enlarging the binding site definition (truncations at 6, 8, 12 Å; no truncation at all) changes the scope of the search to retrieve binding sites that are no more locally but globally similar to the query. It is therefore no surprise that the ROC score increases for the closest group (group 1) and decreases for groups exhibiting only local similarity at the catalytic side (groups 3 and 4).

Table 5- Area under the ROC curve for classification models of 5 952 sc-PDB entries according to Shaper similarity to the 1aq7 PDB entry (bovine trypsin in complex with inhibitor Aeruginosin 98-B).

Group	Ligand-binding site trunaction				
	4 Å	**6 Å**	**8 Å**	**12 Å**	**none**
1[a]	0.868	0.914	0.962	0.964	0.965
2[b]	0.771	0.684	0.764	0.763	0.763
3[c]	0.779	0.667	0.634	0.630	0.634
4[d]	0.684	0.618	0.593	0.606	0.606
5[e]	0.151	0.109	0.063	0.062	0.062

[a] Serine proteases with trypsin fold and trypsin substrate specificity
[b] Serine proteases with trypsin fold and other substrate specificity
[c] Serine protease with subtilisin fold
[d] Serine protease with αβ hydrolase fold
[e] any other sc-PDB entry

A second screen against a reference cavity from a structurally different class (protein kinase Pim-1) confirmed these results (**Table 6**). In this example, we investigated the effect of changing the reference cavity (same binding site but co-crystallized with three chemically different inhibitors) on the screening results. As noted for the previous trypsin screening, a good classification of protein kinases from other protein classes is obtained, whatever the binding site truncation method and the reference cavity (**Table 6**). As previously observed,[30,50,57] we confirm that ATP-binding sites in protein kinases dot not ressemble neither ATP-binding sites in other kinases, nor generic ATP/ADP-binding sites in general (**Table 6**).

When compared to previously reported benchmarks, Shaper was shown to perform as well as SiteAlign[50] and FuzCav[30] in discriminating the different endopeptidase groups from decoy binding sites (**Table 7**). As initially requested, the method is therefore able to combine the advantage of an alignment-dependent method (visualization and interpretation of matched structures) with the speed of an alignment-free method (ca. 10 comparisons/sec on an Intel Xeon E5504 processor).

Table 6 - Area under the ROC curve for classification models of 10 435 sc-PDB entries according to Shaper similarity to the ATP-binding site of Pim-1 kinase. Values are means and standard deviations for three independent screens against three Pim-1 kinase binding sites (1hys, 3cy3, 1yi4)

Group	Ligand-binding site truncaction				
	4 Å	**6 Å**	**8 Å**	**12 Å**	**none**
1[a]	0.883 ± 0.007	0.888 ± 0.008	0.892 ± 0.007	0.886 ± 0.007	0.886 ± 0.010
2[b]	0.428 ± 0.008	0.430 ± 0.008	0.422 ± 0.004	0.419 ± 0.006	0.430 ± 0.003
3[c]	0.460 ±0.012	0.470 ± 0.003	0.473 ± 0.010	0.474 ± 0.008	0.440 ± 0.013
4[d]	0.269 ± 0.002	0.263 ± 0.004	0.261 ± 0.002	0.265 ± 0.004	0.266 ± 0.009

[a] Protein kinases (n=1 138)
[b] Other kinases (n=294)
[c] Other ATP/ADP-binding sites (n= 423)
[d] Other sc-PDB entries (n=8 580)

Table 7 - Comparison of 3 binding site comparison methods, expressed by area under the ROC curve for binary classification models of 5 882 sc-PDB entries according to Shaper similarity to the 1aq7 PDB entry (bovine trypsin in complex with inhibitor Aeruginosin 98-B)

Group	SiteAlign[a]	FuzCav[b]	Shaper[c]
1[d]	0.939	0.906	0.914
2[e]	0.604	0.780	0.684
3[f]	0.462	0.649	0.779
4[g]	0.486	0.662	0.618
5[h]	0.109	0.114	0.109

[a] 3-D alignment-based, slow (2 comparisons/min).[41]
[b] 3-D alignment-free, ultra-fast (1 000 comparison/s).[18]
[c] 3-D alignment based, fast (10 comparisons/s)
[d] Serine proteases with trypsin fold and trypsin substrate specificity
[e] Serine proteases with trypsin fold and other substrate specificity
[f] Serine protease with substilisin fold
[g] Serine protease with αβ hydrolase fold
[h] any other sc-PDB entry

4.6 BINDING SITE SIMILARITY DETECTION AS A FUNCTION OF TARGET SEQUENCE AND STRUCTURE CONSERVATION

The previous application illustrated the ability of Shaper to detect local and global binding site similarities among a class of proteins sharing the same catalytic activity. Detecting remote ligand-binding site similarity among unrelated proteins is undoubtley more difficult. We specifically designed a dataset of promiscuous ligand-binding sites to address this issue and challenged our approach for difficult cases of shared ligand binding irrespective of sequence and structure conservation. We have identified 1 070 pairs of protein-ligand complexes in which the same ligand has been co-crystallized with different proteins. By computing both sequence and structure conservation of the corresponding targets (see Methods), we have classified these pairs in three categories: (i) easy: which means that the ligand is shared by proteins exhibiting both sequence (sequence identity >25%) and structure conservation (CE Z score > 4), (ii) medium, the ligand being shared by proteins showing structure but not sequence conservation, (iii) difficult: the ligand being shared by proteins exhibiting neither sequence nor structure conservation.

Comparison of Shaper with respect to SiteAlign and FuzCav in detecting binding site similarity across these 1 070 pairs show three clear trends (**Figure 6**) : (i) in easy cases, all programs perform very well and recover ca. 85% of the pairs as truly similar; (ii) in cases of medium difficulty, the success rate drops drastically for SiteAlign (52%) and FuzCav (36%) but not for Shaper (76%); (iii) in really difficult cases, only Shaper provides a good performance (46 % of pairs recovered) whereas SiteAlign and FuzCav fails in 90% of the cases (**Figure 6**). Out of the three tools, Shaper is clearly the one that recovers the highest proportion of similar binding sites (69%). The noticeable advantage of Shaper to detect binding site similarity in cases of medium and high difficulty is explained by a better description of binding site attributes, notably the shape of the cavity that is a known important prerequisite for ligand binding. SiteAlign and FuzCav both encode the same pharmacophoric properties than Shaper but only at the C atom of the cavity-lining residue. Conversely, Shaper places multiple pharmacophoric points at the vicinity of all ligand-accessible binding site atoms and therefore gives more importance to ligand-accessible than to buried protein atoms.

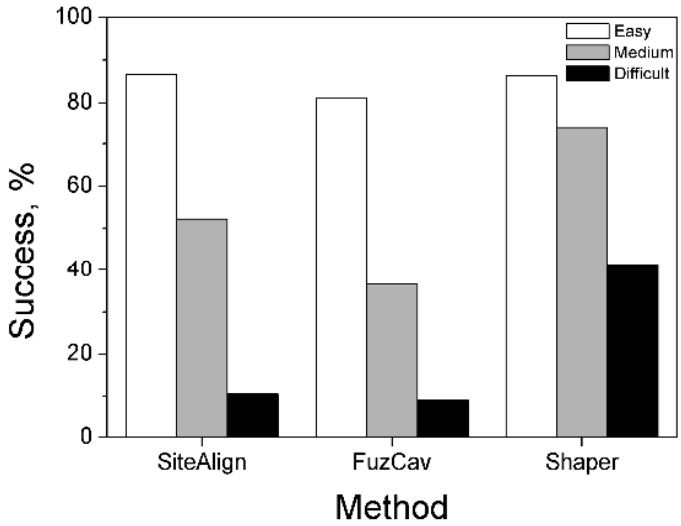


Figure 6 - Comparative ability of three binding site comparison programs (SiteAlign, FuzCav, Shaper) in detecting similarity for ligand-binding sites from different proteins but sharing the same ligand. Success is defined when the pairwise similarity is above each program-specific similarity threshold[18,41] (SiteAlign: d1 <0.6 and d2 <0.2; FuzCav: score >0.16; Shaper: score >0.35). Binding sites are classified in three categories: easy (white bars, conserved sequence and structure), medium (gray bars, different sequence but conserved structure), and difficult (dark bars, different sequence and structure).

4.7 NETWORK OF BINDING SITES FOR A PROTEIN FAMILY

The recent X-ray structure determination of several GPCRs in complex with non-covalent drug-like compounds[19] offered us the opportunity to measure pairwise similarities of 30 ligand-binding sites from 14 different receptors. We particularly paid attention to (i) the comparison of different entries of the same receptor co-crystallized with ligands exhibiting different functional effects (agonist, inverse agonist, antagonist); (ii) the influence of the binding site definition (immediate vicinity of the ligand, full cavity) on the obtained networks.

The network obtained from full cavities is very homogeneous and separate very well entries by receptor name (**Figure 7A**), with the exception of the very similar 1 and 2 adrenergic receptors that are grouped together. The dopamine D3 receptor cavity links the adenosine A2a receptor group to the adrenergic receptor entries. Three receptors (S1P1, Histamine H1, CXCR4) exhibit unique binding site properties and thus are represented as singletons. It is noteworthy that the solvent-exposed peptide-binding site in CXCR4 (3oe0) is not related to the transmembrane non-peptide binding site of the same receptor (3odu). As to be expected from the fine observation of all X-ray structures,[19] it is not possible to distinguish agonist from antagonist (or inverse agonist) binding sites at the default Shaper resolution. However, it is interesting to notice the much more complex network derived from binding sites truncated at a maximal 4 Å-distance from the bound ligand (**Figure 7B**). The later network, although still grouping entries by receptor names offers much more edges (244 vs. 111 for the previous network) between different receptors, mainly those of biogenic amines (1 and 2 adrenergic, dopamine D3, histamine H1, muscarinic M2 and M3). This observed difference is in agreement with the recently established evidence that fine receptor selectivity is principally gained from interaction of ligand moities at the periphery of transmembrane binding sites (notably close to the extracellular loops) and not within the ancestral retinal binding site.[19] The possible definition of binding sites of increasing sizes (from the ligand center of mass) in VolSite permits to delineate the presence or absence of ligand-proximal or more distal relationships, and therefore to estimate whether selective ligands could be designed for receptors from the same family.

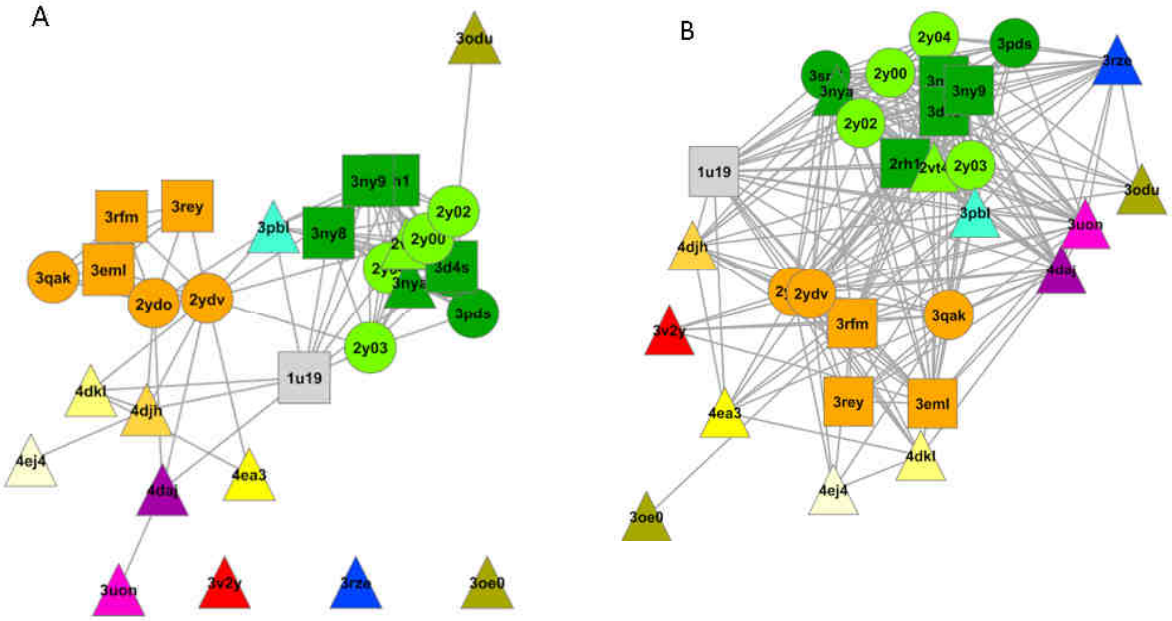


Figure 7 - Network of G protein-coupled receptor binding sites. A) Network of full cavities, B) Network of binding sites truncated at 4 Å of the bound ligand. Nodes are colored by receptor name (Adenosine A2a, orange; 1 adrenergic, light green; 2-adrenergic, dark green; chemokine CXCR3, olive; Dopamine D2, cyan; Histamine H1, blue; Muscarinic M2, light purple; Muscarinic M3, purple; delta opiate, light yellow; kappa opiate, light orange; mu opiate, yellow; nociceptin, bright yellow; rhodopsin, grey; sphingolipid S1P1, red) and shaped according the functional effect of the bound ligand (agonist, circle; inverse agonist, rectangle; antagonist, triangle).

5. CONCLUSION

We herewith present two complementary methods for detecting and comparing protein-ligand binding sites. VolSite first describes cavities from the known position of bound ligands and represent the binding site by grid points bearing pharmacophoric properties complementary to that of the nearest protein atom. VolSite has been applied to pockets occupied by known ligands but could be easily used to systematically scan an entire protein surface and rank detected cavities by decreasing ligandability. A further line of improvement lies in the choice for assigning pharmacophoric properties to cavity points. The current protocol uses simple distance criteria (closest protein atom) to match a pharmacophore feature onto site points. A probabilistic approach, taking into account the density of the different protein atom types at the close vicinity of the cavity point, could avoid polarity mismatches (e.g. hydrophobic point in a very polar environment or vice-versa) or incorrect assignments arising from local uncertainties in protein atomic coordinates. Beside discribing cavities, VolSite descriptors can be directed read by a machine learning classifier (SVM in the present study) for predicting the ligandability (structural druggability) of the corresponding pocket, with accuracy at least similar to that of the best methods reported yet.

The second tool (Shaper) aligns and measures the similarity of two pockets by approximating site points with Gaussians thus enabling to quickly align two sites by optimizing their volume overlap. Shaper was shown to combine the pace of alignment-free site comparison tools and the accuracy and interpretability of alignment-dependent methods. It can be used for clustering a set of binding sites according to their physicochemical properties as well as screening a collection of binding sites for similarity to a query. Interestingly, Shaper was particularly efficient in detecting binding site similarity in absence of sequence of fold conservation. Both methods are relatively insensitive to variations in atomic coordinates and definition of the grid box.

6. COMMENTAIRES

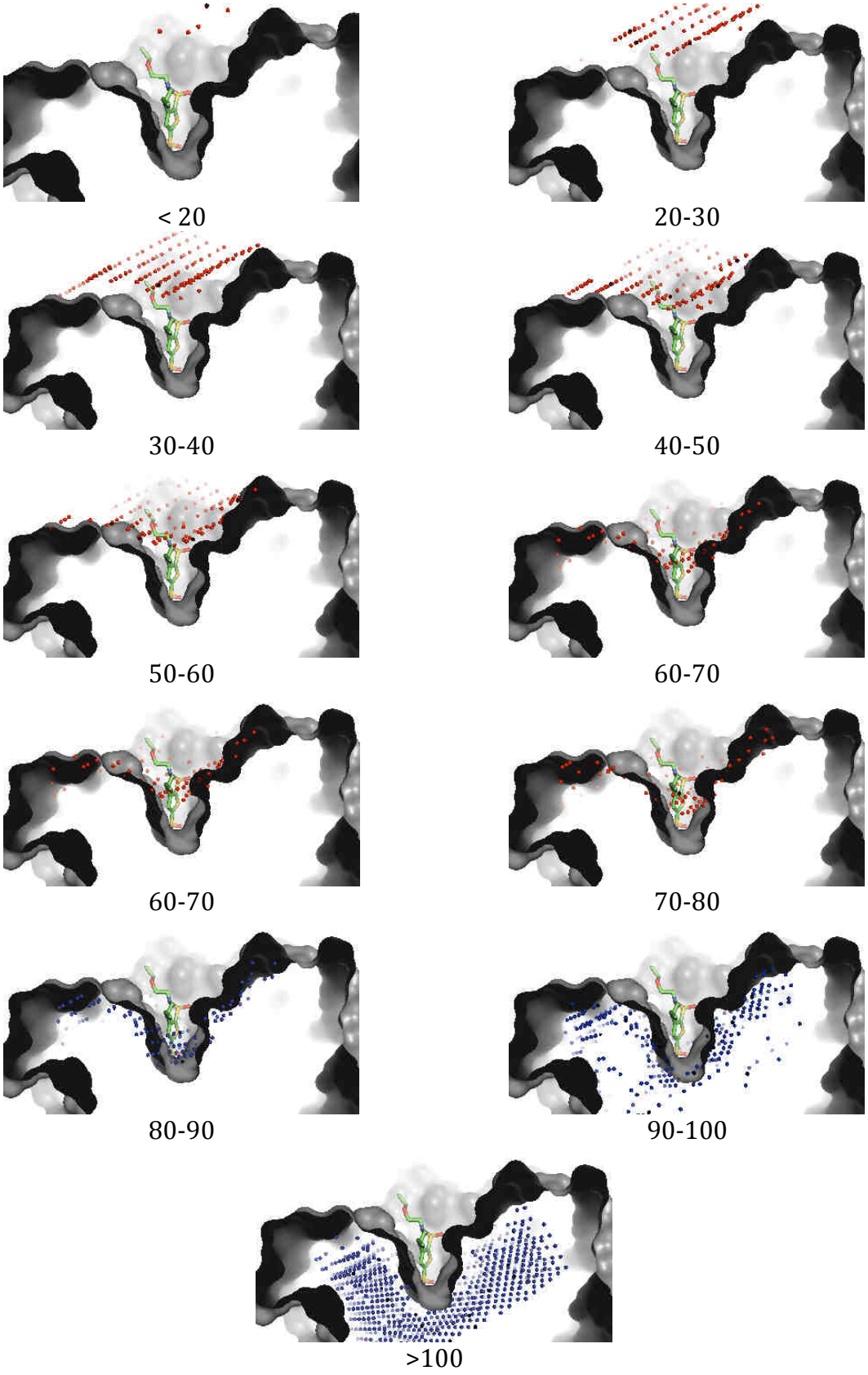
6.1 MODIFICATIONS POST-PUBLICATION

6.1.1 VALEURS D'ENFOUISSEMENT

Le calcul de l'enfouissement de chaque cube est une étape clé pour définir des cavités. Sans ce principe, il nous serait impossible de délimiter correctement les bordures de celle-ci, ou de différencier un point se situant loin de la protéine et pourtant complètement enfoui d'un point en bordure de celle-ci. Cependant, la valeur obtenue ne nous permet pas de savoir si la protéine est équitablement distribuée dans les trois directions de l'espace ou si elle est densément localisée dans une région spécifique de l'espace pour un point donné. Cette caractérisation autoriserait la différenciation des points réellement enfouis de points en surface et d'affiner la représentation d'une cavité. Cela pourrait aussi être ajouté en tant que paramètre dans la prédiction de la droguabilité.

Une autre problématique dans le calcul de l'enfouissement est l'effet de bord dû à la définition du cube principal. Pour rappel, le calcul d'enfouissement s'effectue en générant 120 vecteurs de 8 Å dans les 3 directions de l'espace et en comptant le nombre d'intersections avec la protéine. Cependant, et dans un souci d'efficacité et de rapidité, les intersections sont basées sur les cubes contenant un atome de protéine, et non les atomes eux-mêmes. Par conséquent, lorsque l'on observe un cube à la bordure de la boite, la moitié de ses projections est ignorée. Pour résoudre ce problème, la création d'une marge a été réalisée. Lors de la génération du cube principal, une marge de la longueur des projections est créée : si le cube fait 20 Å d'arête et que les projections sont limitées à 8 Å, une marge de 8 Å sera générée dans les 3 directions de l'espace. Les cubes faisant partie de cette marge sont seulement coloriés « IN » lorsqu'ils contiennent un atome de protéine, et ne participeront pas à la cavité finale. Ils existent seulement pour affiner les valeurs de projection et être indépendant de la taille du cube.

Tableau 8 - Effet de la valeur d'enfouissement sur la sélection des cubes pour la cavité du brinzolamide complexé avec l'anhydrase carbonique II humaine (code PDB:1a42). En gris, la surface de la protéine, en vert le brinzolamide. En rouge les cubes de cavités hors de la protéine et en bleu ceux à l'intérieur de la protéine. Chaque valeur correspond à une tranche d'enfouissement.



Le **Tableau 8** montre l'évolution de l'enfouissement par rapport à ce critère. On peut voir que la nouvelle fonction de calcul est ainsi très graduelle et représente bien la réalité du concept. Cependant, on remarque aussi que le seuil initial d'enfouissement de 40/120 est maintenant trop faible pour nos besoins. L'affinement de ce seuil est une étape clé dans la définition des cavités, et par conséquent du regroupement des points.

6.1.2 AGREGATION

Une fois le calcul de projection terminé, le logiciel possède un ensemble de points de cube qu'il doit regrouper pour définir des cavités. Cette étape, cruciale, est un problème inhérent à la représentation et à la caractérisation de sites actifs. En effet, une cavité est un concept, et ne possède donc pas de critères propres permettant de clairement la définir.

A l'époque de la publication, la méthode de regroupement était initialement conçue en agrégeant entre eux les points à proximité immédiate des autres, et en répétant l'opération jusqu'à ce que tous les points soient pris en compte. Cette méthode, bien qu'elle ait prouvé son efficacité dans les divers tests de criblage virtuel et de prédiction de droguabilité, est limitée à cause de sa simplicité. En effet, si un cube ou un groupe de cubes possède au moins une arête en commun avec un autre cube ou groupe, alors les deux seront regroupés. Cela implique que deux cavités bien distinctes possédant 2 cubes voisins l'un de l'autre seront au final une seule et même cavité (**Figure 8A**). Le cas du brinzolamide en complexe avec l'anhydrase carbonique II humaine en est un exemple typique. On peut observer dans ce cas simple 2 cavités distinctes : la cavité du ligand en rouge, et une seconde, située en surface de la protéine, en bleue. Cette dernière ne peut clairement pas accueillir un ligand tant la cavité est petite et hydrophobe. Cependant, dans le cas du clustering tel que publié, ces deux cavités seront associées en une seule, modifiant ainsi la similarité lors de la comparaison avec d'autres cavités, mais aussi et surtout la droguabilité du site.

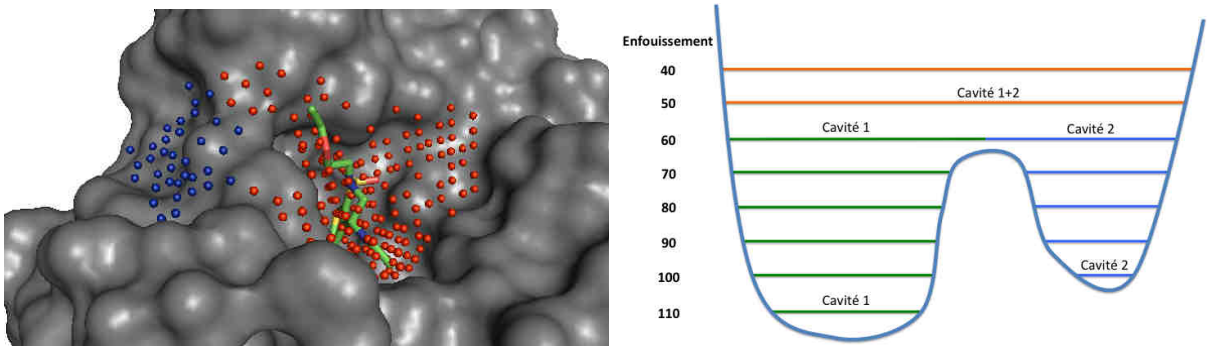


Figure 8 - A gauche : brinzolamide complexé avec l'anhydrase carbonique II humaine (code PDB:1a42). En gris, la surface de la protéine. En vert: le ligand. En rouge, la cavité principale. En bleu, une cavité secondaire détectée. **A droite :** Nouvelle méthode de clustering. La protéine est représentée en bleu foncé. L'algorithme commence par rassembler les points les plus enfouis (ici à 110) formant ainsi la cavité 1 (en vert). On continue à rassembler les points par ordre d'enfouissement décroissant. Lorsqu'un point ne peut être rattaché à un cluster, il définit un nouveau cluster (cavité 2 - en bleu). Lorsque deux clusters partage une surface commune (ici à enfouissement =60) les clusters ne sont pas regroupés tant que la surface n'est pas suffisante. Lorsque cette condition est respecté (ici à enfouissement=50), les cavités sont rassemblées.

Afin d'améliorer la définition de cette étape, une nouvelle méthode de clustering, plus « naturelle » a été mise en place, et repose sur un gradient d'enfouissement. Afin d'illustrer le concept, prenons l'exemple de deux flaques d'eau proches légèrement séparées l'une de l'autre. Lorsque de l'eau est ajoutée dans chacune des flaques, le niveau augmentera, entrainant ainsi un niveau suffisant pour les deux flaques de communiquer. La différence principale qui réside en deux flaques juxtaposées et une grande flaque, consiste dans l'observation de la séparation. Si celle-ci est encore visible car le niveau n'est pas assez élevé, elles seront considérées comme séparées. A contrario, si celle-ci a disparu sous le niveau de l'eau, alors nous n'avons qu'une seule flaque (**Figure 8B**).

On commence en regardant l'environnement des points les plus enfouis, définissant ainsi les clusters initiaux. Pour chaque cube, une densité est calculée et fonction de l'environnement protéique et des points appartenant au même cluster. Chaque cube possède, en 3 dimensions, 27 cubes partageant une arête ou un sommet en commun. Parmi ces 27 cubes, certains caractérisent la protéine, d'autres appartiennent au même cluster, d'autres ne sont pas suffisamment enfouis ou ne sont pas encore traités. Lorsque l'on agrège, on peut ainsi calculer le gain d'association, c'est à dire le nombre d'arêtes ou de sommets en commun gagné via le clustering :

où NA est le nombre de cubes dans le cluster A, NB le nombre de cubes dans le cluster B, N_{Ci} est le nombre d'arêtes ou de sommets en commun gagnés par l'association et $Nsit_i$ le nombre maximal de cubes pouvant être associé au cube i. Cette formule, associée à une dépendance à la taille des clusters permet d'être plus tolérant pour des petits clusters et d'être plus exigeant par rapport à de gros clusters. Ainsi, plus la taille sera importante, plus le nombre d'arêtes ou de sommets en commun devra être important avant de considérer une fusion. Cette nouvelle méthodologie reste encore à être paramétrée sur un jeu de protéine structuralement divers, en utilisant par exemple la classification CATH.[45]

6.2 DETECTION DES CAVITES D'UNE PROTEINE

Une protéine, de part son arrangement spatial, n'est pas lisse mais possède un ensemble de poches, tunnels et cavités. Etre capable de déterminer la poche qui pourra accueillir une molécule de faible poids moléculaire est un enjeu crucial dans la recherche pharmaceutique. La classification de ces poches suivant divers critères tels que orthostérique ou allostérique, hydrophobe ou hydrophile, enfoui ou en surface, en longueur ou sphérique, droguable ou non, permettra ainsi de mieux cerner les règles de reconnaissance moléculaire.

De nombreuses méthodes ont été développées au cours de 50 dernières années avec des légères modifications mais toujours une ligne directrice. Initialement, Lee et Richards utilisèrent la surface de Connolly, concept purement géométrique pour détecter les cavités.

La détection des cavités d'une protéine est un domaine déjà largement analysé et de nombreuses méthodes ont été développées avec pour but la détection et la caractérisation des poches de protéine. L'un des premiers logiciels, POCKET,[58] utilise la surface de Connolly[59] pour détecter les cavités. Depuis, de nombreuses améliorations ont été réalisées :

- l'indépendance à l'orientation initiale de la protéine[60]
- la prise en compte de l'enfouissement comme critère de sélection des points. Chaque point est ainsi définit par son environnement protéique proche.[1,61]
- le regroupement des points par rapport à l'enfouissement permettant de définir ainsi une limite entre la poche et l'extérieur[62,63]

- la caractérisation physico-chimique des poches par l'analyse des atomes de protéine : enfouissement, ouverture, polarité, surface, volume.[52]
- L'ajout de la droguabilité comme critère de sélection des poches[5,56]
- La prise en compte des molécules d'eau
- La possibilité de comparer et d'obtenir un score de similarité basé sur les poches.[64]

Les derniers développements présentés précédemment offrent à VolSite la possibilité de détecter toutes les cavités d'une protéine. En plus, grâce à l'indice de droguabilité, cette méthode permettra de catégoriser les poches en fonction de leur taille, leur hydrophobie et leur enfouissement relatif. Il sera ainsi possible de réaliser une analyse plus fine des cavités issues de la base sc-PDB. L'analyse et la classification de poches en surface aideront à définir de meilleures règles de sélection de cavités droguables. La comparaison des cavités avec la solvatation de celles-ci, dans le cas où elles n'ont pas de ligand, apportera des informations importantes sur l'organisation des molécules d'eau, et permettra ainsi affiner l'assignation des propriétés physico-chimiques.

6.3 ASSIGNATION DE PROPRIETES PHYSICO-CHIMIQUES AUX CUBES

Un point d'amélioration possible de VolSite concerne l'attribution des propriétés physico-chimiques des cubes de la cavité. En effet, celle-ci se base sur la propriété complémentaire à l'atome de protéine le plus proche du centre du cube. Une problématique nous apparait alors : dans le cas où l'atome le plus proche est un atome hydrophobe, et le second atome, accepteur de liaison hydrogène, se situe à 0.1Å de plus que le premier, la propriété finale du cube sera hydrophobe. Qui plus est, si trois accepteurs de liaisons hydrogènes se trouvent à 0.1Å de plus, la propriété restera hydrophobe. Par conséquent, cette attribution est dépendante des coordonnées atomiques dont l'écart est peu significatif par rapport à la déviation moyenne expérimentale issue de la résolution cristallographique. Plusieurs solutions s'offrent à nous tels que la génération d'une chaine d'entiers par cube encodant l'environnement protéique local. Chaque case de cette chaine correspondrait à une propriété physico-chimique (**Figure 9**) : hydrophobe, aromatique, accepteur et donneur de liaison H (séparément), charge positive et négative (séparément). La présence d'une propriété

protéique ne serait ainsi pas définie par l'incrément d'une unité pour chaque case, mais par le rapport entre la distance maximale autorisée pour le type d'interaction et sa distance entre le dit atome et le cube :

Cette densité ainsi définie ne serait pas normalisée afin de mieux prendre en compte la présence de chaque interaction. Les propriétés de la cavité ne seraient donc plus décrites par un pourcentage de propriétés mais par un ratio de propriétés physico-chimiques accessibles. L'alignement et la similarité ne seraient plus calculés par Shaper mais par un algorithme de recouvrement de graphes, définissant des paires si le recouvrement local des propriétés serait supérieur à un certain seuil à échantillonner. La densité d'hydrophobes serait minimisée par un poids plus faible que les autres densités afin de mieux prendre en compte l'importance des propriétés hydrophiles. De plus, cela permettrait d'observer des conservations locales de sites actifs, jusqu'alors peu étudiées.

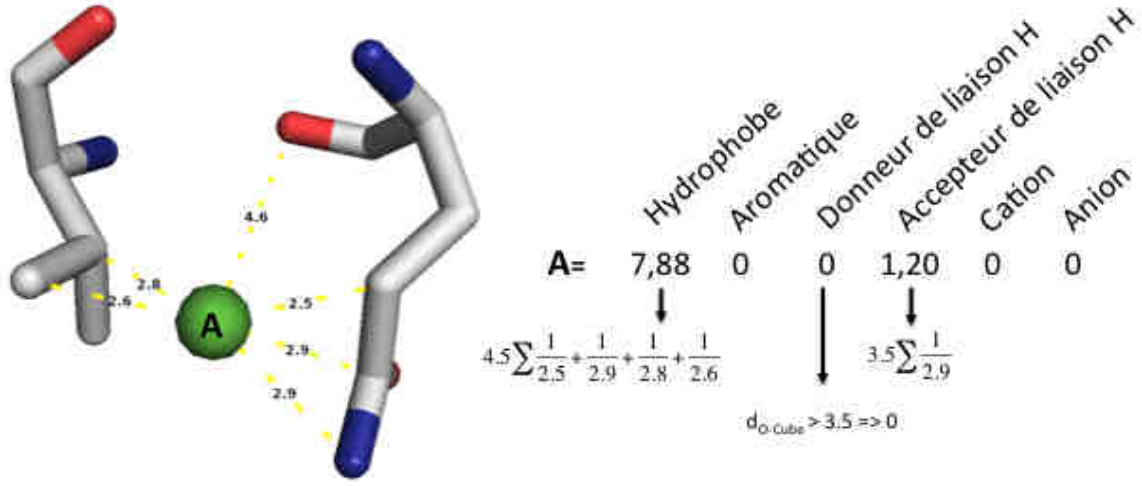


Figure 9 - Nouvelle méthode d'assignation des propriétés physico-chimiques pour un cube de cavité. Le centre du cube A est représenté en vert, l'environnement protéique ainsi que les distances sont plus proches sont calculés. Chaque cube est représenté par 6 cases. Chaque valeur est assignée comme étant la somme des rapports entre la distance maximale autorisée pour chaque interaction et la distance réelle entre l'atome de protéine et le cube observé. Si la distance atome-cube est supérieure à la distance maximale, celle-ci est ignorée.

7. ANNEXES

Supplementary Table 1. List of 73 VolSite descriptors used for predicting structural druggability

Number	Description
1	Total number of cavity points
2	Percent of hydrophobic points (H)
3	Percent of aromatic points (Ar)
4	Percent of H-Bond acceptor points (HBA)
5	Percent of Negative Ionizable points (A-)
6	Percent of Acceptor/Donor points (HBAD)
7	Percent of H-Bond donor points (HBD)
8	Percent of Positive Ionizable points (D+)
9	Percent of Dummy points (/)
10	Percent of (H) points with a projection value[a] between 40 and 50
11	Percent of (H) points with a projection value between 50 and 60
12	Percent of (H) points with a projection value between 60 and 70
13	Percent of (H) points with a projection value between 70 and 80
14	Percent of (H) points with a projection value between 80 and 90
15	Percent of (H) points with a projection value between 90 and 100
16	Percent of (H) points with a projection value between 100 and 110
17	Percent of (H) points with a projection value between 110 and 120
18	Percent of (Ar) points with a projection value between 40 and 50
19	Percent of (Ar) points with a projection value between 50 and 60
20	Percent of (Ar) points with a projection value between 60 and 70
21	Percent of (Ar) points with a projection value between 70 and 80
22	Percent of (Ar) points with a projection value between 80 and 90
23	Percent of (Ar) points with a projection value between 90 and 100
24	Percent of (Ar) points with a projection value between 100 and 110
25	Percent of (Ar) points with a projection value between 110 and 120
26	Percent of (HBA) points with a projection value between 40 and 50
27	Percent of (HBA) points with a projection value between 50 and 60
28	Percent of (HBA) points with a projection value between 60 and 70
29	Percent of (HBA) points with a projection value between 70 and 80
30	Percent of (HBA) points with a projection value between 80 and 90
31	Percent of (HBA) points with a projection value between 90 and 100
32	Percent of (HBA) points with a projection value between 100 and 110
33	Percent of (HBA) points with a projection value between 110 and 120
34	Percent of (A-) points with a projection value between 40 and 50
35	Percent of (A-) points with a projection value between 50 and 60
36	Percent of (A-) points with a projection value between 60 and 70

37	Percent of (A-) points with a projection value between 70 and 80
38	Percent of (A-) points with a projection value between 80 and 90
39	Percent of (A-) points with a projection value between 90 and 100
40	Percent of (A-) points with a projection value between 100 and 110
41	Percent of (A-) points with a projection value between 110 and 120
42	Percent of (HBAD) points with a projection value between 40 and 50
43	Percent of (HBAD) points with a projection value between 50 and 60
44	Percent of (HBAD) points with a projection value between 60 and 70
45	Percent of (HBAD) points with a projection value between 70 and 80
46	Percent of (HBAD) points with a projection value between 80 and 90
47	Percent of (HBAD) points with a projection value between 90 and 100
48	Percent of (HBAD) points with a projection value between 100 and 110
49	Percent of (HBAD) points with a projection value between 110 and 120
50	Percent of (HBD) points with a projection value between 40 and 50
51	Percent of (HBD) points with a projection value between 50 and 60
52	Percent of (HBD) points with a projection value between 60 and 70
53	Percent of (HBD) points with a projection value between 70 and 80
54	Percent of (HBD) points with a projection value between 80 and 90
55	Percent of (HBD) points with a projection value between 90 and 100
56	Percent of (HBD) points with a projection value between 100 and 110
57	Percent of (HBD) points with a projection value between 110 and 120
58	Percent of (D+) points with a projection value between 40 and 50
59	Percent of (D+) points with a projection value between 50 and 60
60	Percent of (D+) points with a projection value between 60 and 70
61	Percent of (D+) points with a projection value between 70 and 80
62	Percent of (D+) points with a projection value between 80 and 90
63	Percent of (D+) points with a projection value between 90 and 100
64	Percent of (D+) points with a projection value between 100 and 110
65	Percent of (D+) points with a projection value between 110 and 120
66	Percent of (/) points with a projection value between 40 and 50
67	Percent of (/) points with a projection value between 50 and 60
68	Percent of (/) points with a projection value between 60 and 70
69	Percent of (/) points with a projection value between 70 and 80
70	Percent of (/) points with a projection value between 80 and 90
71	Percent of (/) points with a projection value between 90 and 100
72	Percent of (/) points with a projection value between 100 and 110
73	Percent of (/) points with a projection value between 110 and 120

[a] number of 8 Å-long rays projected from the site point intersecting a 'IN' cell

Supplementary Table 2. Color force-field to post-process shape matching in Shaper

```
# Pharmacophoric types
TYPE Donor
TYPE Acceptor
TYPE Rings
TYPE Positive
TYPE Negative
TYPE Nulli
TYPE Hydrophobe
#
# SMARTS rules
PATTERN     Hydrophobe      [13C]
PATTERN     Donor           [15O,14N]
PATTERN     Acceptor        [15O,14O]
PATTERN     Rings           [15C]
PATTERN     Positive        [15N]
PATTERN     Negative        [17O]
PATTERN     Nulli           [H]
#
#Matching patterns
INTERACTION     Donor       Donor       attractive gaussian weight=1.0 radius=1.0
INTERACTION     Hydrophobe  Hydrophobe  attractive gaussian weight=1.0 radius=1.0
INTERACTION     Acceptor    Acceptor    attractive gaussian weight=1.0 radius=1.0
INTERACTION     Rings       Rings       attractive gaussian weight=1.0 radius=1.0
INTERACTION     Positive    Positive    attractive gaussian weight=1.0 radius=1.0
INTERACTION     Negative    Negative    attractive gaussian weight=1.0 radius=1.0
INTERACTION     Nulli       Nulli       attractive gaussian weight=1.0 radius=1.0
```

Supplementary Table 3. List of VolSite and Shaper parameters

Parameter	Value	Description
Volsite		
Size	4, 6, 8, 12	Largest distance (in Å) of a cavity point to a ligand heavy atom
Shaper		
Alignment method	1, 2, 3, 4	1: grid, 2: analytic, 3: analytic2, 4:exact (see OEShape documentation)
Radius	1, 1.5	Maximal distance (in Å) between 2 points to score by the color force-field
Acceptor	0, 1	Matching acceptor points (0: false, 1: true)
Donor	0, 1	Matching donor points (0: false, 1: true)
Aromatic	0, 1	Matching aromatic points (0: false, 1: true)
Hydrophobic	0, 1	Matching hydrophobic points (0: false, 1: true)
Negative	0, 1	Matching negative ionizable points (0: false, 1: true)
Positive	0, 1	Matching positive ionizable points (0: false, 1: true)
Null	0, 1	Matching dummy points (0: false, 1: true)
NegAcc	0, 1	Matching negative ionizable to acceptor points (0: false, 1: true)
PosDon	0, 1	Matching positive ionizable to donor points (0: false, 1: true)
Similarity metric		
ColorTanimoto	[0-1]	Tanimoto coefficient of colors overlap
ShapeTanimoto	[0-1]	Tanimoto coefficient of shapes overlap
ComboTanimoto	[0-2]	ColorTanimoto + TanimotoShape
ColorFitTversky	[0-1]	Tversky index of colors overlap (α =0.05, β=0.95)
ShapeFitTversky	[0-1]	Tversky index of shapes overlap (α =0.05, β=0.95)
ComboFitTversky	[0-2]	ColorFitTversky + ShapeFitTversky
ColorRefTversky	[0-1]	Tversky index of colors overlap (α=0.95, β=0.05)
ShapeRefTversky	[0-1]	Tversky index of shapes overlap (α =0.95, β=0.05)
ComboRefTversky	[0-2]	ColorRefTversky + ShapeRefTversky
Tversky1	[0-2]	Combo score (Shape+Color); α=0.1, β=0.9
Tversky2	[0-2]	Combo score (Shape+Color); α=0.2, β=0.8
Tversky3	[0-2]	Combo score (Shape+Color); α=0.3, β=0.7
Tversky4	[0-2]	Combo score (Shape+Color); α=0.4, β=0.6
Tversky5	[0-2]	Combo score (Shape+Color); α=0.5, β=0.5
Tversky6	[0-2]	Combo score (Shape+Color); α=0.6, β=0.4
Tversky7	[0-2]	Combo score (Shape+Color); α=0.7, β=0.3
Tversky8	[0-2]	Combo score (Shape+Color); α=0.8, β=0.2
Tversky9	[0-2]	Combo score (Shape+Color); α=0.9, β=0.1
Similarity	[0-1]	Combo overlap / min(self score)

Supplementary Table 4. GPCR X-ray structures, along with ligand names and functional effects. Volume of the full cavity and druggability were estimated using standard parameters of Shaper

PDB id	Receptor	Res, Å	Ligand	Function	Volume,A^3	Druggability
2ydo	A2a	3.0	Adenosine	Agonist	611	2.37
2ydv	A2a	2.6	NECA	Agonist	526	2.21
3qak	A2a	2.7	UK-432097	Agonist	1026	1.43
3eml	A2a	2.6	ZM241385	Inverse agonist	1063	2.11
3rey	A2a	3.3	XAC	Inverse agonist	1120	2.14
3rfm	A2a	3.6	caffein	Inverse agonist	1005	2.14
2y00	Beta1	2.5	dobutamine	Agonist	941	1.10
2y02	Beta1	2.6	carmoterol	Agonist	810	1.25
2y03	Beta1	2.8	isoprenaline	Agonist	830	1.07
2y04	Beta1	2.4	salbutamol	Agonist	772	1.58
2vt4	Beta1	2.7	cyanopindolol	Antagonist	813	2.04
3pds	Beta2	3.5	FAUC50	Agonist	1059	1.41
3sn6	Beta2	3.2	BI-167107	Agonist	1002	1.59
3nya	Beta2	3.1	alprenolol	Antagonist	796	1.76
2rh1	Beta2	2.4	carazolol	Inverse agonist	833	1.60
3d4s	Beta2	2.8	timolol	Inverse agonist	719	1.90
3ny8	Beta2	2.8	ICI 118,551	Inverse agonist	877	1.82
3ny9	Beta2	2.8	Cpd 2	Inverse agonist	801	1.62
3odu	CXCR4	2.5	IT1t	Antagonist	1070	0.20
3oe0	CXCR4	2.9	CVX15	Antagonist	560	-0.81
3pbl	D3	2.9	Eticlopride	Antagonist	766	1.85
3rze	H1	3.1	Doxepin	Antagonist	654	1.01
3uon	M2	3.0	3-quinuclidinyl-benzilate	Antagonist	766	2.60
4daj	M3	3.4	Tiotropium	Antagonist	961	2.56
4ej4	OPRD	3.4	Naltrindole	Antagonist	1296	0.81
4djh	OPRK	2.9	JDTic	Antagonist	1110	1.53
4dkl	OPRM	2.8	β-FNA	Antagonist	1046	1.61
4ea3	OPRX	3.0	C-24	Antagonist	1312	0.97
1u19	OPSD	2.2	11-cis retinal	Inverse Agonist	398	1.42
3v2y	S1P1	2.8	ML056	Antagonist	793	0.71

8. BIBLIOGRAPHIE

(1) Liang, J.; Edelsbrunner, H.; Woodward, C. *Prot. Sci.* **1998**, *7*, 1884–97.

(2) Laskowski, R. a; Luscombe, N. M.; Swindells, M. B.; Thornton, J. M. *Prot. Sci.* **1996**, *5*, 2438–52.

(3) Pérot, S.; Sperandio, O.; Miteva, M. a.; Camproux, A.-C.; Villoutreix, B. O. *Drug Discov. Today* **2010**, *15*, 656–67.

(4) Hajduk, P. J.; Huth, J. R.; Fesik, S. W. *J. Med. Chem.* **2005**, *48*, 2518–25.

(5) Halgren, T. a *J. Chem. Inf. Model.* **2009**, *49*, 377–89.

(6) Cheng, A. C.; Coleman, R. G.; Smyth, K. T.; Cao, Q.; Soulard, P.; Caffrey, D. R.; Salzberg, A. C.; Huang, E. S. *Nat. Biotechnol.* **2007**, *25*, 71–5.

(7) Luque, I.; Freire, E. *Proteins* **2000**, *Suppl 4*, 63–71.

(8) Dodson, G.; Wlodawer, A. *Trends Biochem. Sci.* **1998**, *23*, 347–52.

(9) Nisius, B.; Sha, F.; Gohlke, H. *J. Biotech.* **2012**, *159*, 123–34.

(10) Henrich, S.; Salo-Ahen, O. M. H.; Huang, B.; Rippmann, F. F.; Cruciani, G.; Wade, R. C. *J. Mol. Recognit.***2010**, *23*, 209–19.

(11) Krasowski, A.; Muthas, D.; Sarkar, A.; Schmitt, S.; Brenk, R. *J. Chem. Inf. Model.* **2011**, *51*, 2829–42.

(12) Schmidtke, P.; Barril, X. *J. Med. Chem.* **2010**, *53*, 5858–67.

(13) Dessailly, B. H.; Nair, R.; Jaroszewski, L.; Fajardo, J. E.; Kouranov, A.; Lee, D.; Fiser, A.; Godzik, A.; Rost, B.; Orengo, C. *Structure (London, England: 1993)* **2009**, *17*, 869–81.

(14) Joachimiak, A. *Curr. Opin. Struct. Biol.* **2009**, *19*, 573–84.

(15) Svergun, D. I. *Bio. Chem.* **2010**, *391*, 737–43.

(16) Montelione, G. T.; Szyperski, T. *Curr. Opin. Drug Discov. Devel.***2010**, *13*, 335–49.

(17) Palczewski, K. *Science* **2000**, *289*, 739–745.

(18) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucl. Acids Res.* **2000**, *28*, 235–42.

(19) Congreve, M.; Langmead, C. J.; Mason, J. S.; Marshall, F. H. *J. Med. Chem.* **2011**, *54*, 4283–311.

(20) Granier, S.; Manglik, A.; Kruse, A. C.; Kobilka, T. S.; Thian, F. S.; Weis, W. I.; Kobilka, B. K. *Nature* **2012**, *485*, 400–4.

(21) Nair, R.; Liu, J.; Soong, T.-T.; Acton, T. B.; Everett, J. K.; Kouranov, A.; Fiser, A.; Godzik, A.; Jaroszewski, L.; Orengo, C.; Montelione, G. T.; Rost, B. *J. Struct. Funct. Genomics* **2009**, *10*, 181–91.

(22) Xie, L.; Xie, L.; Bourne, P. E. *Curr. Opin. Struct. Biol.* **2011**, *21*, 189–99.

(23) Rognan, D. *Molecular Informatics* **2010**, *29*, 176–187.

(24) Rognan, D. *Br. J. Pharmacol.* **2007**, *152*, 38–52.

(25) Kinnings, S. L.; Liu, N.; Buchmeier, N.; Tonge, P. J.; Xie, L.; Bourne, P. E. *PLoS Comput. Biol.* **2009**, *5*, e1000423.

(26) Stauch, B.; Hofmann, H.; Perkovic, M.; Weisel, M.; Kopietz, F.; Cichutek, K.; Münk, C.; Schneider, G. *P.N.A.S* **2009**, *106*, 12079–84.

(27) Defranchi, E.; De Franchi, E.; Schalon, C.; Messa, M.; Onofri, F.; Benfenati, F.; Rognan, D. *PloS one* **2010**, *5*, e12214.

(28) Xie, L.; Wang, J.; Bourne, P. E. *PLoS Comput. Biol.* **2007**, *3*, e217.

(29) Kellenberger, E.; Schalon, C.; Rognan, D. *Current Computer Aided-Drug Design* **2008**, *4*, 209–220.

(30) Weill, N.; Rognan, D. *J. Chem. Inf. Model.* **2010**, *50*, 123–35.

(31) Yeturu, K.; Chandra, N. *BMC bioinformatics* **2008**, *9*, 543.

(32) Yin, S.; Proctor, E. a; Lugovskoy, A. a; Dokholyan, N. V *P.N.A.S* **2009**, *106*, 16622–6.

(33) Das, S.; Kokardekar, A.; Breneman, C. M. *J. Chem. Inf. Model.* **2009**, *49*, 2863–72.

(34) Xiong, B.; Wu, J.; Burk, D. L.; Xue, M.; Jiang, H.; Shen, J. *BMC bioinformatics* **2010**, *11*, 47.

(35) Grant, J. A.; Gallardo, M. a.; Pickup, B. T. *J. Comput. Chem.* **1996**, *17*, 1653–1666.

(36) Marcou, G.; Rognan, D. *J. Chem. Inf. Model.* **2007**, *47*, 195–207.

(37) Meslamani, J.; Rognan, D.; Kellenberger, E. *Bioinformatics (Oxford, England)* **2011**, *27*, 1324–6.

(38) Joachims, T. *Learning to Classify Text Using Support Vector Machines: Methods, Theory, and Algorithms*; LLC, Springer-Verlag: New York, 2002.

(39) Schrodinger Inc SiteMap **2011**.

(40) Tripos SybylX **2011**.

(41) Sofware, O. S. OEChem and OEShape toolkit **2011**.

(42) Grant, J. A.; Pickup, B. T. *J. Phys. Chem.* **1995**, *99*, 3503–3510.

(43) Nicholls, A.; Grant, J. A. *J. Comput. Aided Mol.* **2005**, *19*, 661–86.

(44) Triballeau, N.; Acher, F.; Brabet, I.; Pin, J.-P.; Bertrand, H.-O. *J. Med. Chem.* **2005**, *48*, 2534–47.

(45) Orengo, C. a; Michie, a D.; Jones, S.; Jones, D. T.; Swindells, M. B.; Thornton, J. M. *Structure (London, England: 1993)* **1997**, *5*, 1093–108.

(46) Igarashi, Y.; Eroshkin, A.; Gramatikova, S.; Gramatikoff, K.; Zhang, Y.; Smith, J. W.; Osterman, A. L.; Godzik, A. *Nucl. Acids Res.* **2007**, *35*, D546–9.

(47) Mullan, L. J.; Bleasby, A. J. *Brief Bioinformation* **2002**, *3*, 92–94.

(48) Rost, B. *Protein engineering* **1999**, *12*, 85–94.

(49) Shindyalov, I. N.; Bourne, P. E. *Protein engineering* **1998**, *11*, 739–47.

(50) Schalon, C.; Surgand, J.-S.; Kellenberger, E.; Rognan, D. *Proteins* **2008**, *71*, 1755–78.

(51) Cytoscape:An Open Source Platform for Complex Network Analysis and Visualization.

(52) Nayal, M.; Honig, B. *Proteins* **2006**, *63*, 892–906.

(53) Edfeldt, F. N. B.; Folmer, R. H. a; Breeze, A. L. *Drug Discov. Today* **2011**, *16*, 284–7.

(54) Sheridan, R. P.; Maiorov, V. N.; Holloway, M. K.; Cornell, W. D.; Gao, Y.-D. *J. Chem. Inf. Model.* **2010**, *50*, 2029–40.

(55) Volkamer, A.; Kuhn, D.; Grombacher, T.; Rippmann, F.; Rarey, M. *J. Chem. Inf. Model.* **2012**, *52*, 360–72.

(56) Le Guilloux, V.; Schmidtke, P.; Tuffery, P. *BMC bioinformatics* **2009**, *10*, 168.

(57) Kahraman, A.; Morris, R. J.; Laskowski, R. a; Thornton, J. M. *J. Mol. Biol.* **2007**, *368*, 283–301.

(58) Levitt, D. G.; Banaszak, L. J. *J. Mol. Graph.***1992**, *10*, 229–34.

(59) Connolly, M. L. *J. Appl. Cryst.* **1983**, *16*, 548–558.

(60) Laskowski, R. a *J. Mol. Graph.***1995**, *13*, 323–30, 307–8.

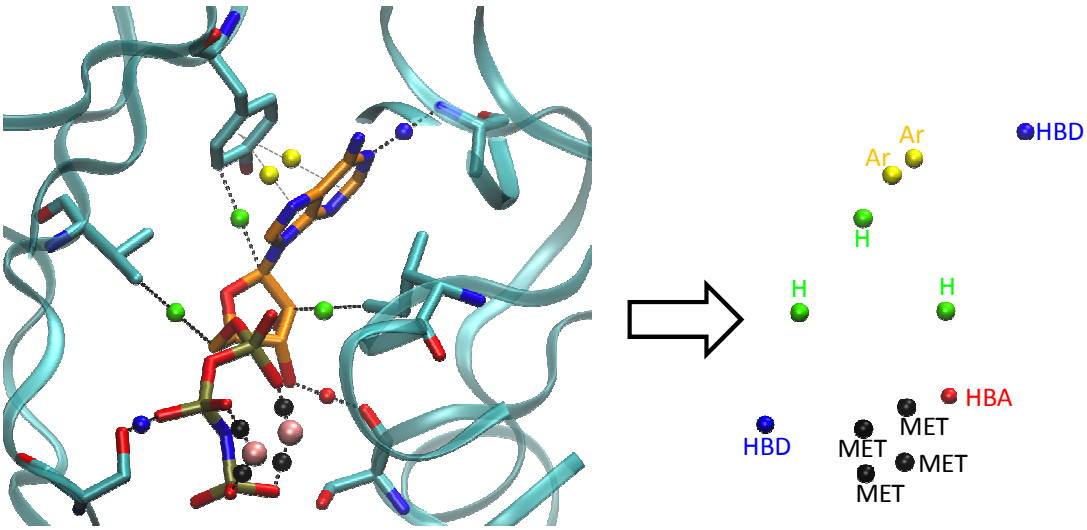
(61) Hendlich, M.; Rippmann, F.; Barnickel, G. *J. Mol. Graph.& modelling* **1997**, *15*, 359–63, 389.

(62) Brady, G. P.; Stouten, P. F. *J. Comput. Aided Mol.* **2000**, *14*, 383–401.

(63) Tripathi, A.; Kellogg, G. E. *Proteins* **2010**, *78*, 825–42.

(64) Weisel, M.; Proschak, E.; Schneider, G. *Chemistry Central journal* **2007**, *1*, 7.

Chapitre 3
Comparaison de modes d'interaction protéine/ligand



Ce chapitre a fait l'objet d'une publication :

Encoding Protein–Ligand Interaction Patterns in Fingerprints and Graphs.

Jérémy DESAPHY, Eric RAIMBAUD, Pierre DUCROT, Didier ROGNAN

Journal of Chemical Information and Modelling, **2013**,53 (3), pp 623-647

1. CONTEXTE

L'obtention d'un ligand de haute affinité pour une cible est réalisable lorsqu'il existe une complémentarité de forme et des propriétés physico-chimiques entre les deux protagonistes. De nombreuses méthodes existantes tentent, et réussissent, à sélectionner des molécules de faible poids moléculaire répondant à ces deux critères. Cependant, rares sont celles qui sont expérimentalement actives. Pour pallier à ce problème, l'une des solutions est d'utiliser les connaissances existantes des molécules actives connues pour la cible d'intérêt. En recherchant des molécules reproduisant le même mode d'interaction que celles actives, on obtient ainsi une relative assurance de la conservation de l'activité, ou tout du moins de réduire le nombre de molécules inactives expérimentalement. On suppose donc que les modes d'interactions sont globalement maintenus pour un même site actif avec des ligands plus ou moins différents. Afin de réaliser cette tâche, il est nécessaire de développer une méthode prenant en compte les deux règles de complémentarité.

La comparaison, à travers des chaines d'entiers, de modes d'interactions entre divers complexes protéine/ligand, ou dans le cadre de criblage virtuel est un domaine de recherche bien entamé qui a commencé en 2004 avec l'apparition de la méthode SIFt (Structural Interaction Fingerprint).[1] Cependant, celui-ci représente les interactions d'un point du site actif au moyen de 7 nombres binaires par résidu mais souffre d'un manque d'interprétation du à la conversion de l'information tridimensionnelle en 1 seule dimension. D'autres méthodes se concentrent ainsi sur la représentation des interactions d'un point de vue du site actif, d'autres d'un point de vue du ligand, mais toutes convertissent l'information sous la forme de chaines d'entiers ou de bits. De plus, dans le cadre d'une analyse de complexes protéine/ligand, les outils existants ne permettent pas de combiner l'information de la protéine et du ligand, pouvant ainsi fausser l'alignement.

Nous nous sommes donc intéressés à l'élaboration d'une nouvelle représentation des complexes protéine/ligand via leurs interactions. Chacune d'elles est caractérisée par 3 points spécifiques, basés sur le ligand, sur la protéine et le centre de chaque interaction. L'obtention d'une information à 3 niveaux permet ainsi d'être à la fois tolérant vis à vis de variations conformationnelles, d'une différence du nombre d'interactions ou d'une modification structurale du ligand.

Une première partie se concentrera sur la conservation des modes d'interaction de complexe protéine/ligand dans la base de données sc-PDB, d'un point de vue général. On tentera ainsi à grande échelle, d'observer la relation entre la conservation des modes d'interactions et la conservation des sites actifs et des ligands.

Une seconde partie se focalisera sur la comparaison de complexes protéine/ligand à travers leurs interactions à travers divers cas d'études. L'une des optiques est l'obtention d'un alignement des complexes prenant en compte ainsi toute l'information disponible entre une protéine et un ligand : interactions non covalentes et complémentarité de forme. Cette comparaison permet aussi de sélectionner des molécules dans le cadre d'un criblage virtuel, qui reproduisent le même mode d'interaction que des molécules actives connues.

Enfin, une dernière partie se concentrera sur des parties plus locales des complexes protéine/ligand. Partant toujours de l'hypothèse d'une complémentarité de forme et de propriétés physico-chimiques, il est ainsi possible de fragmenter le ligand afin de rechercher des modes d'interactions conservés pour des fragments afin d'en trouver des bioisostères.

2. INTRODUCTION

Three-dimensional (3D) structures of protein-ligand complexes provide crucial information to better understand molecular rules governing living cells and assist rational drug discovery. If analyzing a few structures at a graphic desktop is now common practice, mining and comparing a large array of protein-ligand complexes requires a simplification of the 3D information. Among the most useful simplification processes for analyzing protein-ligand interactions is the conversion of atomic coordinates into simpler 1D or 2D fingerprints.[2] Fingerprints are easy to generate, manipulate, compare, and therefore enable a systematic analysis of large datasets. They are largely used to describe and compare molecular objects (small molecular weight ligands,[3] pharmacophores,[4] proteins,[5] protein-ligand binding sites[6]) and represent descriptors utilized by computer-aided drug design programs, notably *in silico* screening tools. Computational chemists frequently manipulate these fingerprints independently in either ligand-based or structure-based approaches to drug design.[7] It would, however, be very interesting to combine both protein-based and ligand-based information in a single descriptor focusing on molecular interactions in order to answer the following questions: Do similar binding sites identically recognize similar ligands? Are protein-ligand interaction patterns conserved across target families? Which chemically different ligand structures or substructures share identical interaction patterns with a single target?

Two main approaches to merge ligand-target pairs in a single descriptor have been proposed up to now. The first one qualitatively describes the protein-ligand pair by annotating ligand descriptors with interaction features. For example, Bajorath et al. reported a way to augment classical ligand fingerprints with protein-ligand interaction-derived information.[2,8,9] The basic underlying idea is that all atoms of a bioactive ligand are not equally responsible for its biological activity. Focusing a chemical fingerprint to protein-interacting ligand atoms (interacting fragment or IF) is likely to enhance the value of such a fingerprint by avoiding the possibility to recruit novel compounds by a pure ligand-based virtual screening approach for wrong reasons (atoms/groups not interacting with a protein pocket). Such IF-annotated fingerprints were shown to outperform conventional fingerprints in standard similarity searches to known ligands

of diverse activity classes.[9] Standard descriptors for protein cavities and their cognate ligands can also be concatenated into a single fingerprint[10,11] and then used as input to train machine learning algorithms to discriminate true from false complexes.[12] In a prospective virtual screening study, this kind of fingerprint was found superior to conventional ligand-based descriptors (2D and 3D) in finding novel non-peptide ligands for the oxytocin receptor.[13] However, the latter descriptors do not describe the physical intermolecular interactions (e.g. hydrogen bond, hydrophobic contact) between ligand and target.

The second possible approach to protein-ligand fingerprinting annotates protein descriptors (usually binding site-lining amino acids) with ligand-interaction features. The interaction fingerprint concept (SIFt: Structural Interaction Fingerprint) was pioneered by Biogen Idec.[1] and consists in converting a 3D protein-ligand complex into a 1D bit string registering intermolecular interactions (hydrophobic, hydrogen bonds, ionic interactions) between a ligand and a fixed set of active site residues. Interactions are computed on the fly using standard topological criteria between interacting atoms (distances, angles) and a bit is switched "on" or 'off" as to whether the interaction occurs or not. The SIFt method was originally designed for analyzing ligand docking poses to protein kinases and shown several promising features: (i) enhancing the quality of pose prediction in docking experiments,[1] (ii) clustering protein-ligand interactions for a panel of related inhibitors according to the diversity of their interactions with a target subfamily,[14] (iii) assisting target-biased library design.[15] The interaction fingerprint concept was further developed by other groups in order to define the directionality of the interactions (e.g. H-bonds donated by the ligand and by the active site are stored in distinct bits),[16] the strength of the interaction,[17] or assign a bit to every active site atom instead of every active site residue.[18] Interaction fingerprints (IFPs) are now part of many docking tools in order to post-process docking poses according to known protein-ligand X-ray structures. Remarkably, IFPs show a great scaffold hopping potential in selecting virtual hits sharing the same interaction pattern than a reference ligand, but with different chemotypes.[19] Since a bit is defined for every active site atom/residue, the method is therefore limited to analyze interactions with highly homologous active sites sharing a fixed number of cavity-lining atoms/residues. To overcome this limitation, cavity-independent fingerprints (APIF)[20] do not consider the absolute but only the relative positions of pairs of protein-ligand interacting atoms and store information in a

294-bit fingerprint according to the interaction type and distance between interacting pairs. Like standard IFPs, APIF scoring by comparison to known references was shown to outperform conventional energy-based scoring functions in docking-based virtual screening of compound libraries. Unfortunately, obtained results are difficult to interpret since deconvoluting APIF into specific protein-ligand features or protein-ligand alignments is not possible. Databases of protein-ligand complexes (e.g. CREDO,[21] PROLIX[22]) focusing on observed interactions in X-ray structures have been described and uses fingerprint representations of interaction patterns to retrieve PDB complexes fulfilling user queries (e.g. number and/or type of protein-ligand interactions, interaction to particular amino acids). However, neither a 3D alignment of protein-ligand complexes nor a generic similarity measure between the two complexes to evaluate, are proposed.

The novel protein-ligand descriptors (fingerprint, graph) and comparisons methods (Ishape, Grim) presented in this study were therefore designed to specifically enable the following features: (i) compare protein-ligand interactions whatever the size and sequence of the target binding sites (e.g. across target families), (ii) quantitatively describe molecular interactions with a specific frame-invariant descriptor, (iii) provide an alternative 3D alignment of protein-ligand complexes to protein-based or ligand-based matches, by focusing on molecular interactions only.

It enables an exhaustive and easily interpretable pairwise comparison of all protein-ligand x-ray structures that may be used in several scenarios: post-process protein-ligand docking poses, find off-targets sharing key interaction patterns to a known ligand, identify bioisosteric fragments with a conserved interaction mode to a given target.

3. METHODS

3.1 DATASETS OF PROTEIN-LIGAND COMPLEXES

All protein-ligand complexes were retrieved from the sc-PDB dataset[23] which archives 9877 high resolution X-ray structures of druggable protein-ligand complexes. For each complex, protein and ligands were separately stored in TRIPOS mol2 file format.[24] Pairwise sc-PDB ligand similarity was expressed by the Tanimoto coefficient on either circular ECFP4 fingerprints[25] in PipelinePilot[26] or MACCS 166-bit structural keys[27] in MOE.[28]

3.1.1 SET 1:900 SIMILAR AND 900 DISSIMILAR PROTEIN-LIGAND COMPLEXES

Pairs of protein-ligand complexes were considered similar if: (i) their pairwise binding site similarity (expressed by the Shaper similarity score[29]) was higher than 0.44 and ii) their pairwise ligand similarity (expressed by a Tanimoto coefficient on ECFP4 fingerprints) was between 0.55 and 0.75. This selection protocol led to a set of 7426 pairs of similar complexes, out of which 900 pairs were finally selected (**Supplementary Table 1**) by retrieving all possible non-redundant Uniprot names. The same number of pairs of dissimilar protein-ligand complexes was retrieved assuming that their pairwise binding site similarity and their pairwise ligand similarity were lower than 0.20. This cut-off was chosen arbitrarily to be sure that both active sites and bound-ligands were really dissimilar. This selection protocol led to 3524800 pairs of dissimilar complexes, out of which 900 pairs were randomly selected (**Supplementary Table 1**) avoiding duplicates in protein names.

3.1.2 SET 2: SC-PDB FRAGMENTS

The 9877 ligands of the current sc-PDB release were submitted to a retrosynthetic fragmentation protocol using 11 RECAP[30] rules embedded in Pipeline Pilot.[26] About 78% of the ligands (7769) could be fragmented into 20839 building blocks (15828 cyclic, 5011 acyclic). Each fragment was annotated with descriptors from its parent ligand (HET identifier, fragment number) and PDB target (Uniprot target name, KEGG BRITE functional class[31]). PDB codes and HET codes of the fragments are given in **Supplementary Table 2**.

3.1.3 SET 3 : CCDC/ASTEX SUBSET OF PROTEIN-LIGAND COMPLEXES

The CCDC/Astex set of 95 clean high-resolution (< 2.0 Å) protein-ligand X-ray structures[32] was downloaded from the Cambridge Crystallographic Data Centre.[33] 45 entries, already present in the sc-PDB were discarded, 14 additional entries were removed since the corresponding target (mainly antibodies) was completely absent in the sc-PDB thus leading to a final set of 36 protein-ligand complexes (**Supplementary Table 3**). Hydrogen atoms were added in SYBYL[24] and their atomic coordinates changed to manually optimize protein-ligand interactions first and intramolecular interactions in a second step. Ionization and tautomeric states of cavity-lining residues were accordingly updated whenever necessary. Proteins and ligands were separately stored in ready-to-dock file mol2 file format.

3.1.4 SET 4 : DUD-E TARGET AND LIGAND SETS

Active and decoy ligand sets for 10 targets covering 5 different protein families: G protein-coupled receptors (adenosine A2A receptor: AA2AR, adrenergic beta2 receptor: ADRB2), nuclear hormone receptors (androgen receptor: AND, glucocorticoid receptor: GCR); other enzymes (adenosine deaminase: ADA, prostaglandin G/H synthase 2: PGH2); proteases (angiotensin-converting enzyme: ACE, renin: reni); protein kinases (fibroblast growth factor receptor 1: FGFR1, RAC-alpha serine/threonine-protein kinase: AKT1) were downloaded in 3D mol2 file format from the DUD-E[34] website (http:// http://dude.docking.org/). Since several pdb entries selected as DUD-E atomic coordinates for the host protein were already present in the sc-PDB and that one of our rescoring procedure (Grim) relies on existing protein-ligand complexes, PDB entries not present in the sc-PDB and of the highest possible resolution (2pwh, 2am9, 1p93, 1a4l, 3zqz, 3sfc, 3tt0, 4ekl) were selected as host coordinates for docking. For two targets (ADRB2, PGH2), we decided to keep the original DUD-E PDB entry (3ny8, 3nt1) but removed it from the set of references for further scoring.

3.2 DETECTION OF PROTEIN-LIGAND INTERACTIONS

7 pharmacophoric properties (hydrophobic, aromatic, H-bond donor, H-bond acceptor, positive ionizable, negative ionizable, metal; **Supplementary Table 4**) for protein and ligand atoms are assigned by parsing the atom and bond connectivity fields of the mol2 files. Protein-ligand interactions are then detected on the fly with respect to the above-defined pharmacophoric types and previously defined topological criteria[16] (**Supplementary Table 5**). The protein-ligand interaction (**Figure 1A**) is characterized by the two interacting atoms and an interaction pseudoatom (IPA) located at three possible positions: (i) the geometric center of interacting atoms (*Centered mode*), (ii) the interacting protein atom (*InterProt mode*), (iii) the interacting ligand atom (*InterLig mode*). IPAs can be computed using one of the three possible modes to enable a mapping of the interaction on either ligand or protein atoms (InterLig and InterProt modes, respectively) or more naturally at the mid-distance of interacting atoms (Centered mode).

Since hydrophobic atoms are by far the most frequent, two additional rules have been defined to limit hydrophobic IPAs (**Figure 1B**). First, only one IPA can be generated between a single ligand atom and a single protein amino acid, whatever the number of interacting atoms. In case a ligand atom verifies the condition of a hydrophobic interaction with two different atoms of the same protein residue, only the shortest distance is kept to assign the corresponding IPA. All hydrophobic IPAs are then iteratively clustered with a hierarchical agglomerative clustering using a 1.0 Å distance criterion and an average linkage method. Second, two aromatic rings not engaged in an aromatic interaction (edge-to-face or face-to-face) will define a hydrophobic interaction between their two closest atoms at the condition that the corresponding hydrophobic distance rule is verified (**Figure 1C**). When the aromatic interaction condition is verified, an aromatic IPA is set between centroids of the corresponding aromatic rings. In *InterLig* mode only,

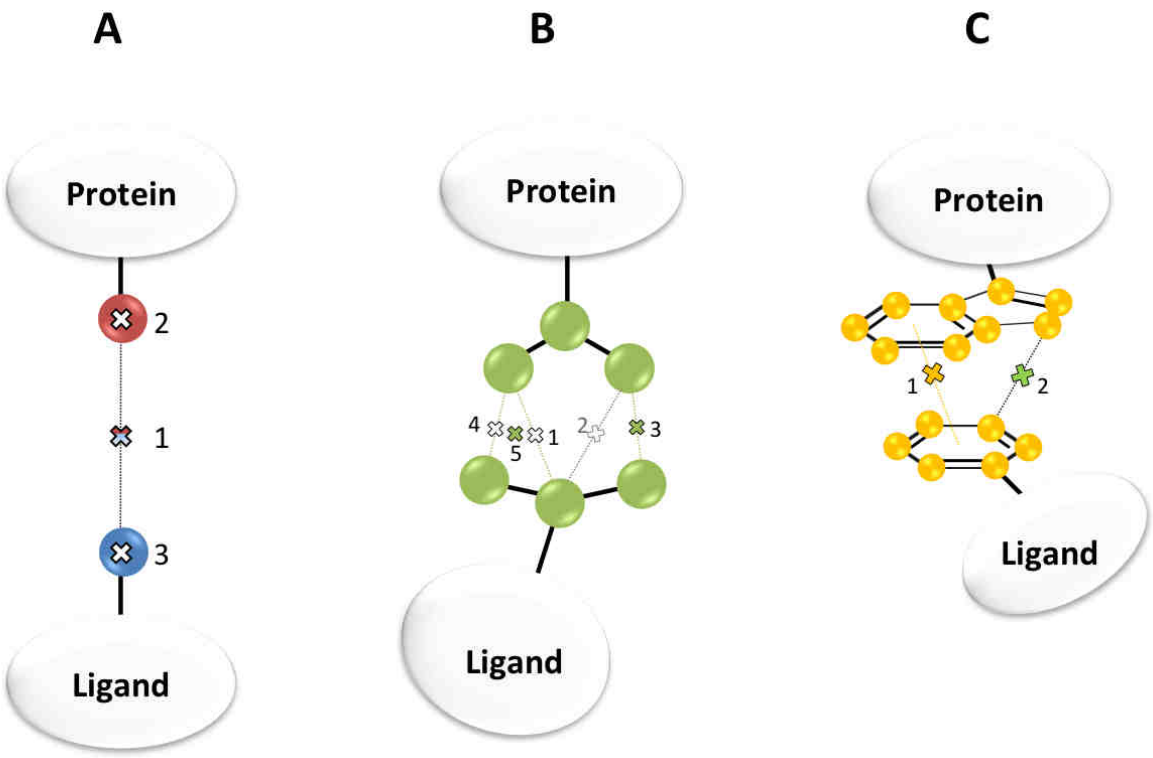
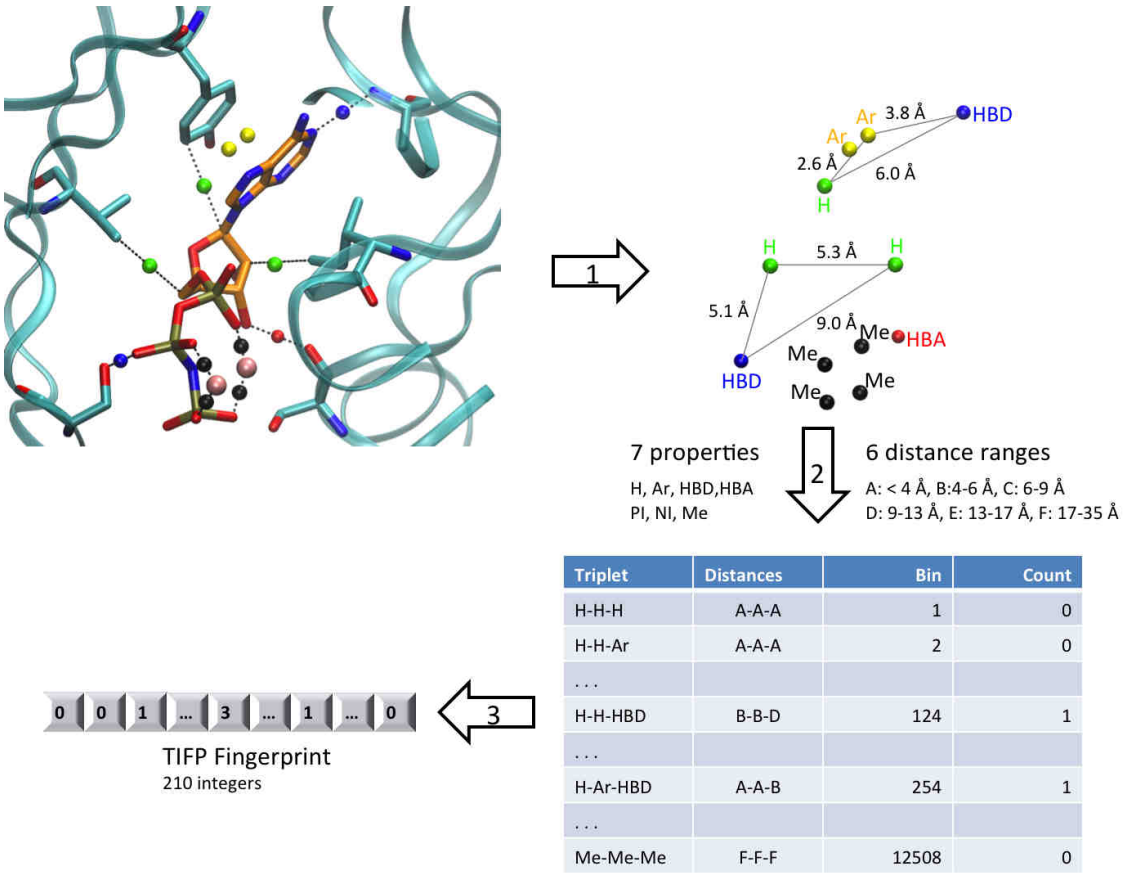


Figure 1- Definition of interaction pseudoatoms (IPAs). (A) Case of hydrogen bonding, ionic interaction, and metal complexation. Two atoms of complementary pharmacophoric properties (red and blue balls) fulfilling interaction rules describe an interaction (dotted line) at a pseudoatom (cross) located at three possible positions: (i) geometric center of interacting atoms (1), (ii) interacting protein atom (2), (iii) interacting ligand atom (3). (B) Case of hydrophobic interactions between protein and ligand hydrophobic atoms (green balls). Only the shortest interaction (number 1) between a single ligand atom and many protein atoms is kept (interaction 2 is not conserved). Remaining hydrophobic IPAs (1, 3, 4) are then clustered to yield IPAs 3 and 5. (C) Case of aromatic interactions between protein and ligand aromatic atoms (yellow balls). Protein and ligand aromatic atoms verifying the aromatic interaction rules define an aromatic interaction (yellow dotted line) with an aromatic IPA set at the mid-distance between both aromatic ring centroids (IPA 1). For aromatic interacting atoms, not respecting the aromatic interaction rule but fulfilling a hydrophobic interaction, a hydrophobic interaction (green dotted line) and IPA (IPA 2) is defined between the closest protein and ligand atoms.

a last pruning step avoids property redundancy on ligand atoms by keeping a single occurrence of interaction type per ligand atom. For example, if a ligand atom makes one aromatic and two hydrophobic interactions, only one aromatic and one hydrophobic *InterLig* IPA are created. Finally, aromatic edge-to-face and face-to-face interactions are merged to represent only one aromatic IPA. When all interactions have been detected, IPAs are exported in a mol2 file format.

3.3 FINGERPRINTING TRIPLETS OF INTERACTION PSEUDOATOMS

Starting from a set of IPAs (whatever the mapping mode), the TIFP fingerprint encodes protein-ligand interactions by a vector of 210 integers (**Figure 2**). Each integer of the vector registers the count of unique IPA triplets (7 properties and 3 related distances) occurring at binned inter-feature distances. Please note that we only consider 7 properties since the two aromatic interactions (edge to face, face-to face) are merged. The distances between IPAs are currently discretized in 6 intervals ([0-4 Å[, [4-6 Å[, [6-9 Å[, [9-13 Å[, [13-17 Å[, [17 + Å[).



Triplet	Distances	Bin	Count
H-H-H	A-A-A	1	0
H-H-Ar	A-A-A	2	0
. . .			
H-H-HBD	B-B-D	124	1
. . .			
H-Ar-HBD	A-A-B	254	1
. . .			
Me-Me-Me	F-F-F	12508	0

Figure 2 - Generating a fingerprint of IPA triplets. From atomic coordinates of the protein–ligand complex (PDB code: 1j7u), interactions are detected on the fly and described by IPAs featuring seven interaction types (hydrophobic, H, green balls; aromatic, Ar, yellow balls; H-bond donor, HBD, blue balls; H-bond acceptor, HBA, red balls; positive ionizable, PI, not featured here; negative ionizable, NI, not featured here; metal complexation, Me, dark). All possible triplets (three properties, three distances) of IPAs are generated (step 1) and matched to a triplet list (step 2). The count of each triplet type is encoded through a fingerprint of 12 508 integers that is further pruned to 210 integers (step 3) to feature the most frequently occurring triplets.

Starting from the first interval, all triplet combinations are counted and stored, until the last interval is processed. Given 7 pharmacophoric types and 6 distance ranges, the total number of triplets is thus equal to 7^3 x 6^3 = 74088. To generate the shortest possible fingerprint, redundant triplets (property redundancy, isoscele and equilateral triangles) are removed. Last, the geometrical validity of the pharmacophoric triplet is checked by applying the triangle inequality rule stating that one distance cannot be longer than the sum of the two others. The full fingerprint accounts for 18179 possible triplets stored in 12508 bins. The higher number of triplets with regards to bins is simply due to the redundancy which is observed at the triplet (e.g. ABC and BAC triplets) and not at the bin level (both triplets could be assigned to a single bin).

To speed up fingerprint calculations and comparisons, a two-step compression was done as follows. Full length fingerprints were computed for the 9877 protein-ligand complexes of the sc-PDB dataset (2011 release) and the count status of every triplet was computed. First, 17612 weakly populated triplets corresponding to 12119 bins (count < 10) were removed. Second, 361 triplets (226 bins) with a frequency between 10 and 20 were merged into 185 triplets (46 bins) of the same composition but with no distance information. The remaining 206 triplets (164 bins), with a frequency higher than 20, were kept unchanged. The final size of the compressed TIFP fingerprint amounts to 391 triplets in 210 separate bins. The similarity between two TIFP fingerprints was expressed by a Tanimoto coefficient as follows:[35]

$$Tc = \frac{\sum_{j=1}^{N} x_{jA} x_{jB}}{\sum_{j=1}^{N} (x_{jA})^2 + \sum_{j=1}^{N} (x_{jB})^2 - \sum_{j=1}^{N} x_{jA} x_{jB}}$$

where x_{jA} is the value of the bin j in the reference fingerprint A and x_{jB} the value of the bin j in the comparison fingerprint B.

3.4 SHAPE MATCHING OF IPAS (ISHAPE)

IShape uses an algorithm very similar to that recently described in Shaper, a tool to align protein-ligand binding sites.[29] It relies on OEChem and OEShape toolkits[36] which present the advantage to describe molecular shapes by a smooth Gaussian function and to align two molecular objects (IPAs) by optimizing the overlap of their corresponding volumes.[37–39] During the alignment, a reference IPA set (*Centered* mode only) is kept

rigid while the set of IPAs to fit (fit object) undergoes rigid body rotations and translations. To speed-up calculations, the 'Grid' volume overlap method was chosen to represent the volume of the target IPAs and all atom radii were set to that of carbon (1.7 Å). Once the best shape alignment has been achieved, it is scored by a 'Color Force Field' (a color being a pharmacophoric feature) similar to that used by the ligand matching tool ROCS[36] to account for pharmacophoric properties matching. In other words, the alignment proposed by the simple shape matching is scored to account for pharmacophoric feature superposition.

The force field (**Supplementary Table 6**) consists in SMARTS patterns for 7 pharmacophoric properties (H, hydrophobic; Ar, Aromatic; HBA, H-bond acceptor; HBD, H-bond donor; A-, negative ionizable; D+, positive ionizable; Me, metal complexation) and 7 pattern matching rules (H to H, Ar to Ar, HBA to HBA, HBD to HBD, A-to A-, D+ to D+, Me to Me) to score the shape-based alignment by pharmacophoric similarity. Color matches were considered for interaction points up to 1.5 Å apart with a single weight for all matching rules. The similarity SimA,B between IPAs A (reference) and B (fit) was calculated by a Tversky index as follows:

$$\mathrm{Sim}_{A,B} = \frac{O_{A,B}}{0.95\, I_A + 0.05\, I_B + O_{A,B}}$$

where $O_{A,B}$ is the overlap between colors of IPAs A and B, and I non-overlapped colors of each entity A and B. Classification models based on pairwise similarity values were assessed by computing the area under the receiver operating characteristic (ROC) curve,[40] and the F-measure as follows:

$$\mathrm{Recall} = \frac{TP}{TP+FN} \qquad \mathrm{Precision} = \frac{TP}{TP+FP} \qquad F-\mathrm{Measure} = \frac{2*(\mathrm{recall})(\mathrm{precision})}{\mathrm{precision}+\mathrm{recall}}$$

where TP are true positives, FN false negatives, FP false positives and FN false negatives. The best similarity threshold is found by the maximum of the F-measure curve when the threshold was varied from 0 to 1 with an increment of 0.01. ROC and Boltzmann enhanced discrimination of ROC (BEDROC) curves were computed using the CROC program.[41] For computing BEDROC values, the alpha-parameter was set to the default value of 20.

3.5 GRAPH MATCHING OF IPAS (GRIM)

At the noticeable difference of IPA fingerprinting, Grim matching simultaneously considers the three IPA definition modes (InterLig, InterProt, Centered) in order to take into account the ligand, the protein and the protein-ligand interactions at the same time. Starting from two sets of IPAs (reference, target), Grim first creates a list of possible IPA matches. A match is done if reference and target IPAs have the same label (same interaction type) and represent the same 'point of view', i.e. *Centered*, *InterLig* or *InterProt*. For example, centered IPAs of reference and target will be compared but centered IPAs of reference and InterLig IPAs of the target will not. A product graph is then created from the reference and target graphs in which each successfully matched pair defines consequently a vertex. A weight is added to vertex which is inversely proportional to the observed frequency among the 284186 IPAs generated from the 9877 protein-ligand complexes of the sc-PDB dataset. Assigned weights were as follows: hydrophobic IPA (0.299), aromatic IPA (0.990), H-bond acceptor (0.930), H-bond donor (0.834), negative ionizable (0.993), positive ionizable (0.966), metal complexation (0.985). It should be recalled that the different weights assigned to hydrogen bonds (acceptor vs. donor) comes from a previously observed bias in the sc-PDB in which donors occur more frequently from protein than from ligand atoms.[42]

An edge is observed between two vertices of the product graph after computing distances between the two reference IPAs and the two target IPAs. If the difference is below a given threshold (**Supplementary Table 7)** an edge is created. Hence, the largest cliques are detected using the Bron-Kerbosch algorithm[43] with pivoting and pruning improvements.[44] Each IPA of the target is matched with the corresponding reference IPA using a quaternion-based characteristic polynomial.[45] It returns both the translation vector and the rotation matrices to match target and reference graphs as well as a Graph-alignment score (Grscore). Cliques are then scored by decreasing Grscore, a score which was empirically determined by fitting six Grim parameters to the previously described IShape similarity score on the set of 1800 protein-ligand complexes (900 similar and 900 dissimilar) as follows:

GrScore=	0.5006		
	+0.0151	NLig	number of matched InterLig IPAs
	+0.0039	NCenter	number of matched Centered IPAs
	+0.0143	NProt	number of matched InterProt IPAs
	+0.2098	SumCl	$\frac{\sum \text{pair weights in clique}}{\sum \text{all possible pair weights}}$
	-0.0720	RMSD	root-mean square deviation of the matched clique
	-0.0003	DiffI	absolute value of the difference in the number of IPAs between reference and query

3.6 DOCKING

All ligands from sets 2-4 were docked into their original X-ray structure with the Surflex-Dock (v.2601) software.[46] Protomols were first generated from the list of cavity-lining residues defined as any amino acid within a 6.5 Å-radius sphere centered on the bound-ligand center of mass. Compounds were then docked with default settings (excepted for the "–pgeom" option) of the docking engine keeping the best 20 poses according to the native Surflex-Dock scoring function. All poses were finally re-ranked by decreasing Tanimoto similarity value of the TIFP and decreasing Grcore to either the native X-ray pose (self-docking) or any sc-PDB pose with the same sc-PDB target name (target-directed docking).

4. RESULTS AND DISCUSSION

4.1 FINGERPRINTING INTERACTION PATTERNS IN SC-PDB COMPLEXES

All 9877 protein-ligand complexes of the sc-PDB were parsed to detect protein-ligand interactions and defined 284186 interaction pseudo-atoms in total (**Table 1**). As to be expected, about 70% of these interactions represent apolar hydrophobic contacts (**Figure 3**). Due to the prevalence of anionic compounds (mostly nucleotides) among sc-PDB ligands, H-bonds donated by protein atoms are significantly more abundant than H-bonds donated by ligands (16% vs. 6%, respectively). The same statistical discrepancy also occurs when comparing salt bridges (3% with a protein cation, 1% with a protein anion; **Figure 3**). For a protein-ligand complex, the number of interactions goes from 2 (*E. coli* dihydroorotate dehydrogenase in complex with orotic acid, PDB code: 1f76) to 78 (orange carotenoid protein R155L mutant in complex with beta-caroten-4-one; PDB code: 3mg3). A protein makes in average 28 +/- 10 interactions with its ligand.

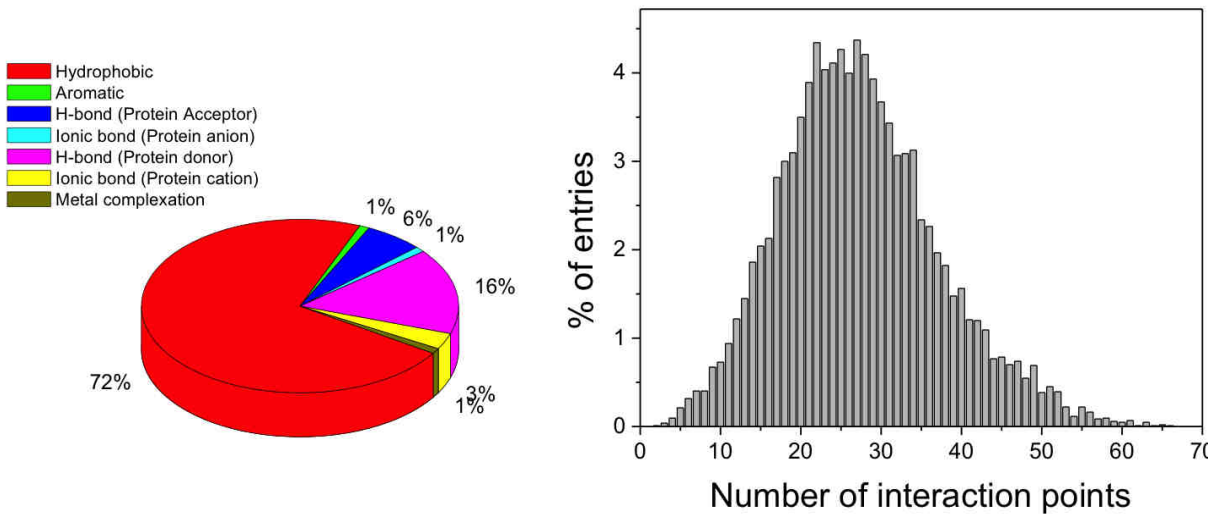


Figure 3 - Analysis of IPAs from 9877 protein-complexes of the sc-PDB dataset. **A)** Distribution of interaction types, **B)** Distribution of interaction points.

However, when positioning the pseudoatoms on the ligand (*InterLig* mode), fewer interactions are output (182262 in total, **Table 1**) due to the additional filtering process of hydrophobic IPAs (see Methods). Using the *InterLig* mode, a sc-PDB complex has on average 18 +/- 6 IPAs, with a minimum of 1 and a maximum of 44.

Table 1 - Pharmacophoric property distribution of interactions in the sc-PDB dataset (n= 9877)

Position	Tot[a]	Hyd[b]	Ar[c]	HBA[d]	NI[e]	HBD[f]	PI[g]	Me[h]
Centered	284 186	72%	1%	6%	1%	16%	3%	1%
Ligand	182 262	66%	1%	8%	1%	18%	4%	2%
Protein	284 186	72%	1%	6%	1%	16%	3%	1%

[a] total number of protein-ligand interacting pseudoatoms (IPAs)
[b] IPA with hydrophobic property
[c] IPA with aromatic property
[d] IPA with H-bond acceptor property (from protein side)
[e] IPA with negative ionizable property (from protein side)
[f] IPA with H-bond donor property (from protein side)
[g] IPA with positive ionizable property (from protein side)
[h] IPA with metal complexation property

4.2 A RELIABLE SIMILARITY METRIC TO COMPARE PROTEIN-LIGAND INTERACTION PATTERNS

A dataset of 900 pairs of similar protein-ligand complexes and 900 pairs of dissimilar complexes (see Methods) was set-up to investigate the possibility to quantitatively assess the similarity of protein-ligand complexes from either TIFP fingerprints (alignment-free comparison) or the set of matched IPAs (Ishape: shape-based alignment, Grim: graph-based alignment). First, a binary classification of all 1800 pairs using a ROC curve analysis of three possible lists, ordered by decreasing Tc (TIFP), Sim (IShape) and Grscore (Grim) values, clearly indicates that all three metrics discriminate similar from dissimilar complexes (**Figure 4A, Table 2**). Almost perfect area under the ROC curves are obtained with either Grim or IShape comparisons (0.96 and 0.95, respectively). The corresponding value using the simple TIFP fingerprint similarity is still very high but significantly lower (0.90) and therefore indicates some limited but true noise in the fingerprints that does not exist in IPAs themselves. Plotting the performance of a binary classification model (complexes are either predicted similar or dissimilar) as a function of the similarity threshold used for deciding upon similarity, gives an explanation for the higher permissivity behavior of the TIFP score. Hence, the

corresponding classification models exhibit recall values slowly decaying and precisions slowly increasing when the similarity Tc value increases (**Figure 4B**). The gap between the similarity value at the highest F-measure (Tc=0.318) and that at the maximal precision (Tc=0.911) is very large. Conversely, using the Grscore as a metric of complex similarity produces classification models whose performance (recall, precision, F-measure) are optimal in a very narrow similarity threshold window (0.59 <Grscore< 0.65,**Figure 4D**) and therefore more robust. In between, the IShape similarity score varies between 0.41 at the F-measure optimum and 0.63 at the precision optimum (**Figure 4C**).

The greater noise in the TIFP fingerprint comparison with respect to either shape or graph matching arises from two main reasons (i) the information loss upon converting 3D information into 1D data, (ii) the relative importance of hydrophobic triplets in TIFP fingerprints (with either 2 or 3 features) which is minored in the graph alignment-based scores. Hence, our clique detection method uses a weight on pharmacophoric features which is inversely proportional to their abundance in the sc-PDB (hydrophobes are less important than polar features in the clique ranking).

Current benchmarks on a Intel® Core™2 Duo E8500 processor (3.16 GHz, 6 M cache) indicate that all three comparison methods are fast enough (10, 20 and 35 ms/comparison for TIFP, Ishape and Grim, respectively) to be applied to large scale comparisons.

Table 2 - Quantitative estimation of pairwise similarity for 1800 pairs (900 similar, 900 dissimilar) of protein-ligand complexes

Statistics	TIFP[a]	IShape[a]	Grim[a]
AUROC[b]	0.908	0.959	0.954
BEDROC[c]	0.973	0.999	0.999
Best threshold[d]	0.318	0.407	0.594
F-measure[e]	0.830	0.892	0.909
Precision100% [f]	0.911	0.627	0.659

[a] similarity measured by the Tanimoto coefficient from TIFP fingerprints (TIFP), the IShape Sim similarity index (IShape) and the Grscore (Grim).
[b] area under the ROC curve for a binary classification (similar, dissimilar)
[c] Boltzmann enhanced discrimination of ROC
[d] Similarity score enabling the best possible classification
[e] F-measure of the best classification model
[f] Similarity score enabling a perfect classification (precision of 100%)

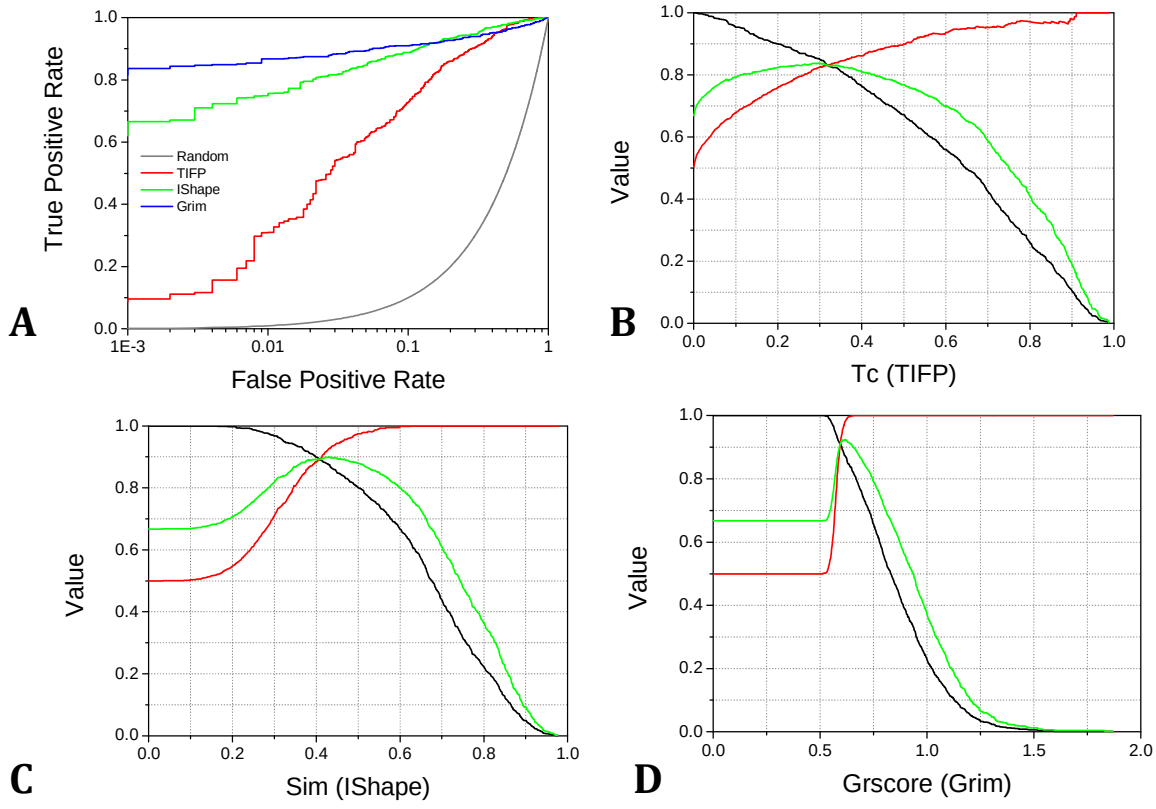


Figure 4 - Statistical evaluation of protein-ligand complex similarity scores. **A)** ROC plot obtained by sorting 1800 sc-PDB pairs of complexes by decreasing similarity values (red, TIFP Tc score; green, Ishape Sim score; blue, Grim Grscore). True positives are pairs of similar protein-ligand complexes predicted similar whereas false positives are pairs of dissimilar complexes predicted similar. Accuracy of random picking is represented by a gray line. **B-D)** Variation of statistical parameters (recall, dark line; precision, red line; F-measure, green line) for a binary classification model (similar/dissimilar) of all 1800 pairs, according to the TIFP similarity threshold value (panel B), IShape similarity score (panel C) and Grim Grscore (panel D).

4.3 INTERACTION PATTERN SIMILARITY DEPENDS TIGHTLY ON BINDING SITE SIMILARITY

The above-described similarity offers us the opportunity to check across all sc-PDB complexes a possible dependence between the similarity of protein-ligand interactions and the corresponding ligand and/or protein binding site similarities. The similarity of two protein-ligand complexes was then computed with three different metrics: the pairwise similarity of their ligands (Tanimoto coefficient on ECFP4 and MACCS fingerprints), the pairwise similarity of their binding sites (Shaper similarity), and the pairwise similarity of their interaction patterns (IShape similarity). Ligands were considered similar if their Tanimoto coefficient was above 0.4 for circular ECFP4 fingerprints, or above 0.7 for MACCS keys. Binding sites were considered similar if their pairwise Shaper similarity was above 0.35.[29] Last, interaction patterns were considered similar if their IShape similarity was above 0.41, as previously suggested in **Table 2**. Plotting these three possible values against each other clearly shows that interaction pattern similarity is not related to ligand similarity whatever the descriptor used (**Figure 5A,B**) but strongly correlates with binding site similarity (r=0.876, sd=0.11; **Figure 5C**). As to be expected, there are very few cases of similar interaction patterns between dissimilar ligands and dissimilar binding sites (lower left quadrant,**Figure 5D**). In most of the cases, this relates to small molecular weight ligands exhibiting a simple and promiscuous hydrophobic interaction pattern (see a prototypical example **Figure 6A**). Cases for which similar ligands exhibit similar interaction patterns to dissimilar binding sites are still rare (upper left quadrant, **Figure 5D**). This situation mainly occurs either when one of the two binding sites undergoes a significant change (mutation, monomer vs. dimer-lining interface) without altering ligand recognition, or for primary metabolite-binding sites (e.g. GDP-binding sites, **Figure 6B**) which have evolved to share conserved features even in absence of sequence and fold conservation. Interestingly, as far as binding sites are similar, interaction patterns are conserved irrespectively of the corresponding ligand similarity (in 93 and 88% of cases for similar and dissimilar ligands, respectively; **Figure 5D**). Despite a bias due to the still limited ligand diversity among sc-PDB ligands, this observation suggests that a single interaction mode to a single druggable cavity remains the rule because a few key

interactions to a few key residues need to be fulfilled to achieve significant binding. Careful inspection of ligand structures revealed that dissimilar ligands sharing both a conserved cavity and interaction pattern are usually sharing a common substructure, which is the main anchoring moiety to their target (e.g. trypsin inhibitor binding to the catalytic site, **Figure 6C**).

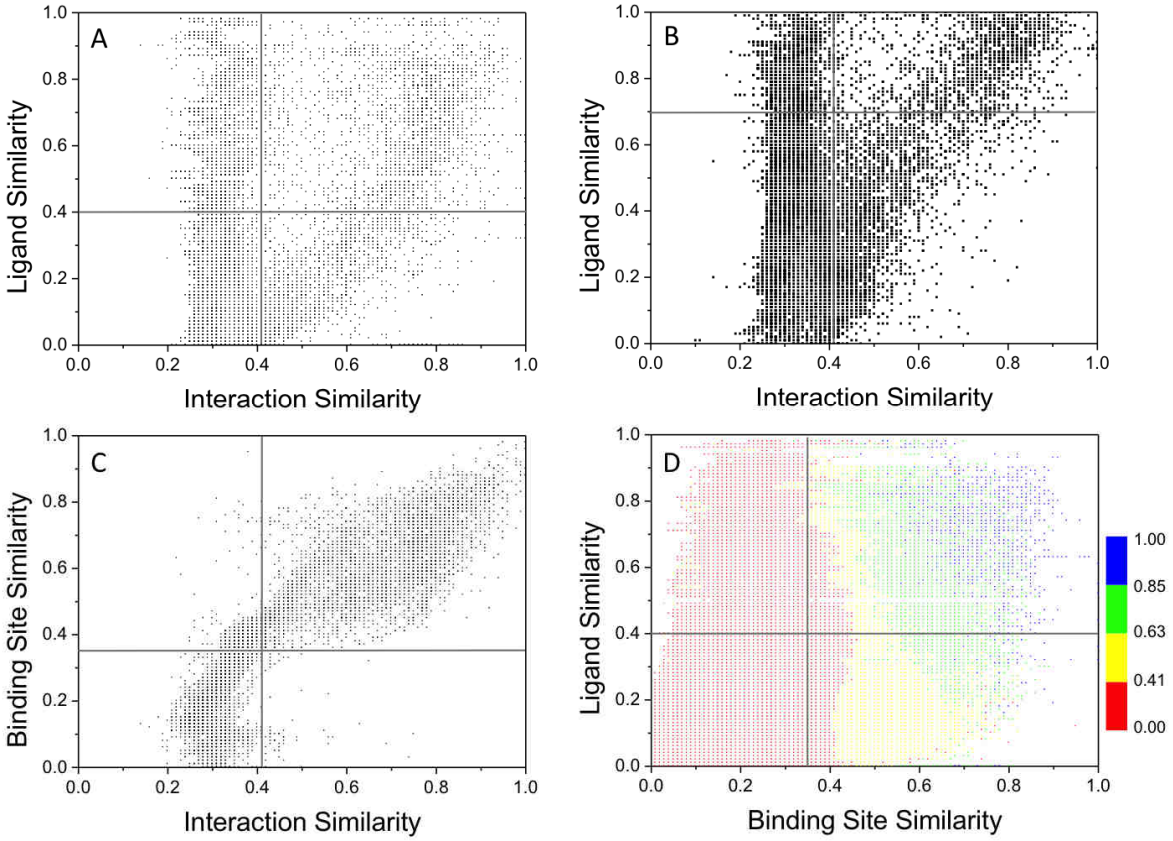


Figure 5 - Relationships between ligand similarity, binding site similarity and interaction pattern similarity for 9877 sc-PDB entries. **A)** Ligand similarity (Tanimoto coefficient on ECFP4 fingerprints) vs. interaction pattern similarity (IShape similarity score). B) Ligand similarity (Tanimoto coefficient on MACCS public keys) vs. interaction pattern similarity (IShape similarity score). **C)** Binding site similarity (Shaper[29] similarity score) vs. interaction pattern similarity (Ishape similarity score). **D)** Ligand similarity (Tanimoto coefficient on ECFP4 fingerprints) vs. binding site similarity (Shaper[29] similarity score). Data are colored according to the interaction pattern similarity score (IShape similarity).

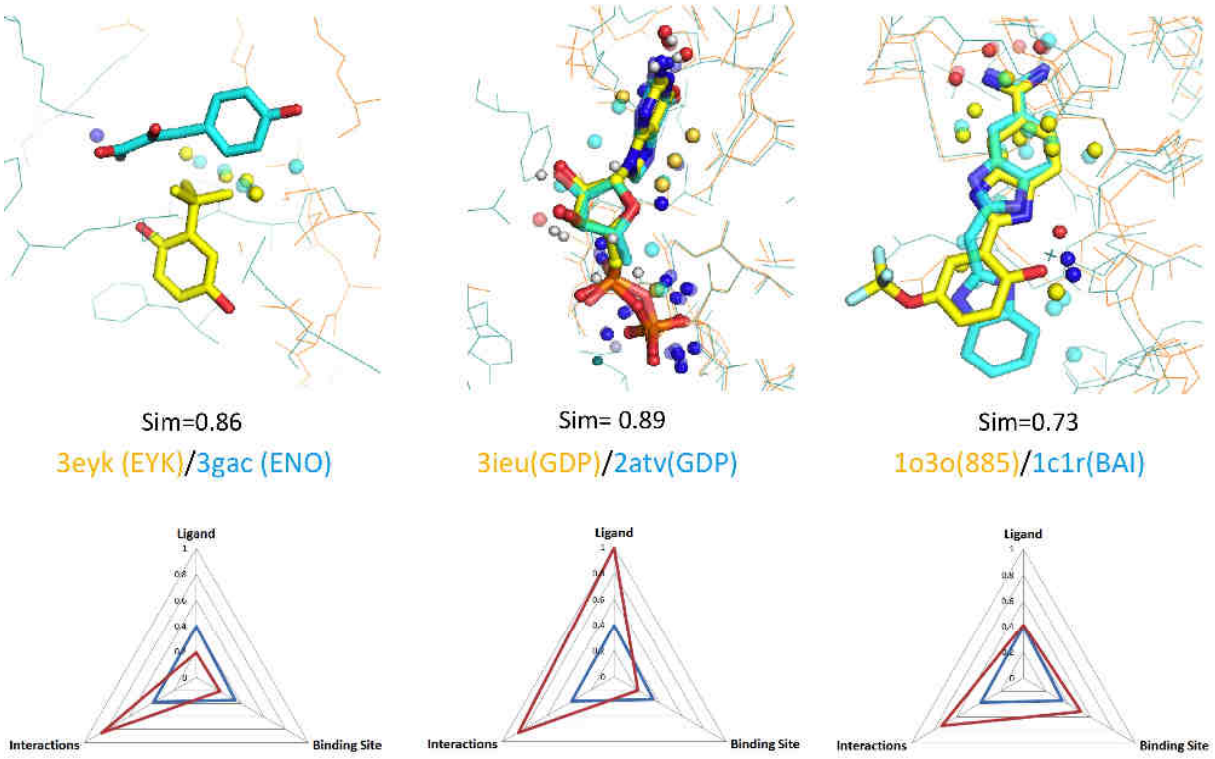


Figure 6 - Ishape alignment of protein-ligand complexes**.** Interaction points are labeled for the two complexes, according to their pharmacophoric properties (hydrophobic, yellow or cyan; H-bond donor, blue; H-bond acceptor, red, negative ionizable, blue; positive ionizable, red; metal chelation, white). Bound ligands heavy atoms are displayed by cpk-colored sticks. The radar plot on lower panel indicates the pairwise similarity of the corresponding ligands (Tanimoto coefficient on ECFP4 fingerprints), binding sites (Shaper similarity) and interaction patterns (IShape similarity). The blue line indicates the similarity threshold for the three metrics, the red line indicates the similarity values for the current protein-ligand complex. **A)** IShape alignment of an influenza hemagglutinin-inhibitor complex (PDB code: 3eyk, HET code: EYK, yellow) and of a macrophage inhibitory factor-substrate complex (PDB code: 3gac, HET code: ENO, cyan). **B)** IShape alignment of protein ERA-GDP complex (PDB code: 3ieu, HET code: GDP, yellow) and of a RAS-like estrogen-regulated growth inhibitor RERG-GDP complex (PDB code: 2atv, HET code: GDP, cyan). **C)** IShape alignment of two trypsin-inhibitor complexes (PDB code: 1o3o, HET code: 885, yellow; PDB code: 1c1r, HET code: BAI, cyan).

4.4 SOME APPLICATIONS OF INTERACTION PATTERN FINGERPRINTS AND GRAPHS

4.4.1 INTERACTION-BASED ALIGNMENT OF PROTEIN-LIGAND COMPLEXES

When aligning protein-ligand complexes, modelers usually have the choice between two scenarios: (i) align protein-bound ligands hoping that protein atoms will overlap adequately, (ii) align protein atoms (main chain or all heavy atoms) hoping that ligand atoms will match. IShape and Grim enables to merge both options by focusing on interaction patterns, thereby optimizing both target and ligand alignment in a single step. A prototypical example of the advantage in aligning interaction patterns is provided **Figure 7** in the alignment of two complexes between the inhibitor brinzolamide and two carbonic anhydrase isoforms (carbonic anhydrase II, PDB code 1a42; carbonic anhydrase IV, PDB code 3znc). A ligand-based alignment is not satisfactory because of the flexibility of the two alkyl side chains that induces a significant shift of the ligand and the proteins (**Figure 7A**). A sequence-based structural alignment of both proteins (using the SYBYL *'Align Structure by Homology'* method) better matches the two proteins structures with however some mismatches in loops enclosing the binding site and in the bicyclic scaffold of the inhibitor (**Figure 7B**). The best compromise is obtained by the interaction pattern alignment generated by Grim (**Figure 7C**) which optimally matches all partners (inhibitor, protein, zinc) at the same time since it simultaneously takes the three kinds of IPAs (*Centered*, *InterLig*, *InterProt*) into consideration during the graph alignment procedure.

In most cases, the alignments produced by IShape (shape-based alignment) and Grim (graph-based alignment) are similar. We however recommend the usage of Grim, which is insensitive to the difference in the number of IPAs between the reference and the fit object. Significant variations in size of either the target or the ligand will produce two interaction patterns, one being a subset of the other. A local alignment (Grim) will therefore usually outperforms a global match (IShape) in these conditions, as shown in the following two examples. In the first one, the same target (dUTPase) is complexed to two related ligands (α,β-imido-dUTP in 3ehw, dUDP in 1duc) but in different oligomeric states (one molecule at the interface of a trimer in 3ehw, monomer in 1duc).

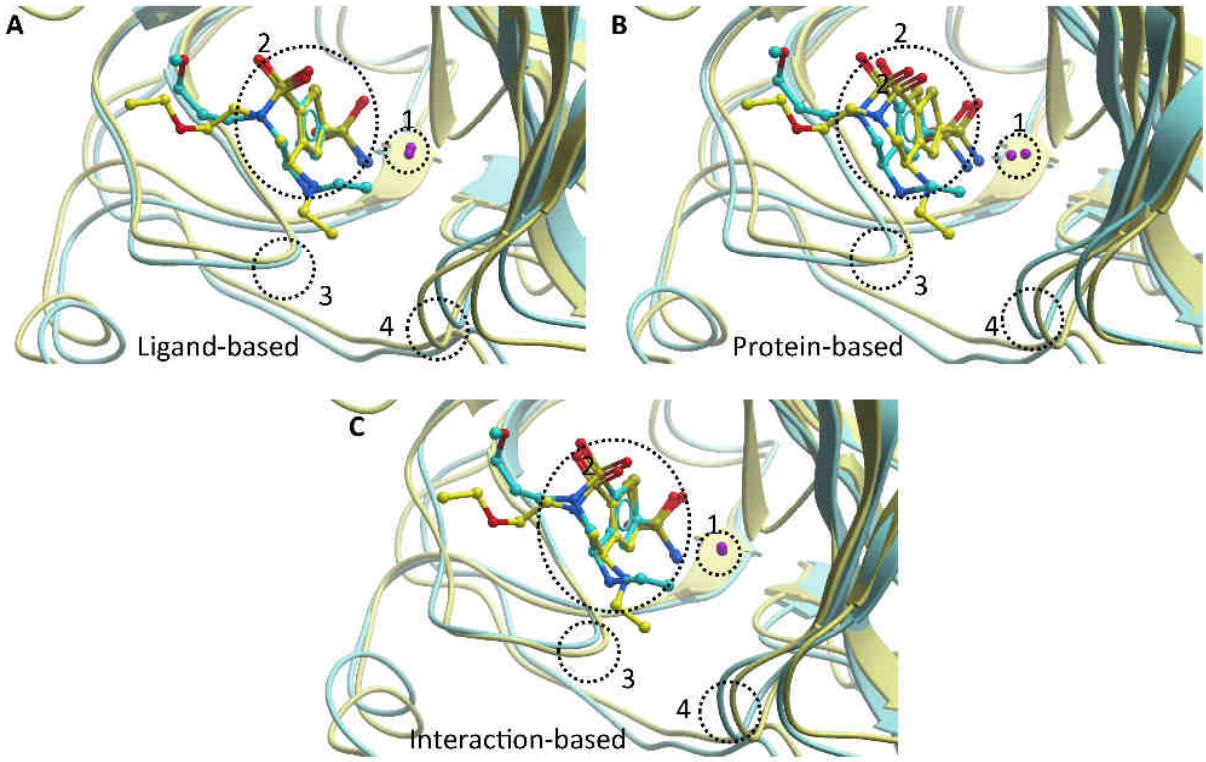


Figure 7 - Ligand-based (panel **A**), protein-based (panel **B**) and interaction-based (panel **C**) alignments of two complexes of brinzolamide bound to human carbonic anhydrase II (yellow ribbons, pdb id: 1a42) and murine carbonic anhydrase IV (cyan ribbons, pdb id: 3znc). The bound inhibitor is displayed by cpk-colored ball and sticks (yellow carbon atom, 1a42-bound inhibitor; cyan carbon atom, 3znc-bound inhibitor). The catalytic zinc ion is displayed by a magenta ball. Four areas of interest are circled: 1, catalytic ion; 2. scaffold of the bound inhibitor, 3. and 4. site-enclosing loops .

As a consequence, the 3ehw ligand exhibits many more interactions (46 interactions, 36 IPAs) than the 1duc ligand (14 interactions, 11 IPAs). Whereas Grim accommodates fairly well this discrepancy by finding the largest common subgraph and perfectly aligning both complexes (**Figure 8A**), IShape fails to align the two protein-ligand complexes by proposing a global shape matching that does not coincide with a proper alignment (**Figure 8B**). In the second example, variations occur at the ligand level with one ligand (PDB code 1pax) being a substructure of the second one (PDB code 1uko), both compounds being co-crystallized with chicken and human poly(ADP-ribose) polymerase. IPA numbers being quite different in both cases (17 in 3ehw, 37 in 1uk0), the same erroneous alignment is produced by IShape whereas Grim perfectly overlaps both protein-ligand structures (**Figure 8C,D**).

Visual inspection of overlaid protein-ligand complexes is a common task for modelers involved in structure-based lead optimization programs. Aligning these molecular objects by focusing on interactions and not structures permits to compare ligands (from a single series) in complex with a single protein, but also multiple ligands

complexed to homologous targets. One of the two alignment tools proposed here (Grim) is particularly interesting since it is insensitive to large variations in one of the two partners and should therefore be of interest for target family-based ligand optimization as well as for structure-based fragment growing for example.

4.4.2 POST-PROCESSING DOCKING POSES

The very first motivation in designing the TIFP fingerprint was to remove the dependency of standard IFPs to the active site definition. To check the comparative performance of the conventional IFP and the newly defined TIFP in rescoring docking poses, a dataset of 42 protein-ligand complexes[16] (set 2) in which ligands have been chosen to cover fragment-like space was retained. We previously reported for this dataset that IFP rescoring (selecting the pose that has the highest IFP similarity to the X-ray solution) was superior to conventional docking scores in self-docking experiments.[16] The 42 fragments were docked again to their cognate protein X-ray structure with Surflex-Dock[46] and 20 poses were generated and stored for every ligand. Poses were scored according to six scoring functions: (i) the native docking score (Surflex score), (ii) the Chemscore empirical scoring function,[47] (iii) IFP similarity to the native X-ray pose (IFP-xray score), (iv) interaction pattern graph similarity to the X-ray pose (Grim-xray), (v) TIFP similarity to the X-ray pose (TIFP-xray score); (vi) interaction pattern graph similarity to that of any known sc-PDB ligand of the same target (Grim-scPDB). The top scored pose for every scoring scheme was retained and its root-mean square deviation (rmsd) to the X-ray pose computed. Considering the docking successful if the top-ranked pose deviates less than 2.0 Å from the X-ray solution, we confirmed that the quality of poses generated by Surflex-Dock is quite remarkable in ca. 90% of the cases. We also confirm that rescoring according to IFP similarity statistically enhance the quality of the top ranked posed (65% of success vs. 58% for the Surflex score and 44% for Chemscore rescoring; **Figure 9A**). Noteworthy, transforming interaction fingerprint (IFP) into interaction fingerprint patterns (TIFP) slightly alter the quality of the rescoring with only 60% of success, thereby not providing any significant advantage with respect to energy scoring at least for this dataset and the default Surflex scoring function. We suspect, as previously reported before, that some noise has been introduced in TIFP representation because of the large proportion of hydrophobic

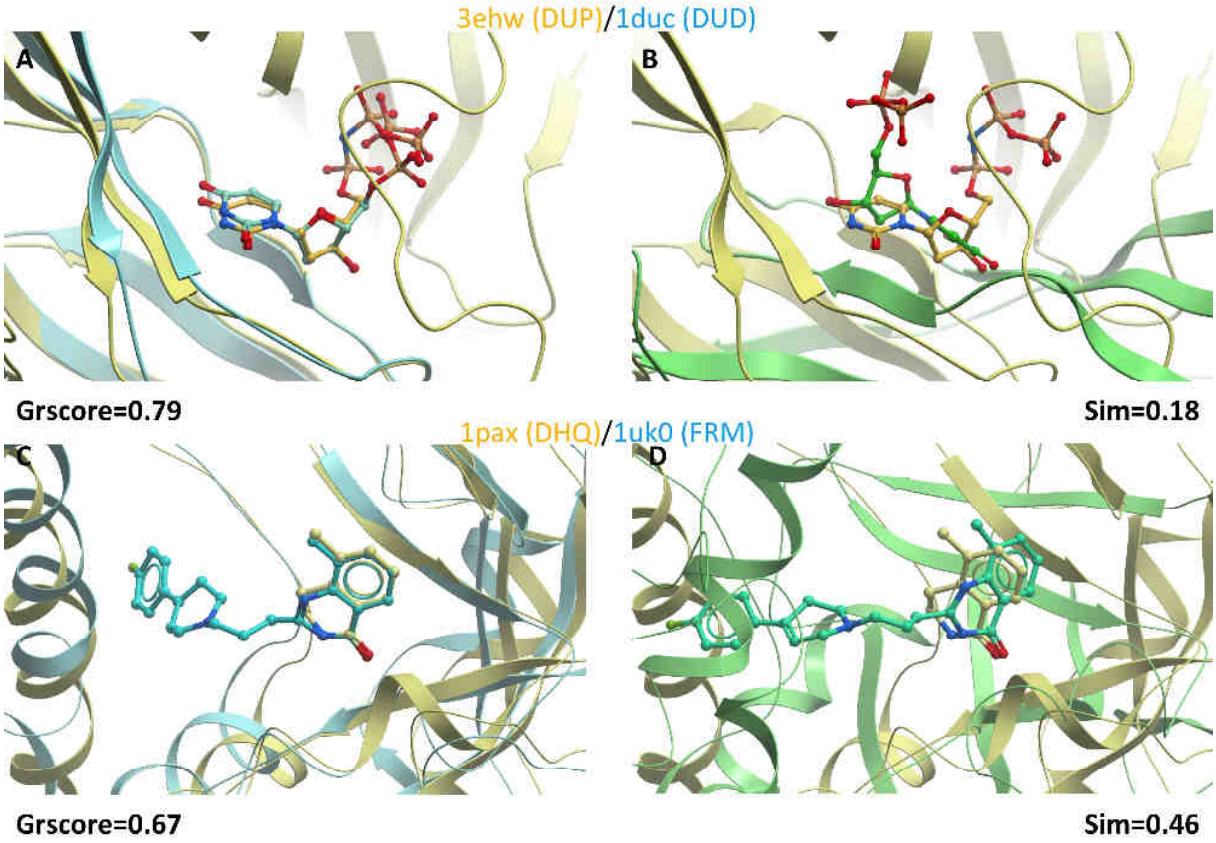


Figure 8 - IShape versus Grim alignments for difficult cases. **A)** Grim alignment of two related ligands (DUP: α-β-imido-dUTP, DUD: dUDP) co-crystallized with dUTPase as a trimer (PDB code: 3ehw, yellow ribbons) or monomer (PDB code: 1duc, cyan ribbons). Bound ligands are represented by cpk-colored ball and sticks (carbon in 3ehw, yellow; carbon in 1duc, cyan; oxygen, red; nitrogen, blue; phosphorus, orange. The Grim Grscore is indicated below the alignment. **B)** Ishape alignment of two related ligands (DUP: α-β-imido-dUTP, DUD: dUDP) co-crystallized with dUTPase as a trimer (PDB code: 3ehw, yellow ribbons) or monomer (PDB code: 1duc, green ribbons). Bound ligands are represented by cpk-colored ball and sticks (carbon in 3ehw, yellow; carbon in 1duc, green; oxygen, red; nitrogen, blue; phosphorus, orange. The Ishape similarity score is indicated below the alignment. **C)** Grim alignment of two inhibitors (DHQ, FRM) co-crystallized with poly(ADP-ribose) polymerase (PDB code: 1pax, yellow ribbons; PDB code: 1uk0, cyan ribbons). Bound ligands are represented by cpk-colored ball and sticks (carbon in 1pax, yellow; carbon in 1uk0, cyan; oxygen, red; nitrogen, blue; fluorine, light green). The Grim Grscore is indicated below the alignment. **D)** Ishape alignment of two inhibitors (DHQ, FRM) co-crystallized with poly(ADP-ribose) polymerase (PDB code: 1pax, yellow ribbons; PDB code: 1uk0, cyan ribbons). Bound ligands are represented by cpk-colored ball and sticks (carbon in 1pax, yellow; carbon in 1uk0, cyan; oxygen, red; nitrogen, blue; fluorine, light green). The Ishape similarity score is indicated below the alignment.

interaction-containing triplets which dominates the similarity calculation. To avoid this flaw, poses were rescored by interaction pattern graph similarity with Grim. Grim rescoring significantly improves the performance of the rescoring (70% of success, **Figure 9A**) notably for high precision poses (< 1 Å rmsd). Graph-based rescoring is very efficient provided that a good pose is available within the pool of possible solutions and does not suffer from the previously reported TIFP drawback since graph nodes are weighted according to the scarcity of the encoded interaction. If we do not consider the best rmsd score which indeed is the best possible score according to the quality of generated poses, Grim rescoring with respect to the X-ray pose (Grim_xray) is the only method that is statistically superior to the native Surflex-Dock scoring (Kolmogorov-Smirnov test, D=0.3156; p=0.016)

To remove all possible bias, we discarded the true X-ray pose as a reference for interaction pattern similarity and used instead all existing sc-PDB ligands co-crystallized with the target under investigation (Grim-scPDB rescoring). This scoring procedure is only possible with either interaction pattern fingerprints or graphs and not doable with conventional IFPs since defining a unique binding site of constant composition is almost impossible for most targets. Reference sc-PDB entries were selected according to the recommended Uniprot name of the target of interest and provided for each ligand a variable number of references (from none to 115). If the target was absent in the sc-PDB (10 out of 42 cases), the closest sc-PDB target was manually selected as a surrogate (see full list of references in **Supplementary Table 2**). In four cases (PDB entries 1pu7, 1pu8, 1qpr, 1qy2) no surrogate could be found and no rescoring solution could be proposed. As to be expected, rescoring by similarity of interaction pattern to target-specific sc-PDB ligand sets (Grim_scPDB score) is inferior to rescoring with respect to the true X-ray solution (**Figure 9A**, Kolmogorov-Smirnov test: D=0.2821, P=0.0073). It however presents the considerable advantage of a target-specific rescoring taking into account all available structural information and does not necessitate the selection of a particular reference for comparing target-ligand interactions. For the dataset under investigation, we should however acknowledge that this advantage with respect to conventional energy-based scoring (Surflex, Chemscore) vanishes for pose prediction with rmsd higher than 1.5 Å.

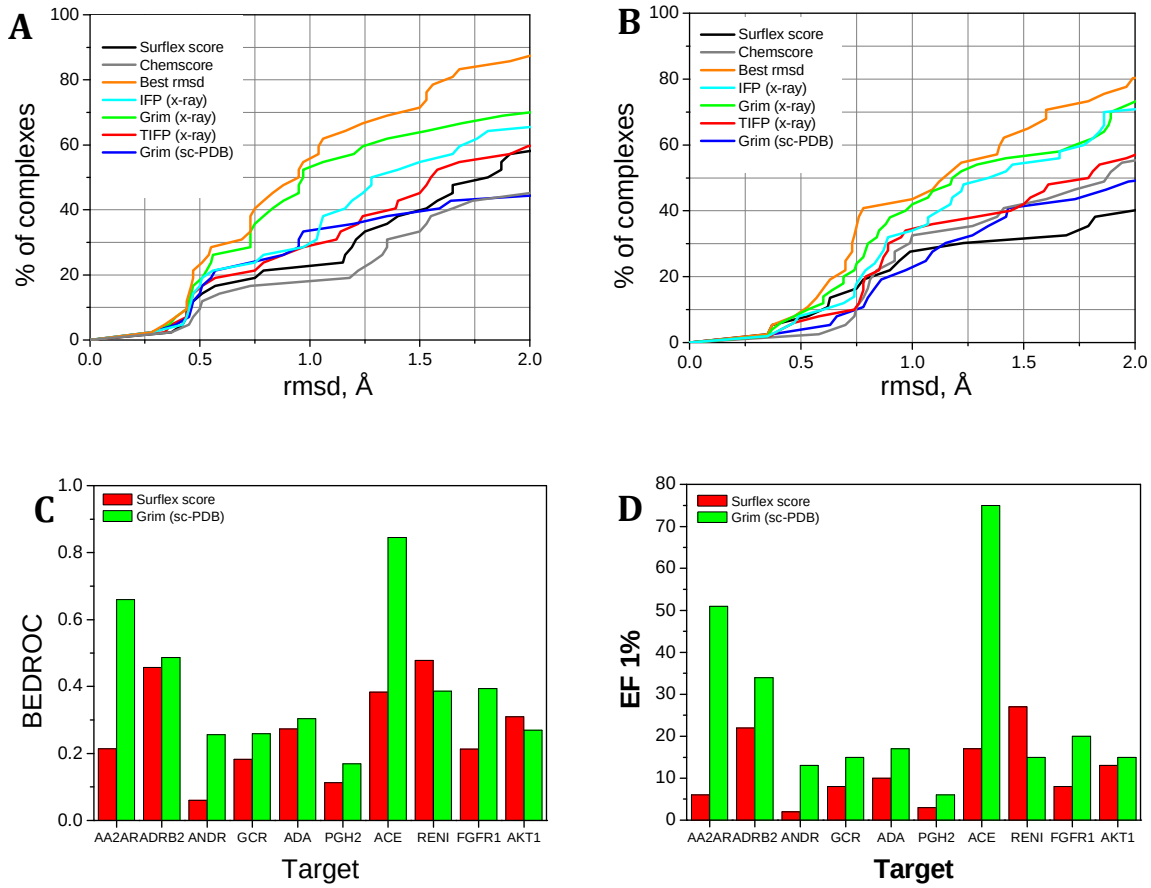


Figure 9 - Post-processsing docking poses by similarity of interaction fingerprints (IFP) and interaction patterns (TIFP, Grim). **A)** Posing accuracy or 42 low molecular weight ligands[16] (set 2) obtained upon Surflex-Dock docking. For each complex, top ranked poses are stored according to the native Surflex-Dock score (black line), the Tanimoto similarity of standard interaction fingerprints to that of the native X-ray pose (IFP-xray, cyan line), the Tanimoto similarity of interaction fingerprint triplets to that of the X-ray pose (TIFP-xray, red line), the similarity of interaction pattern graphs to that of the X-ray pose (Grim-xray, green line), or the similarity of interaction pattern graphs to that of X-ray poses of all sc-PDB ligands sharing the same target (Grim-scPDB, blue line) . **B)** Posing accuracy of 36 ligands from the CCDC/Astex subset. For each complex, top ranked poses are stored according to the native Surflex-Dock score (black line) or the similarity of interaction pattern graphs to that of X-ray poses of all sc-PDB ligands sharing the same target (blue line). **C)** Surflex-Dock vs. Grim scoring of 10 poses for a set of DUD-E[34] actives and decoys and 10 representative targets: Adenosine A2A receptor (AA2AR), Beta2 adrenergic receptor (ADRB2), Androgen receptor (ANDR), Glucocortocoid receptor (GCR), Adenosine deaminase (ADA), Prostaglandin G/H synthase 2 (PGH2), Angiotensin-converting enzyme (ACE), Renin (RENI), Fibroblast growth factor receptor 1 (FGFR1),RAC-alpha protein kinase (AKT1). The top-ranked pose according to the Surflex-Dock score was retained. For Grim rescoring, scores were fused by ligand and protein-ligand sc-PDB reference complexes of the same target and the best Grscore retained. The discrimination of actives from inactives is measured by the area under the BEDROC[49] curve. **D)** Enrichment in true actives at a constant 1% false positive rate upon scoring docking poses by either the native Surflex-Dock score or the Grim Grscore. Targets and ligand sets are identical to that indicated in panel C.

On the second dataset of 36 CCDC/Astex complexes, rather similar trends were observed (**Figure 9B**). Surflex provided adequate poses for 80% of the ligands. Rescoring using the true X-ray pose evidently provides a significant advantage with a better performance of graph rescoring (Grim) and conventional IFP scoring (70% of good poses) with respect to TIFP scoring (57% of success only) or energy-based scoring (40% and 55% of success for Surflex and Chemscore, respectively). Grim-scPDB rescoring also produces better poses than the conventional Surflex score but not Chemscore rescoring for this dataset (**Figure 9B**). We clearly acknowledge that a much larger docking set may be necessary to really appreciate the benefit of Grim rescoring on the quality of docking poses for known actives.

We next ask the question whether a real advantage is also found in virtual screening experiments for which Grim-scPDB rescoring may be particularly adapted to distinguish true actives from decoys. For that purpose, ten targets of pharmaceutical interest (**Table 3**) covering 5 major target families (proteases, G protein-coupled receptors, other enzymes, protein kinases, nuclear hormone receptors), along with a set of prepared actives and decoy ligands, were chosen from the DUD-E dataset.[34] To avoid any possible bias in Grim-scPDB rescoring, caution was given to select for each target an X-ray structure absent from the sc-PDB dataset, or to remove it from the pool of references. Starting from the same set of docking poses generated for each target by Surflex-Dock, the respective ability of the native Surflex-Dock score and of the Grscore to discriminate actives from chemically similar decoys was further inspected. In 5 out of 10 cases (AA2AR, ANDR, GCR, ACE, FGFR1), the area under the ROC curve was significantly improved (more than 0.1 unit) upon Grim rescoring (**Table 3**). In three cases (ADRB2, PGH2, RENI) both scoring methods could be considered as equally potent in segregating actives from decoys. A slight advantage of the Surflex-Dock native scoring function could only be found in the remaining two cases (ADA, AKT1). In virtual screening, it is however of utmost importance to enrich the list of top scorers (to be experimentally confirmed) in true actives. Following accepted recommendations,[48] we therefore focused the analysis in early enrichment in true actives by computing two important statistical parameters: the Boltzmann enhanced discrimination of the ROC (BEDROC) curve[49] as well as the enrichment in true actives at the low false positive rate of 1% (EF1). Both metrics demonstrates an enhanced advantage in rescoring docking poses with Grim in 8 out of 10 cases when considering the BEDROC metric, and in 9 out of 10

cases when considering the EF1 value (**Table 3**,**Figure 9B**). From 1.5 to 4 times more true actives at a constant 1% false positive rate are found with the Grim rescoring method, therefore demonstrating its usefulness in virtual screening scenarios. For 7 out of the 10 targets, the benefit in Grim rescoring was related to the number of protein-ligand X-ray reference structures. However, it is currently impossible to draw general conclusions since such retrospective virtual screening studies are known to be very dependent on the chosen ligand set (actives and decoys) and protein coordinates (e.g. active vs. inactive state of a receptor). The herein proposed Grim rescoring mode is however very interesting since it enables a user-independent scoring strategy capitalizing on existing knowledge about known ligand binding modes to the target of interest. Since sc-PDB protein-ligand interaction patterns only have to be computed once and are further stored in a look-up table, post-processing is relatively straightforward and only necessitates (in addition to the set of docked poses) the name of the target of interest.

Table 3 - Area under the ROC plot of a binary classification (active, inactive) of docked poses to the X-ray structure of 10 representative targets

	PDB code	DUD-E Actives	DUD-E Decoys	ROC[a]		BEDROC[b]		EF1[c]		sc-PDB references
				SF-Dock	Grim	SF-Dock	Grim	SF-Dock	Grim	
G Protein-Coupled receptors										
Adenosine A2A receptor (AA2AR)	3pwh	482	31500	0.736	0.911	0.214	0.660	6	51	4
Beta2 adrenergic receptor (ADRB2)	3ny8	231	15000	0.854	0.846	0.457	0.487	22	34	3
Nuclear hormone receptors										
Androgen receptor (ANDR)	2am9	269	14350	0.470	0.730	0.060	0.256	2	13	29
Glucocortocoid receptor (GCR)	1p93	258	15000	0.557	0.742	0.183	0.259	8	15	8
Other enzymes										
Adenosine deaminase (ADA)	1a4l	93	5450	0.828	0.749	0.274	0.304	10	17	20
Prostaglandin G/H synthase 2 (PGH2)	3nt1	435	23150	0.620	0.626	0.113	0.169	3	6	7
Proteases										
Angiotensin-converting enzyme (ACE)	3zqz	282	16900	0.840	0.952	0.383	0.845	17	75	14
Renin (RENI)	3sfc	104	6958	0.878	0.850	0.478	0.386	27	15	29
Protein kinases										
Fibroblast growth factor receptor 1 (FGFR1)	3tt0	139	8700	0.721	0.836	0.213	0.394	8	20	7
RAC-alpha protein kinase (AKT1)	4ekl	293	16450	0.759	0.709	0.310	0.270	13	15	10

[a] area under the ROC curve for a binary classification of ligands (actives, decoys) from their docked poses scored by either the Surflex-Dock score (SF-Dock) or the Grscore of the interaction pattern graphs (Grim)

[b] area under the Boltzmann-enhanced discrimination of receiver operating characteristic (BEDROC) curve.

[c] Percent of actives found when 1% of decoys have been retrieved.

4.4.3 SCAFFOLD HOPPING WITH INTERACTION PATTERN CONSERVATION

Since the TIFP fingerprint generically describe spatial interactions between a ligand and its target, it can be applied to search for truly bioisosteric scaffolds to any reference of known binding mode. All 9877 sc-PDB ligand were thus fragmented according to retrosynthetic RECAP rules, therefore generating a set of 20839 fragments whose TIFP fingerprint was deduced, after pairwise atom matching, from that of the full fingerprint from the parent ligand. To establish the proof-of-concept, we selected as reference the 1-N-(pyrimidin-2-yl)benzene-1,3-diamine scaffold (**Figure 10**) from a tyrosine-protein kinase inhibitor (HET code: MPZ, PDBid: 1y57). This scaffold presents the prototypical interaction observed between most ATP-competitive inhibitors and the hinge region of protein kinases. A bioisostere to this scaffold was here defined as any sc-PDB fragment fulfilling a TIFP similarity higher than 0.50 and a chemical similarity (expressed by a Tanimoto coefficient on ECFP4 fingerprint) below 0.25. Out of the 27 selected fragments, 23 originate from protein kinase inhibitors (**Supplementary Table 8**) and also interact with the above-described kinase hinge region. Aligning the corresponding fragments from their interaction pattern (IPAs in *Centered* mode) confirm that all hits are really bioisosteric to the reference, the H-bond acceptor and donor atoms (to the hinge region) being well matched (**Figure 10**). Interestingly, some hits could be retrieved from ligands interacting with unrelated proteins (methyltransferase, ATPase, pyruvate decarboxylase, sugar epimerase) albeit with scaffolds (adenine, aminopyrimidine) frequently observed in protein kinase inhibitors. Of course, the likehood of finding bioisosteric scaffolds is higher among ligands sharing the same target or target class (e.g. estrogen receptor alpha-binding scaffolds, **Figure 11A**). However, fragment interaction patterns may be conserved among completely unrelated proteins exhibiting only subpocket similarities (e.g. transcriptional regulator-bound 3-aminooxolan-2-one and chymotrypsinogen-bound tyrosine, **Figure 11B**).

Traditional sources for finding bioisosteric groups rely on existing structure-activity knowledge.[50,51] Computational approaches to find potential replacements may be derived from pairwise 2D and 3D similarity searches.[52,53]

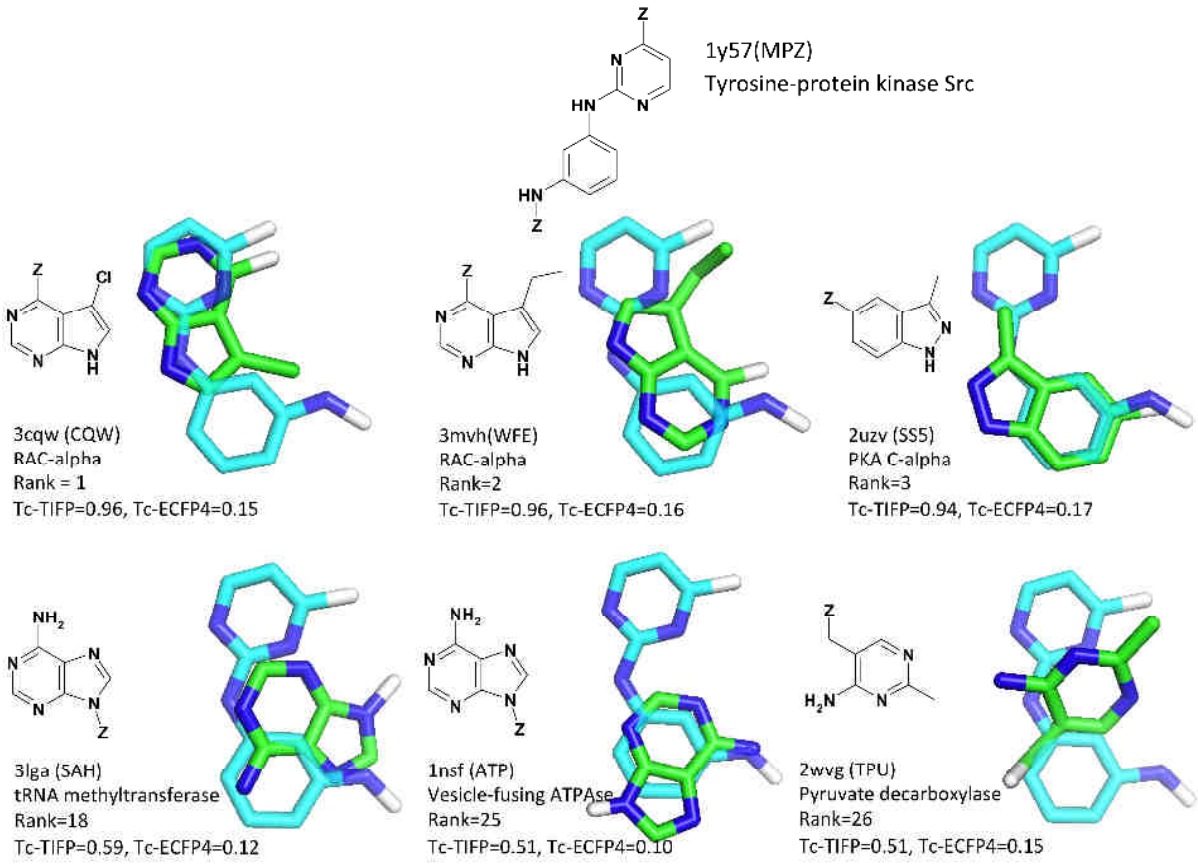


Figure 10 - Search for bioisosteric scaffolds to the 1-N-(pyrimidin-2-yl)benzene-1,3-diamine fragment bound to tyrosine-protein kinase Src (PDB code 1y57, HET code MPZ). In upper panel are displayed the top 3 scored scaffolds bound to a protein kinase (according to TIFP fingerprint similarity) and aligned with Grim to the reference (reference carbon atoms in cyan, aligned carbon atoms of the selected scaffold in green, oxygen and nitrogen atoms in blue and red, respectively). PDB and HET codes of the aligned ligands, common name of the aligned scaffold-bound protein, rank of the selected scaffold (scored by decreasing TIFP similarity), and pairwise protein-scaffold similarities (TIFP similarity, ECFP4 similarity) are indicated for every hit. In the lower panel are displayed the top 3 scaffolds bound to a non-protein kinase target. The Z atom indicates the branching points generated by RECAP fragmentation

A few methods focusing on existing protein-ligand 3D structures have been reported but are restricted to different ligands complexed with the same target,[54] or require the prior knowledge of similar binding sites.[55,56] To the best of our knowledge, we report here the first method considering biosiosteric searches from a set of existing protein-ligand interactions in the PDB with no a priori on either ligand and/or binding site similarity.

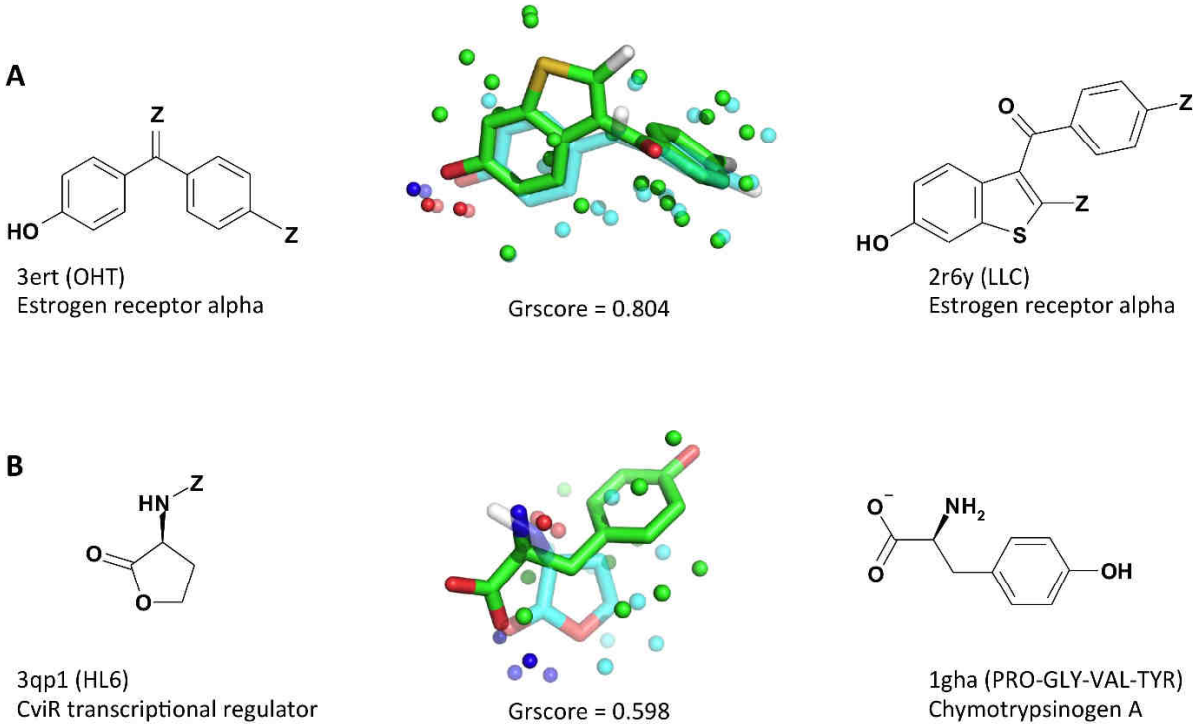


Figure 11 - Grim alignment of the top-ranked bioisosteric scaffold to two references, one fragment from the estrogen receptor alpha ligand OHT (panel A, pdb id: 3ert), one from the CviR transcriptional protein ligand HL6 (panel B, pdb id: 3qp1). The structure of the most bioisosteric scaffold (HET code, pdb id) and its target protein are indicated on the right hand side. Query (cyan carbon atoms) and top-ranked scaffolds (green carbon atoms) are aligned according to Grim along with the fitted IPAs (green, hydrophobic interactions of the query; cyan, hydrophobic interactions of the hit; red: hydrogen bond (protein acceptor); blue, hydrogen bond (protein donor)). IPAs of the query are displayed by transparent spheres, IPAs of the hits are displayed by solid spheres. The similarity scores of the interaction pattern graphs (Grscore) are indicated below the alignments.

5. CONCLUSION

We herewith propose a generic fingerprint (TIFP) of protein-ligand interaction patterns as well as two computational methods (IShape, Grim) to efficiently compare and align protein-ligand complexes. The TIFP fingerprint currently describes standard intermolecular interactions but could be easily extended to less frequent, weaker but sometimes important interactions like weak hydrogen bonds (C-H…O), halogen bonds, or cation(donor)-pi interactions. It enables an ultrafast comparison of protein-ligand complexes but suffers from the predominance of hydrophobic contacts among most PDB-ligand X-ray structures. We therefore prefer to directly manipulate the interaction pattern as a list of pseudoatoms describing the interactions engaged between the ligand and the target. Interaction patterns were computed for 10 000 protein-ligand complexes of the sc-PDB dataset, therefore providing a framework for two important applications: (i) post-processing docking poses while taking into account all known interactions with the target of interest, (ii) search for truly bioisosteric scaffolds to a reference by prioritizing substructures exhibiting conserved interaction patterns. Last, the proposed alignment tools (notably the Grim method) enables to directly fit protein-ligand complexes from the corresponding interaction patterns, without having to choose between a ligand-based or a target-based point of view. With the rapid growth of structural information in the Protein Data Bank, such methods are believed to play an important role in assisting molecular modelers to visualize common features or differences among protein-ligand complexes of biological interest.

6. ACKNOWLEDGMENT

The Calculation Center of the IN2P3 (CNRS, Villeurbanne, France) and the GENCI (Project x2010075024) are acknowledged for allocation of computing time. The Institut de Recherches Servier (Croissy/Seine, France) is warmly acknowledged for a doctoral grant to J.D. and for useful discussions.

7. COMMENTAIRES

7.1 DETECTION DES INTERACTIONS

Comme nous avons pu l'observer au cours du premier chapitre, il existe de nombreuses interactions non-covalentes possibles entre un ligand et une protéine. En l'état, seules les interactions clés telles que les liaisons Hydrogène, les interactions ioniques ou les contacts apolaires et aromatiques ont été considérées. Nous sommes ainsi partis de l'hypothèse que ces interactions sont suffisantes pour reconnaître un mode d'interaction au sein d'une famille de protéine ou de ligand. Cependant, il existe certains cas où cette simplification ne suffit plus.

Les liaisons Hydrogène faibles avec un groupement apolaire peuvent jouer un rôle crucial dans la reconnaissance moléculaire, même si son rôle dans l'affinité reste faible. Dans le cas des protéines kinases, la partie « Hinge » réalise généralement au moins une liaison Hydrogène forte avec le ligand (**Figure 12A**), mais il peut effectuer quelque fois une liaison Hydrogène faible (**Figure 12B)**. Ce groupement CH, initialement hydrophobe, se retrouve ainsi donneur de liaison H lorsque le Carbone est lié à un Azote ou un Oxygène dans un hétérocycle aromatique.

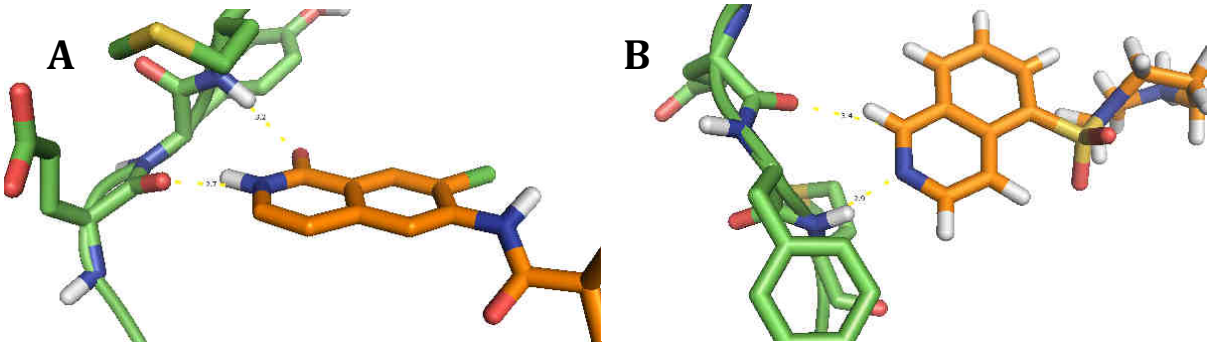


Figure 12 - A : Cas typique d'une molecule réalisant 2 liaisons Hydrogène avec le hinge (Code PDB:3NCZ). **B**: Cas d'une molecule réalisant 1 liaison Hydrogène forte et une faible (Code PDB:2F2U).

Une étude a ainsi été réalisée pour montrer l'influence des règles de détection des interactions sur la sélection de molécules. On utilise comme référence la structure d'une protéine kinase co-cristallisée avec un ligand actif. Ce dernier effectue 2 liaisons hydrogène fortes avec la charnière. Un criblage par docking de 660 molécules a été effectué avec le logiciel Gold et 10 poses ont été retenues par molécule. Toutes ont été testées expérimentalement avec des mesures d'IC50 et des pourcentages d'inhibition. Chacun des modes d'interactions de ces poses a ensuite été comparé à celui du ligand

d'origine avec Grim et deux sélections ont été effectuées : la première correspond à la meilleure pose pour chaque ligand selon GoldScore et la seconde selon Grim. Enfin une annotation fonctionnelle binaire de chaque pose a été calculée à partir des règles suivantes : Pourcentage d'enfouissement > 70%, Nombre de liaisons Hydrogène avec le hinge ≥ 2, Nombre total de liaisons Hydrogène ≥ 3. On affecte alors à chaque molécule la valeur 1 si elle respecte les conditions mentionnées ou 0 sinon.

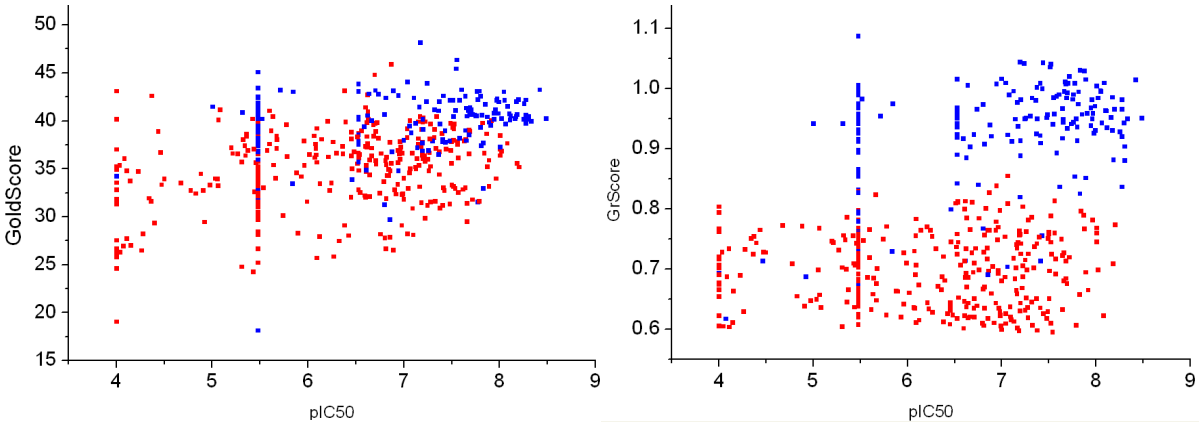


Figure 13 - A gauche: distribution des meilleurs scores de Gold par molécule en fonction du pIC50 de celle-ci. A droite: distribution des meilleurs scores de Grim par molécule en fonction du pIC50. Les points sont coloriés en bleu lorsqu'il respecte les règles géométriques et en rouge dans le cas contraire

Les distributions entre l'activité et le score de graphe ou le score de docking ne montrent aucune corrélation, distribution attendue puisque les interactions protéine/ligand ne sont pas les seules composantes de l'activité (**Figure 13**). Cependant, une nette tendance se dégage à travers l'annotation fonctionnelle, puisque l'essentiel des molécules répondant aux divers critères se situent au dessus d'une valeur de 0.85 en GrScore, alors que le score de docking ne permet pas cette discrimination. Malgré tout, une partie des molécules actives (pIC50 > 6) est mal scorée par Grim et mal annotée. Après observation de ces molécules, il s'avère qu'elles n'effectuent en effet qu'une seule liaison Hydrogène forte mais aussi une liaison Hydrogène faible, non détectée par Grim.

Dans le cadre d'un criblage virtuel, beaucoup de molécules actives seraient ainsi éliminées à cause de certaines interactions importantes qui sont ignorées. La sélection est par conséquent toujours dépendante des modes d'interactions du/des ligand(s) de référence. Il est donc nécessaire, suivant les cas d'études, de prendre en compte des interactions plus spécifiques, et de réfléchir sur les règles d'appariements. Cependant, il est important de souligner qu'il ne faut pas faire de généralité car ces cas restent atypiques.

7.2 GRIM VS ISHAPE

7.2.1 COMPARAISONS

IShape et Grim diffèrent tout d'abord par l'information qu'ils utilisent au départ. IShape nécessite des points d'interactions centrés, tandis que Grim a la possibilité de prendre tous les positionnements (centered, InterLig et InterProt). Cela n'implique pas qu'IShape ne peut pas gérer ces positionnements, mais seulement qu'il n'a pas été conçu de cette façon. La raison en est plus historique que technique puisqu'IShape a été le tout premier logiciel créé à partir des points d'interactions. Il serait intéressant d'observer dans les cas difficiles trouvés avec Grim ou même ceux d'IShape, si en utilisant tous les positionnements, IShape serait capable d'améliorer l'alignement de complexes protéine/ligand. Cependant, IShape effectue un alignement global des complexes, ce qui implique une moyenne de recouvrement et non un alignement partiel mais optimal des points d'interactions.

7.2.2 SIMILARITE

La notion de valeur de similarité est aussi très différente entre les 2 logiciels. IShape fournit des valeurs bornées entre 0 et 1 alors que Grim n'est pas limité et peut aller de 0.3 à 2.

Cette différence de borne est essentiellement due à la représentation qu'il en est fait. Qu'importe la taille des objets initiaux, IShape tentera de maximiser un recouvrement. Ce dernier est soit nul, donnant une valeur de similarité de 0, partielle, entre 0 et 1, soit optimal et donc de 1. La dépendance à la taille peut se ressentir via la métrique utilisée, ici le Ref Tversky, qui donnera plus d'importance à la référence qu'à la comparaison. Si la référence est plus grande que la comparaison, même avec un recouvrement optimal, la partie non recouvrée baissera le score. Inversement, si la référence est plus petite, le score sera de 1 pour un recouvrement optimal. Cela se vérifie en regardant la distribution moyenne de score en fonction du nombre de points d'interaction centrés, réalisé sur l'ensemble des entrées de la sc-PDB-2011 (**Figure 14A**). La distribution diminue de façon linéaire par rapport au nombre de points (R^2=0.94) montrant ainsi l'extrême dépendance à la taille de la référence.

Le cas de Grim est différent, puisqu'il deviendra plus exigeant à mesure que le nombre de points d'interactions augmente. Lorsque l'on regarde la distribution

moyenne des scores de Grim en fonction du nombre de points d'interactions « Merged » de référence (**Figure 14B**), on remarque que le score moyen augmente pour de petits jeux d'interactions mais commence à diminuer lorsque le nombre de points de la référence dépasse 75. On notera aussi qu'il existe une corrélation forte entre le nombre de points centré et « Merged » (Q^2=0.98), permettant ainsi de pouvoir comparer les divers graphiques. Ainsi, plus le nombre de points augmente, plus le nombre d'appariements possibles augmente et plus il est nécessaire pour Grim d'être exigeant en terme de score. De plus, lorsque l'on compare les valeurs maximales moyennes possibles entre IShape et Grim (**Figure 14C,D**) par rapport au nombre de points de référence, IShape tend à diminuer la valeur similarité alors que Grim l'augmente de façon linéaire.

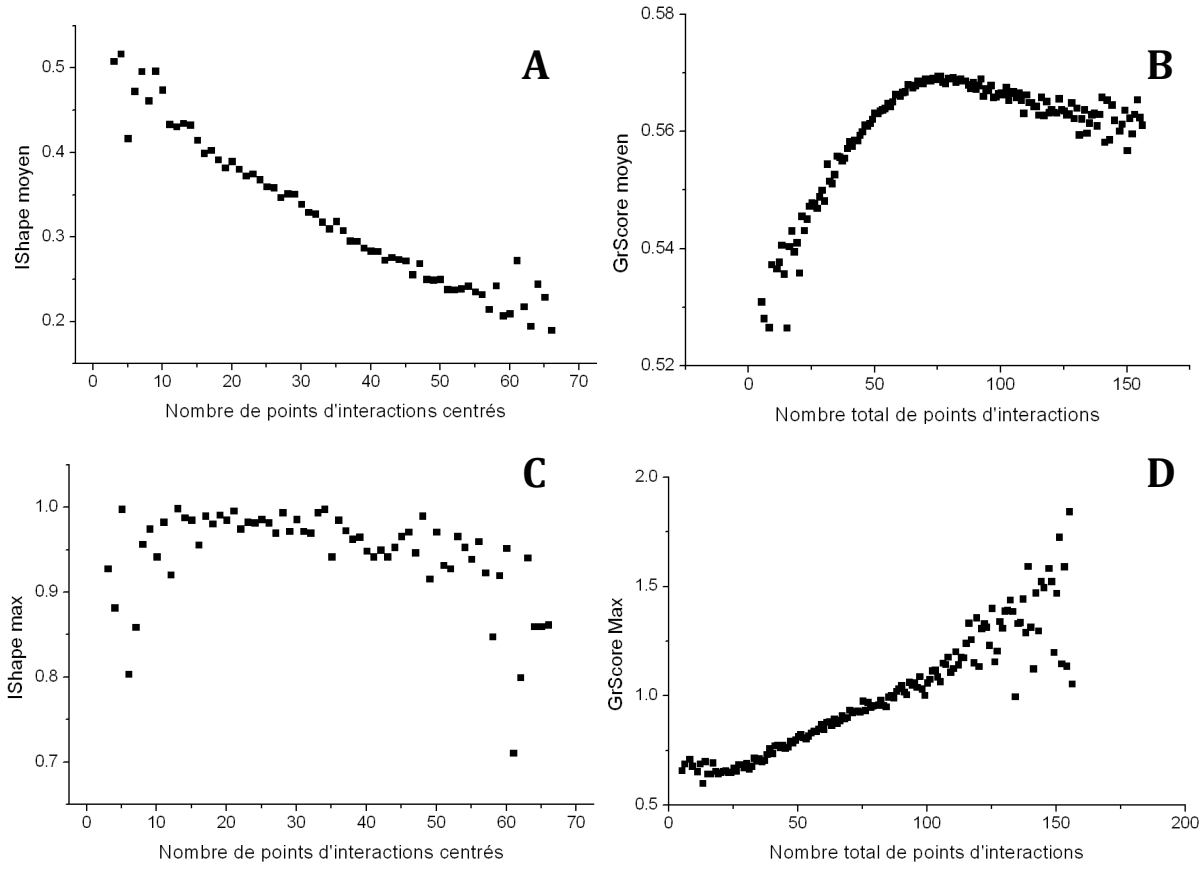


Figure 14 - Distribution de score moyen en function du nombre de points d'interactions sur l'ensemble des entrées de la sc-PDB. A) Score moyen d'IShape en function du nombre de points d'interactions centrés. B) Score moyen de GrScore en function du nombre de points d'interactions Merged. C) Score maximal moyen d'IShape en function du nombre de points d'interactions centrés. D) Score maximal moyen de GrScore en function du nombre de points d'interactions Merged.

Grim doit être choisi par rapport à IShape dans le cadre de comparaison de mode d'interactions de complexe protéine/ligands. En effet, IShape est très dépendant de la référence et tend à diminuer les scores pour compenser l'effet de taille, alors que Grim linéarise le score. Il serait donc important de recalculer la matrice complète de similarité de paires de modes d'interactions et d'observer l'incidence sur la corrélation avec la similarité de paires de sites actifs et de ligands.

7.3 ALIGNEMENT DE COMPLEXES PROTEINE/LIGAND

Le cas de la **Figure 7** montre l'apport de la combinaison de toutes les parties (protéine et ligand) pour l'alignement de complexes protéine/ligand. Divers cas n'ont cependant pas été testés et méritent pourtant notre attention. Prenons le cas de l'alignement basé seulement sur la protéine : la procédure employée utilise l'ensemble des carbones alphas pour aligner les protéines. Cependant, seul le site de liaison est réellement utile dans notre cas. Trois nouveaux tests ont donc été réalisés : le premier utilise la même méthode d'alignement que pour aligner les protéines, mais seulement avec les résidus du site actif (**Figure 15A**). Les deux autres sont basées sur VolSite et Shaper en utilisant les cavités à 6 Å (**Figure 15B**) et les cavités All (**Figure 15C**).

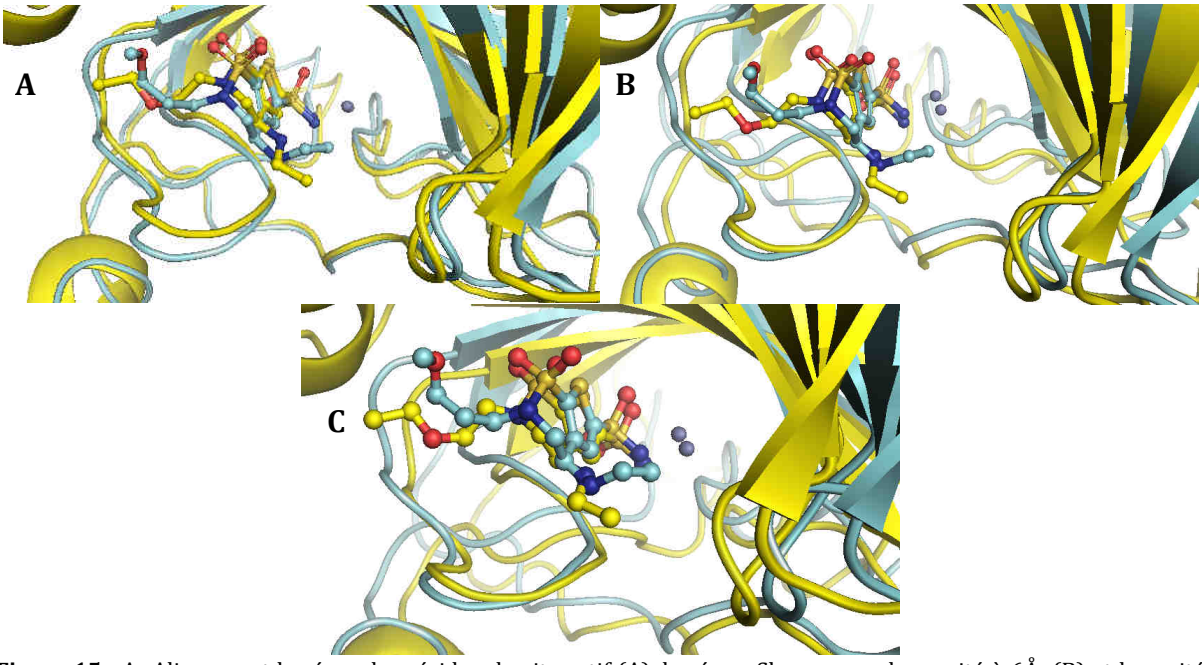


Figure 15 - A- Alignement basé sur les résidus du site actif (A), basé sur Shaper avec les cavité à 6Å (B) et la cavité complète (C), de deux complexes du brinzolamide lié à la carbonique anhydrase II humaine (en jaune) et de la carbonique anhydrase IV (cyan).

La sélection des seuls résidus du site de liaison améliore l'alignement par rapport à la protéine entière. Cela peut s'expliquer par la présence d'un brin bêta proche du site de liaison modifiant l'alignement global, qui est ignoré lorsque l'on ne se focalise que sur le site de liaison. L'alignement obtenu est ainsi très similaire à celui fournit par Grim.

Ce brin bêta modifie de façon importante la cavité du site et par conséquent les alignements issus de VolSite et Shaper. Dans le cas présent, utiliser la cavité du site de liaison ne permet pas de correctement aligner les deux complexes protéine/ligand.

La plupart des méthodes utilisées pour aligner deux complexes utilisent le site de liaison ou plus comme référence. Lorsque la référence ou la comparaison est incomplète ou monomère/dimère, la définition du site de liaison peut grandement changer et par causalité l'alignement. Après observation visuelle, aucune interaction n'est réalisée entre le brinzolamide et ce brin beta, ce qui permet à Grim de l'ignorer et de ne pas influencer l'alignement.

7.3.1 POST-TRAITEMENT D'ARRIMAGE MOLECULAIRE

Lors de post-traitement, Grim compare le mode d'interaction des poses de docking par rapport à celui/ceux du/des ligand(s) de référence. Pour réaliser cela, Grim utilise un recouvrement de graphes pour apparier les points d'interactions, puis effectue un alignement afin d'obtenir le score GrScore. Cependant, la position des points d'interactions avant et après alignement avec Grim peut avoir changé, entraînant des scores ne correspondant pas au mode d'interaction de la pose (**Figure 16**). Cette erreur arrive cependant dans des cas très spécifiques où le nombre de points appariés est très faible, et correspond à des scores de Grim proches de 0.5 où le rmsd de l'alignement des points matchés (à ne pas confondre avec le rmsd avant et après alignement) est une composante importante du score.

Pour pallier à ce problème, il est nécessaire de se rappeler que le recouvrement de graphes fournit l'ensemble des cliques maximales possibles, et donc plusieurs solutions pour l'alignement des points d'interactions. Une solution possible serait ainsi de regarder toutes les cliques, toujours ordonnées par GrScore décroissant, et de sélectionner la première possédant un rmsd 'post-alignement' inférieur à un certain seuil. Si aucune clique n'est trouvée, alors les modes d'interactions entre la référence et le ligand arrimé ne coïncident pas.

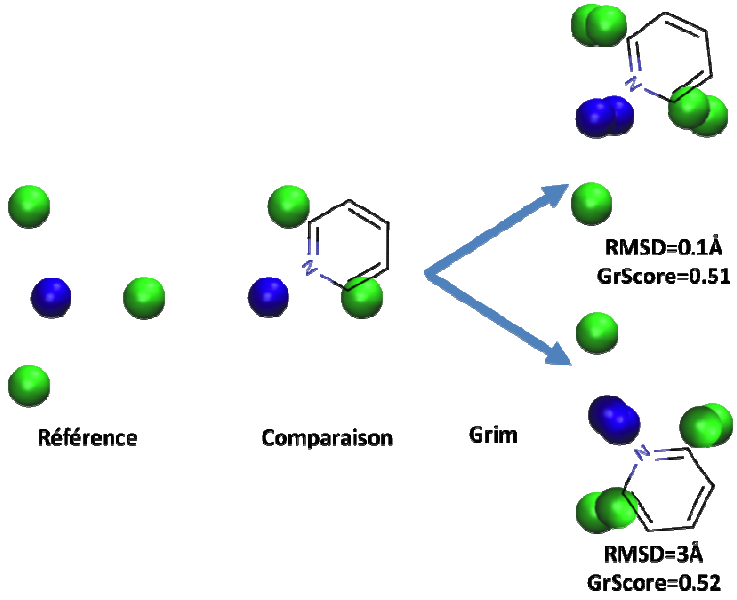


Figure 16 - Effet de l'alignement de Grim lors de la comparaison des modes d'interactions des poses de docking. En function des cliques obtenues, le meilleur score de l'alignement peut ne pas correspondre à la pose de docking initiale (en bas) mais être à un score plus faible (en haut)

Pour tester cette hypothèse, une post-analyse d'arrimage moléculaire a été réalisée avec Grim sur un jeu de 6660 poses. La distribution du rmsd en fonction du score de docking (**Figure 17**) montre très clairement un rmsd post-alignement faible lorsque l'on tend vers des GrScore plus important. Ainsi, pour un GrScore au dessus de 0.8, le rmsd ne dépasse pas les 0.06 Å, montrant la stabilité du processus d'alignement. Dans la gamme de GrScore de 0.5 - 0.7, le rmsd possède une grande amplitude, en allant de 0.04 Å à 1.8 Å. Deux cas de figure sont possibles : avec un rmsd élevé, le mode d'interaction de la pose ne correspond pas à sa position dans le site de liaison. Avec un rmsd faible, le mode d'interaction correspond à sa position dans le site de liaison, mais ne représente qu'une faible proportion du mode d'interaction du ligand de référence.

Ce traitement permet ainsi d'éliminer des poses scorées pour de mauvaises raisons, avec l'avantage de ne pas interférer dans la sélection des bonnes poses. Par conséquent, cela constitue un excellent critère pour filtrer des molécules.

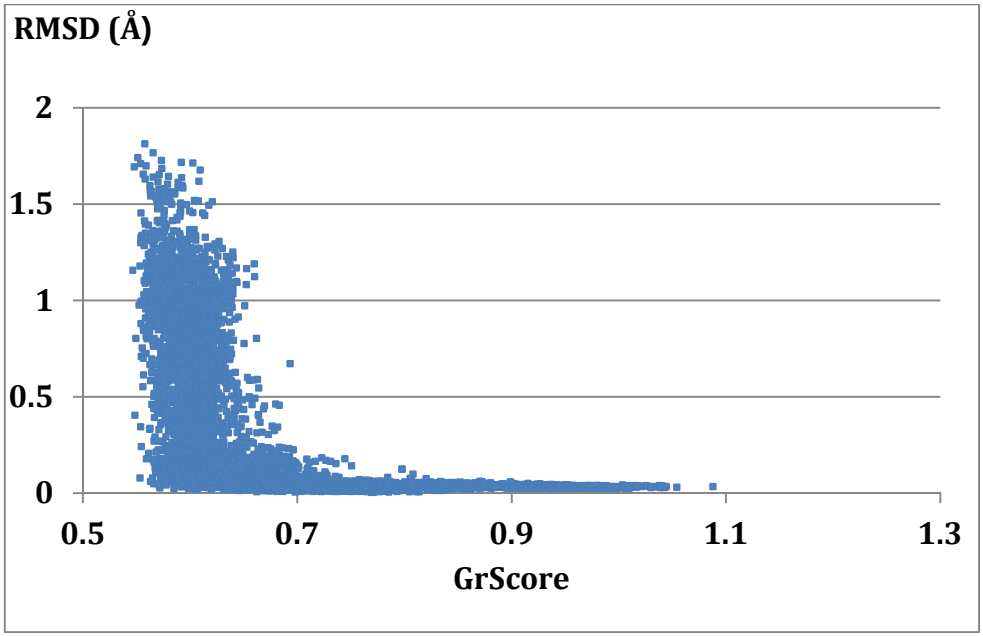


Figure 17 - Distribution des rmsd post-alignement en fonction du GrScore pour 6660 poses d'arrimage moléculaire

Une seconde amélioration concerne les interactions polaires. Le GrScore donne un poids important aux interactions polaires (0.97) appariées par rapport aux interactions apolaires (0.33). Cependant, on peut facilement constater que 3 interactions apolaires matchées équivalent à une interaction apolaire. Il nous est donc difficile de savoir si la similarité des modes d'interactions est essentiellement due à une reconnaissance de forme à travers les interactions apolaires, ou plutôt à une conservation des propriétés physico-chimiques. Lors d'un post-processing, il est intéressant de connaître les deux, mais il est surtout important de savoir la contribution de chacun d'eux. Ainsi, l'une des modifications importantes serait de connaître le nombre de points d'interactions polaires matchés comme complément du score de graphe.

7.4 CONCLUSION GENERALE

L'utilisation des modes d'interactions est une méthode basée sur la connaissance combinant à la fois l'information du ligand et de la protéine, en ne sélectionnant que les informations communes aux deux parties. Dans le cas d'une protéine, tous les résidus du site de liaison ne sont pas forcément en contact avec le ligand, et tous les atomes du ligand n'interagissent pas avec la protéine. De nombreuses améliorations sont possibles, tant sur les types d'interactions détectés que sur la comparaison de celles-ci. Il est cependant important de ne pas tomber dans une surexpression de l'information car cette méthodologie reste très dépendante de la cible observée. De plus, l'un des inconvénients lors de la comparaison réside dans la recherche de similarité des modes d'interactions, mais il serait intéressant de rechercher aussi la dissimilarité afin de trouver des interactions nouvelles ou des interactions manquantes sur lesquelles il serait possible de se focaliser.

8. BIBLIOGRAPHIE

(1) Deng, Z.; Chuaqui, C.; Singh, J. *J. Med. Chem.* **2004**, *47*, 337–44.

(2) Tan, L.; Batista, J.; Bajorath, J. *Chem Biol Drug Des* **2010**, *76*, 191–200.

(3) Willett, P. *Drug Discov. Today* **2006**, *11*, 1046–53.

(4) Mason, J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C.; Labaudiniere, R. F. *J. Med. Chem.* **1999**, *42*, 3251–64.

(5) Leslie, C.; Eskin, E.; Noble, W. S. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* **2002**, 564–75.

(6) Rognan, D. *Mol Inform* **2010**, *29*, 176–187.

(7) Schneider, G. *Nature reviews. Drug discovery* **2010**, *9*, 273–6.

(8) Crisman, T. J.; Sisay, M. T.; Bajorath, J. *J. Chem. Inf. Model.* **2008**, *48*, 1955–64.

(9) Tan, L.; Lounkine, E.; Bajorath, J. *J. Chem. Inf. Model.* **2008**, *48*, 2308–12.

(10) Bock, J. R.; Gough, D. a *J. Chem. Inf. Model.* **2005**, *45*, 1402–14.

(11) Weill, N.; Rognan, D. *J. Chem. Inf. Model.* **2009**, *49*, 1049–62.

(12) Wang, F.; Liu, D.; Wang, H.; Luo, C.; Zheng, M.; Liu, H.; Zhu, W.; Luo, X.; Zhang, J.; Jiang, H. *J. Chem. Inf. Model.* **2011**, *51*, 2821–8.

(13) Weill, N.; Valencia, C.; Gioria, S.; Villa, P.; Hibert, M.; Rognan, D. *Mol Inform* **2011**, *30*, 521–526.

(14) Chuaqui, C.; Deng, Z.; Singh, J. *J. Med. Chem.* **2005**, *48*, 121–33.

(15) Deng, Z.; Chuaqui, C.; Singh, J. *J. Med. Chem.* **2006**, *49*, 490–500.

(16) Marcou, G.; Rognan, D. *J. Chem. Inf. Model.* **2007**, *47*, 195–207.

(17) Kelly, M. D.; Mancera, R. L. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1942–51.

(18) Mpamhanga, C. P.; Chen, B.; McLay, I. M.; Willett, P. *J. Chem. Inf. Model.* **2006**, *46*, 686–98.

(19) Venhorst, J.; Núñez, S.; Terpstra, J. W.; Kruse, C. G. *J. Med. Chem.* **2008**, *51*, 3222–9.

(20) Pérez-Nueno, V. I.; Rabal, O.; Borrell, J. I.; Teixidó, J. *J. Chem. Inf. Model.* **2009**, *49*, 1245–60.

(21) Schreyer, A.; Blundell, T. *Chem Biol Drug Des* **2009**, *73*, 157–67.

(22) Weisel, M.; Bitter, H.-M.; Diederich, F.; So, W. V.; Kondru, R. *J. Chem. Inf. Model.* **2012**, *52*, 1450–61.

(23) Meslamani, J.; Rognan, D.; Kellenberger, E. *Bioinformatics* **2011**, *27*, 1324–6.

(24) Certara SYBYL **2012**.

(25) Rogers, D.; Hahn, M. *J. Chem. Inf. Model.* **2010**, *50*, 742–54.

(26) Accelrys Software Inc PipelinePilot **2012**.

(27) Durant, J. L.; Leland, B. a; Henry, D. R.; Nourse, J. G. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273–80.

(28) Chemical Computing Group MOE **2011**.

(29) Desaphy, J.; Azdimousa, K.; Kellenberger, E.; Rognan, D. *J. Chem. Inf. Model.* **2012**, *52*, 2287–2299.

(30) Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. *J. Chem. Inf. Model.* **1998**, *38*, 511–522.

(31) Tanabe, M.; Kanehisa, M. In *Current Protocols in Bioinformatics*; Wiley, Ed.; New York, 2012.

(32) Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T. M.; Mortenson, P. N.; Murray, C. W. *J. Med. Chem.* **2007**, *50*, 726–41.

(33) CCDC/Astex validation set for docking software.

(34) Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. *J. Med. Chem.* **2012**, *55*, 6582–94.

(35) Willett, P.; Barnard, J. M.; Downs, G. M. *J. Chem. Inf. Model.* **1998**, *38*, 983–996.

(36) Sofware, O. S. ROCS **2012**.

(37) Grant, J. A.; Gallardo, M. a.; Pickup, B. T. *J. Comput. Chem.* **1996**, *17*, 1653–1666.

(38) Grant, J. A.; Pickup, B. T. *J. Phys. Chem.* **1995**, *99*, 3503–3510.

(39) Nicholls, A.; Grant, J. A. *J. Comput. Aided Mol.* **2005**, *19*, 661–86.

(40) Triballeau, N.; Acher, F.; Brabet, I.; Pin, J.-P.; Bertrand, H.-O. *J. Med. Chem.* **2005**, *48*, 2534–47.

(41) Swamidass, S. J.; Azencott, C.-A.; Daily, K.; Baldi, P. *Bioinformatics* **2010**, *26*, 1348–56.

(42) Barillari, C.; Marcou, G.; Rognan, D. *J. Chem. Inf. Model.* **2008**, *48*, 1396–410.

(43) Bron, C.; Kerboscht, J. **1973**, *16*.

(44) Johnston, H. C. *Intl. J. Comput. Inf. Sci.***1976**, *5*, 209–238.

(45) Theobald, D. L. *Acta. Cryst.. Section A, Foundations of crystallography* **2005**, *61*, 478–80.

(46) Jain, A. N. *J. Med. Chem.* **2003**, *46*, 499–511.

(47) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. *J. Comput. Aided Mol.* **1997**, *11*, 425–45.

(48) Jain, A. N.; Nicholls, A. *J. Comput. Aided Mol.* **2008**, *22*, 133–9.

(49) Truchon, J.-F.; Bayly, C. I. *J. Chem. Inf. Model.* **2007**, *47*, 488–508.

(50) Wirth, M.; Zoete, V.; Michielin, O.; Sauer, W. H. B. *Nucl. Acids Res.* **2013**, *41*, D1137–43.

(51) Ujváry, I. *Pesticide Science* **1997**, *51*, 92–95.

(52) Schuffenhauer, A.; Gillet, V. J.; Willett, P. *J. Chem. Inf. Model.* **2000**, *40*, 295–307.

(53) Wagener, M.; Lommerse, J. P. M. *J. Chem. Inf. Model.* **2006**, *46*, 677–85.

(54) Kennewell, E. a; Willett, P.; Ducrot, P.; Luttmann, C. *J. Comput. Aided Mol.* **2006**, *20*, 385–94.

(55) Wood, D. J.; de Vlieg, J.; Wagener, M.; Ritschel, T. *J. Chem. Inf. Model.* **2012**, *52*, 2031–43.

(56) Moriaud, F.; Doppelt-Azeroual, O.; Martin, L.; Oguievetskaia, K.; Koch, K.; Vorotyntsev, A.; Adcock, S. a; Delfaud, F. *J. Chem. Inf. Model.* **2009**, *49*, 280–94.

Chapitre 4
Recherche de fragments bioisostériques

1. INTRODUCTION

De longues étapes d'optimisation et de tests sont nécessaires entre la découverte de la première molécule active et le début des études cliniques. En effet, il est très souvent nécessaire de modifier la structure de la molécule afin d'améliorer certaines propriétés. Ainsi, le remplacement d'une partie lipophile par une autre plus polaire peut améliorer la solubilité du composé par exemple. Rigidifier une partie de la molécule est quelque fois capital pour améliorer son affinité pour la cible ou améliorer sa métabolisation. Enfin, ces modifications peuvent être une étape importante pour rendre la molécule brevetable. Pour réaliser ces objectifs, il est généralement demandé au chimiste médicinal de modifier la structure chimique de la molécule sur la base de son expérience et du contexte de la protéine cible. Cependant, cette phase manuelle nécessite beaucoup de temps, d'argent et de connaissances pour arriver à terme, la rendant ainsi limitée et très dépendante du chimiste. Une alternative pour pallier à cette problématique est de développer et d'utiliser des méthodes in-silico, basées sur les résultats expérimentaux d'autres projets pharmaceutiques afin d'orienter les chimistes médicinaux vers les modifications structurales à apporter au composé afin d'obtenir la ou les propriétés désirées.

Nous avons pu observer au cours du chapitre précédent qu'il n'existait pas de corrélation entre la similarité de paires de ligand et la similarité de paires de mode d'interaction. Cela indique que pour une même protéine cible, des ligands structuralement différents peuvent réaliser des interactions similaires. Dans ces conditions, cela implique naturellement qu'à un niveau local, certains atomes ou groupes d'atomes sont capables d'interagir de façon similaire avec le site de liaison. Il est alors possible de moduler les propriétés physico-chimiques d'un composé en changeant une partie de la molécule, tout en possédant l'assurance de ne pas perdre le mode d'interaction initial. Ce concept de paires, appelé bioisostères, a été initialement proposé par Friedman en 1951.[1] Celui-ci a alors déclaré que deux atomes ou groupes d'atome sont considérés comme bioisostères si leur arrangement spatial et physico-chimique est similaire et s'ils confèrent le même type d'activité biologique.

Afin de réaliser une recherche exhaustive de ces bioisostères potentiels, il est tout d'abord nécessaire de décomposer la molécule (**Figure 1**). Dans le cadre de potentiels candidats médicaments, les atomes sont généralement agencés autour d'un noyau (« scaffold »),[2] généralement composé d'un ou plusieurs hétérocycles, qui sont substitués par d'autres atomes ou groupes d'atomes, appelé des substituants ou chaînes latérales[3] (**Figure 1B**). Il peut exister dans certains cas un lien (« linker ») au sein du noyau qui lie les cycles. Alternativement, ce noyau peut être représenté plus simplement en perdant l'information des types atomiques, ce que l'on appelle alors un squelette (« framework »). Ces représentations purement structurales peuvent aussi être physico-chimiques. On ne représente plus les molécules par leurs atomes mais par les interactions potentielles qu'ils peuvent effectuer avec une protéine cible (**Figure 1C/D**).[4] Ainsi, la discrétisation d'une molécule en un ensemble caractéristique est une problématique à part entière due à la grande diversité des molécules, et de l'enchevêtrement des cycles, des liens et des substituants. Dans le cadre de ce projet, lors de la fragmentation d'une molécule, certaines liaisons spécifiques sont clivées et les atomes indiquant cette séparation seront appelés ici les points d'attache.

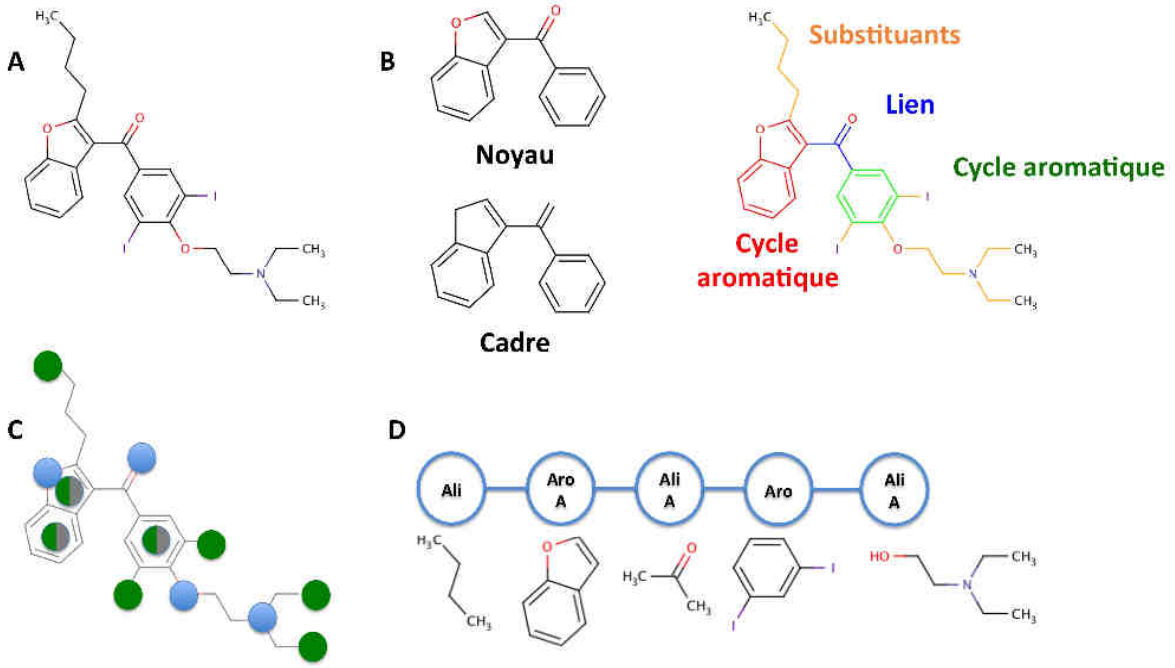


Figure 1- Différentes représentations de l'amiodarone. **A** : structure originelle. **B** : Le noyau représente le cœur aromatique de la molécule tandis que le squelette est une simplification du noyau en ignorant les types atomiques. La molécule peut aussi être dissociée suivant ses substituants, ses cycles aromatiques et le/les liens entre ces cycles. **C** : Représentation pharmacophorique de l'amiodarone. Les atomes ou groupes d'atomes sont caractérisés par les potentielles interactions qu'ils peuvent réaliser avec la protéine, tel que les contacts hydrophobes (en vert), des liaisons hydrogènes (en bleu) ou des interactions entre aromatiques (en gris). **D** : L'amiodarone représentée sous la forme d'un graphe de propriétés (Aro : aromatique, Ali : hydrophobe, A :liaison Hydrogène).

De la décomposition d'une molécule, il est alors possible d'observer l'occurrence de tels fragments dans les composés organiques. Plusieurs options ont été étudiées, en passant par l'analyse structurale des molécules jusqu'à l'analyse locale des sites de liaisons. BIOSTER est la première base de donnée récapitulant l'ensemble des bioisostères expérimentaux connus.[5] Basée sur une analyse manuelle de la littérature, cette base fournit en 2012 près de 25000 remplacements bioisostériques. Wagener et Lommersee[6] ont développé une procédure convertissant les fragments en une empreinte pharmacophorique et ont utilisé un sous-ensemble de BIOSTER pour valider leur protocole. Cette méthode permet de plus de distinguer des bioisostères issus du cœur de la molécule, du lien, ou des substituants. Schuffenhauer[7] a étudié la capacité de discrimination de paires bioisostériques par rapport à des leurres à travers des empreintes et des champs électrostatiques. Il a ainsi montré qu'il était possible d'observer des relations bioisostériques entre deux fragments à partir de ces méthodologies. Holliday[8] s'est basé sur la base de données BIOSTER pour décrire les substituants selon les propriétés physico-chimiques de chacun de ses atomes. La comparaison de ces propriétés entre deux substituants permet d'inférer une mesure de distance entre ces derniers afin de retrouver de potentiels bioisostères. L'utilisation de graphes condensés tels qu'observables dans la **Figure 1D** peut aussi être une méthode de recherche bioisostérique dans le cadre de criblage virtuel.[4]

La recherche de bioisostères peut aussi être réalisée à travers l'analyse de bases de données publiques de molécules. Ertl a ainsi traité plus de 3 millions de molécules à la recherche des substituants organiques les plus communs dans les molécules. Des propriétés physico-chimiques tels que l'hydrophobie, le volume, le nombre de donneurs et d'accepteurs de liaisons hydrogène ont ensuite été calculés pour chaque substituant afin de dériver une liste de bioisostères potentiels facilement interrogeable à travers une interface web.[9,10] De plus, Ertl a aussi étudié le remplacement des noyaux et des cycles aromatiques en analysant l'ensemble des molécules disponibles à travers des bases de données publiques.[7,11] Sa représentation des noyaux est ici purement structurale, et génère une chaîne de valeurs encodant le nombre d'atomes et de types de liaison, la distance topologique par rapport aux points d'attache des fragments, le coefficient de partage octanol/eau ou encore le nombre d'atomes et leurs propriétés associées en fonction de leurs distances aux points d'attache. De ces deux méthodes, Ertl a développé le logiciel IADE (Intelligent Automatic DEsign of bioisosteric analogs)[12] qui

fragmente un ligand de référence et qui génère à partir d'un ensemble de substitutions possibles, de nouvelles entités moléculaires aux propriétés analogues.

Le concept de paires moléculaires appariées (MMPs) consiste dans l'observation et la prédiction des modifications physico-chimiques et biologiques liées au changement d'un atome ou groupe d'atomes au sein d'une paire de molécules. Griffen illustre ce principe en observant l'effet de la substitution d'une partie de la molécule sur l'activité du composé dans le cas de diverses séries chimiques.[13] SwissBioisostere est un exemple de base de données de remplacement moléculaire[14] utilisant le principe des MMPs. Cette base, qui contenait en 2012 près de 6 millions de remplacements, permet d'obtenir des statistiques sur les modifications de chaque fragment, leur effet sur l'activité, le coefficient de partage octanol/eau, la surface polaire ainsi que la masse moléculaire. Cependant, les temps de calcul augmentent très vite lorsque la masse du fragment est importante.

Une des applications des MMPs est l'interface VAMMPIRE.[15] Celle-ci utilise les complexes cristallographiques protéine/ligand pour lesquelles une activité est fournie ainsi que l'ensemble des molécules de ChEMBLdb actives envers l'une des cibles. Ces dernières sont sélectionnées, alignées et une minimisation d'énergie est réalisée si elles diffèrent par seulement un atome avec la molécule co-cristallisée avec la cible protéique. L'environnement protéique local est alors enregistré pour chaque molécule autour de la zone substituée. Une interface web permet de sélectionner un fragment pour lequel toutes les substitutions connues, leurs environnements protéiques associés et leurs effets sur l'activité seront affichés. Une telle méthodologie possède l'avantage de ne plus se limiter à une représentation purement structurale des bioisostères mais aussi de permettre une analyse plus fine de l'environnement protéique lié à de telles modifications.

La recherche de bioisostères peut aussi être réalisée à partir des pharmacophores d'interaction protéine/ligand. Ainsi, on ne considère plus la structure du fragment mais les interactions qu'il effectue avec la protéine. Wood et Wagener ont développé KRIPO, *Key Representation of Interactions in POckects,* qui est une méthode basée sur les pharmacophores pour rechercher de nouveaux bioisostères.[16] La méthode de fragmentation clive les molécules suivant toutes les liaisons acycliques non terminales et génère toutes les combinaisons possibles de fragments.[17] Basée sur la PDB,

cette méthode analyse les interactions protéine/ligand afin de générer des pharmacophores d'interactions. Ces derniers, convertis en empreinte à partir de triplets permettent de réaliser facilement des mesures de similarité de mode d'interactions entre deux fragments. Lorsqu'au moins 5 points pharmacophoriques sont appariés entre deux fragments, ces derniers sont considérés comme bioisostériques puis alignés.

Afin de mieux orienter la recherche de bioisostères, nous proposons ici de ne plus utiliser l'information structurale des fragments mais de se focaliser sur leur mode d'interactions. En effet, puisque des fragments bioisostériques partagent le même mode d'interaction, la comparaison de ces derniers permet d'ouvrir la perspective de trouver des nouvelles paires de fragments, de masses moléculaires plus ou moins différentes répondant aux critères de bioisostérie.

Dans une première partie, nous décrirons un nouveau protocole de fragmentation basé sur le clivage de liaisons autour de cycles aromatiques ou aliphatiques. La deuxième partie se concentrera sur l'analyse des modes d'interaction de chaque fragment, la comparaison de ces interactions, ainsi que l'alignement des fragments induit par cette approche. Enfin, nous discuterons d'une nouvelle interface web spécialement développée dans cette optique.

2. MATERIEL ET METHODES

2.1 CONSTRUCTION DE LA BASE DE DONNEES DE FRAGMENTS

2.1.1 PREPARATION DES COMPLEXES PROTEINE/LIGAND

La sc-PDB (screening-PDB)[18] est une banque de dépôts de complexes protéine/ligand construite à partir de la PDB (Protein Data Bank).[19] Partant des données natives de la PDB, un premier filtre permet de sélectionner les entrées :

- Avec une résolution cristallographique inférieure à 2.5Å,
- Possédant au moins une chaine protéique composée d'un minimum de 35 acides aminés
- Incluant au moins une molécule de faible masse avec les caractéristiques suivantes :
    - une masse molaire comprise entre 140 et 810 g/mol
    - au moins un atome de carbone, un d'oxygène ou d'azote,
    - moins de 20 angles dièdres,
    - une surface enfouie dans la protéine à plus de 50%.

Pour chaque entrée, une première étape d'analyse convertit la sélénométhionine et la sélénocystéine en méthionine et cystéine, respectivement. Lorsqu'un atome ou résidu possède plusieurs taux d'occupation, seules les coordonnées correspondant au meilleur taux sont conservées. Une seconde étape consiste dans l'identification des molécules à partir du code HET de leurs résidus. Cela permet d'éliminer les groupes prosthétiques tel que Heme ou les clusters Fer-Soufre, de détecter les peptides (composés par 8 acides aminés au plus), les cofacteurs, la protéine (composée d'au minimum 8 acides aminés) et les molécules non polymériques. Seuls ces dernières, les cofacteurs et les peptides sont considérés par la suite comme ligand. S'il n'existe qu'un cofacteur et aucune molécule synthétique ou peptide dans l'entrée, le cofacteur est sélectionné comme ligand. Dans le cas contraire, le peptide ou la molécule synthétique sera le ligand et le cofacteur restera cofacteur.

Lorsque les entrées sont traitées et leur ligand sélectionné, le site de liaison entourant le ligand est extrait en sélectionnant tous les acides aminés possédant au

moins un atome lourd à moins de 6 Å d'un atome lourd du ligand. Les cavités du site de liaison sont générées avec le logiciel VolSite afin d'inférer une valeur de droguabilité. Toute cavité considérée par la méthode comme non droguable, c'est à dire avec une valeur de prédiction négative, est éliminée.

Les fichiers natifs de la PDB contiennent peu d'information sur la connectivité et aucune information sur la géométrie ou l'état d'hybridation des atomes. Dans un premier temps, toutes les paires d'atomes dont la distance est inférieure à la somme des rayons de van der Waals sont considérées comme étant liées par une liaison covalente. Une tolérance de 0.4 Å est ajoutée afin de prendre en compte les approximations liées aux structures cristallographiques, ainsi que les modifications éventuelles liées aux effets électroniques. Dans un second temps, une procédure d'appariement de graphes est appliquée entre chaque résidu de l'entrée et un patron de ce résidu. Pour ce dernier, les codes SMILES[20] de chaque code HET sont récupérés à partir du fichier d'information de la cristallographie macromoléculaire (mmCIF), puis sont convertis au format MOL2 avec le logiciel dbtranslate de Tripos.[21] L'état d'ionisation est corrigé avec Filter d'OpenEye,[22] et les types MOL2 de chaque atome sont vérifiés avec le logiciel GOLD.[23] Chaque ligand des complexes protéine/ligand est comparé à son patron via une procédure d'appariement de graphe afin d'inférer le type MOL2 de chaque atome, ainsi que le type de liaison : liaison simple, double, triple, aromatique, amide.

L'ajout des hydrogènes est effectué avec SybylX 2.0 et l'orientation des hydrogènes polaires avec le logiciel Hyde de BioSolveIt. Si des molécules d'eau sont présentes au sein du site de liaison, seules celles réalisant au moins deux liaisons hydrogène avec la protéine sont gardées. La protéine, le ligand et le site sont enregistrés pour chaque entrée au format MOL2 dans des fichiers séparés. Au final, ce sont 8077 entrées comprenant 2377 protéines distinctes et 5233 ligands distincts qui ont été sélectionnées pour générer la version 2012 de la base sc-PDB.

2.1.2 FRAGMENTATION

Tous les ligands de la sc-PDB 2012 ont été fragmentés selon la procédure décrite par la **Figure 2**. Cette méthode s'inspire d'un algorithme développé par Brenk dans le cadre de la recherche de cœurs interagissant avec le hinge des protéines kinases.[24] Tous les cycles, aromatiques ou aliphatiques, sont détectés selon un algorithme de réduction de graphes.[25] Dès lors que deux cycles partagent un atome en commun, ils sont fusionnés. Chaque cycle est alors pris comme référence et ses substituants sont analysés selon 2 critères. Dans un premier cas, le substituant ne contient aucun cycle et reste lié au cycle de référence. Dans un second cas, le substituant possède au moins un cycle et le substituant est alors clivé au niveau de la première liaison entre deux atomes apolaires. Si aucune liaison n'est clivée, alors la liaison avec l'atome en alpha du cycle du substituant est clivée. Dès lors, la liaison est rompue, et chaque atome est remplacé dans son fragment opposé par *un point d'ancrage*, que l'on nomme Z. Si 2 fragments sont identiques et représentent les mêmes atomes, on ne garde qu'une seule occurrence du fragment. Ce protocole appliqué à l'ensemble des ligands de la sc-PDB fournit 35081 fragments.

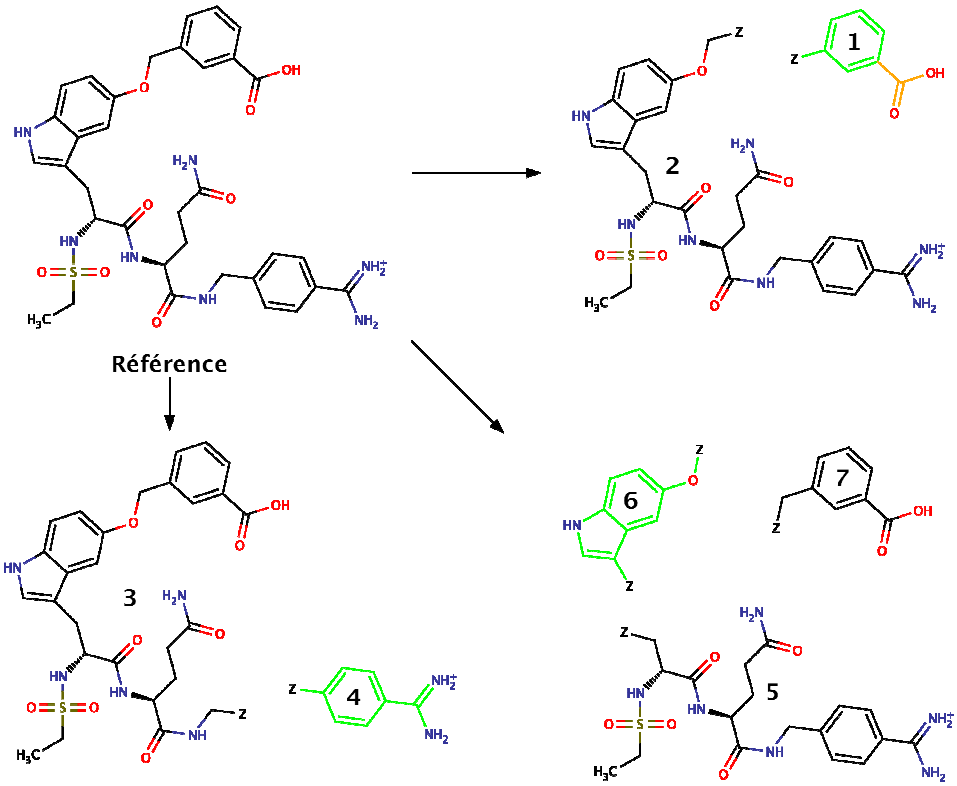


Figure 2 - Exemple de fragmentation de l'acide 3-[[3-[(2R)-3-[[(2S)-5-amino-1-[(4-carbamimidoylphenyl) methylamino]-1,5-dioxo-pentan-2-yl]amino]-2-(ethylsulfonylamino)-3-oxo-propyl]-1H-indol-5-yl]oxymethyl] benzoique. A gauche le ligand de référence. Chaque cycle utilisé comme référence est représenté en vert. Cas du fragment 1 : l'acide carboxylique (en orange) n'étant pas un cycle, il est rattaché au cycle de référence

2.1.3 ATTRIBUTION DES MODES D'INTERACTIONS

Pour chaque fragment, les interactions non-covalentes qu'il effectue avec sa protéine cible sont détectées en utilisant l'outil *ints* de la suite **IChem**,[26] avec les options par défaut (**Tableau 1**). La représentation *Merged*, incluant *InterLig*, *Centered* et *InterProt*, est enregistrée pour chaque fragment au format MOL2. Les empreintes d'entiers TIFP sont aussi générées avec l'outil fgps de la suite **IChem** en utilisant les paramètres par défaut. Par conséquent, une chaine de 210 entiers est sauvegardée pour chaque fragment.

Seuls les fragments réalisant au moins 5 interactions non-covalentes et au moins 2 interactions aromatiques ou polaires sont gardés, réduisant ainsi le nombre de fragments à 19002.

Tableau 1 - Critères géométriques d'angle et de distance pour chaque interaction non-covalente détectée entre la protéine et le ligand

Interaction	Distance (Å)	Angle (degrés)
Liaison Hydrogène	3.5	180 ± 60
Contact Apolaire	4.5	/
Liaison Ionique	4.0	/
Metal/Accepteur	2.8	/
Aromatique Face/Face	4	180 ± 30
Aromatique Côté/Face	4	90 ± 60

2.1.4 CONVERSION DES FRAGMENTS EN 2-DIMENSIONS

Afin d'obtenir une représentation réaliste des fragments en 2 dimensions qui n'est pas biaisée par les points d'ancrage, il est nécessaire de réaliser la conversion en plusieurs étapes. Tout d'abord, les fragments, initialement en 3-dimensions au format MOL2 sont convertis au format SDF avec SybylX. Le logiciel Corina[27] est ensuite utilisé afin de transformer les fragments en une représentation en 2 dimensions. L'état de protonation est corrigé par le logiciel Filter d'OpenEye.[22] Enfin, une analyse visuelle de chaque fragment permet de corriger de possibles erreurs de valence ou d'état de protonation, qui sont ensuite appliqué à tous les fragments avec le logiciel Standardizer de ChemAxon.

2.2 MESURE DE SIMILARITE

2.2.1 SIMILARITE STRUCTURALE

Les empreintes circulaires ECFP4[28] des 19002 fragments en deux dimensions sont générées avec le logiciel Pipeline Pilot.[29] Les empreintes ECFP4 sont ensuite converties en une chaine binaire de 1024 bits. Puisque ce logiciel ne considère pas les atomes Z correspondants à nos points d'ancrages, ces derniers sont d'abord remplacés par des atomes de carbone. La composante « Merged Molecules » basée sur la comparaison de chaines SMILES canoniques permet d'obtenir la liste des fragments distincts, au nombre de 7564. La similarité entre deux empreintes ECFP4 est calculée entre tous les fragments avec le coefficient de Tanimoto[30] :

$Tc = \frac{c}{a+b-c}$ où a : Nombre de bits égal à 1 dans la chaîne de référence
b : Nombre de bits égal à 1 dans la chaîne de comparaison
c : Nombre de bits égal à 1 et commun aux deux chaînes

Le remplacement des points d'ancrage n'est réalisé que pour la génération des empreintes ECFP4 et n'est pas utilisé dans la suite du matériel et méthode.

2.2.2 SIMILARITE DE MODES D'INTERACTION

Pour les empreintes TIFP, 2 valeurs de similarité sont calculées en utilisant le coefficient de Tanimoto :

$$Tc = \frac{\sum_{j=1}^{N} w_j x_{jA} x_{jB}}{\sum_{j=1}^{N} w_j \left(x_{jA}\right)^2 + \sum_{j=1}^{N} w_j \left(x_{jB}\right)^2 - \sum_{j=1}^{N} w_j x_{jA} x_{jB}}$$

où x_{jA} correspond à la valeur de la case j dans la chaine d'entier de référence A, x_{jB} la valeur de la case j dans la chaine d'entier de la comparaison B et w_j le poids associé à chaque case j. La première valeur, TIFP, mesure la similarité globale (avec des poids définis à 1 pour chaque case) tandis que la seconde, nommée TIFPPol, affecte des poids inversement proportionnels aux nombres d'hydrophobes dans le triplet : 1 pour un triplet purement hydrophile, 0.8 pour un triplet avec 1 point hydrophobe, 0.6 avec 2 points hydrophobe et 0.4 pour un triplet lipophile.

De plus, tous les modes d'interactions *Merged* sont comparés avec la méthode Grim (GRaph Interaction Matching)[26] en utilisant les paramètres par défaut.

2.3 INTERFACE SCPDB-FRAG

L'ensemble des fragments, de leurs interactions, et des mesures de similarité est rassemblé sous la forme d'une base de données facilement interrogeable à travers une interface web : scPDB-Frag. Celle-ci permet entre autre de rechercher pour un fragment d'intérêt l'ensemble de ses bioisostères potentiels, c'est à dire l'ensemble des fragments structuralement différent selon les empreintes ECFP4 mais partageant le même mode d'interaction selon les empreintes TIFP, TIFPPol et/ou Grim.

2.3.1 BASE DE DONNEES

Les données sont actuellement divisées en 2 parties : la base de données relationnelle ainsi que les données brutes de mesures de similarité.

La base de données, implémentée sous MySQL, a pour rôle de répertorier l'ensemble des fragments existants avec les caractéristiques structurales, d'interactions et de nomenclatures. Elle est constituée de 9 tables dont le descriptif est fourni dans la **Figure 3**.

L'aspect structural est entièrement géré par la suite JChem Base de ChemAxon. Celle-ci permet d'enregistrer la structure des fragments et de réaliser des recherches structurales dans l'interface web. JChemProperties, JChemProperties_CR sont deux tables propres à JChem Base tandis que FragList et FragList_UL sont les tables créées pour nos besoins. Tous les fragments structuralement uniques générés dans la partie **2.2.1** sont chargés dans la table FragList via JChemManager. Etant donné que les fragments peuvent être présents de multiples fois au sein des divers ligands de la sc-PDB, cette procédure minimise le coût en terme d'espace disque.

Toutes les données des modes d'interactions des 19002 fragments sont sauvegardées dans la table FragToPDB. Celle-ci contient l'identifiant du fragment, son code PDB associé, l'identifiant du fragment unique (celui de FragList) ainsi que le nombre d'interactions non covalentes par type d'interaction.

En dernier lieu, la table scPDB contient les informations relatives à chaque entrée de la sc-PDB, incluant le numéro E.C., le nom sc-PDB, l'identifiant UniProt, liant ainsi le fragment à sa cible. La table EC_Name apporte des informations sur la nomenclature des enzymes, tandis que KEGG_Classif fourni une classification des protéines. Une dernière table permet d'obtenir des statistiques sur le nombre de visites sur l'interface web.

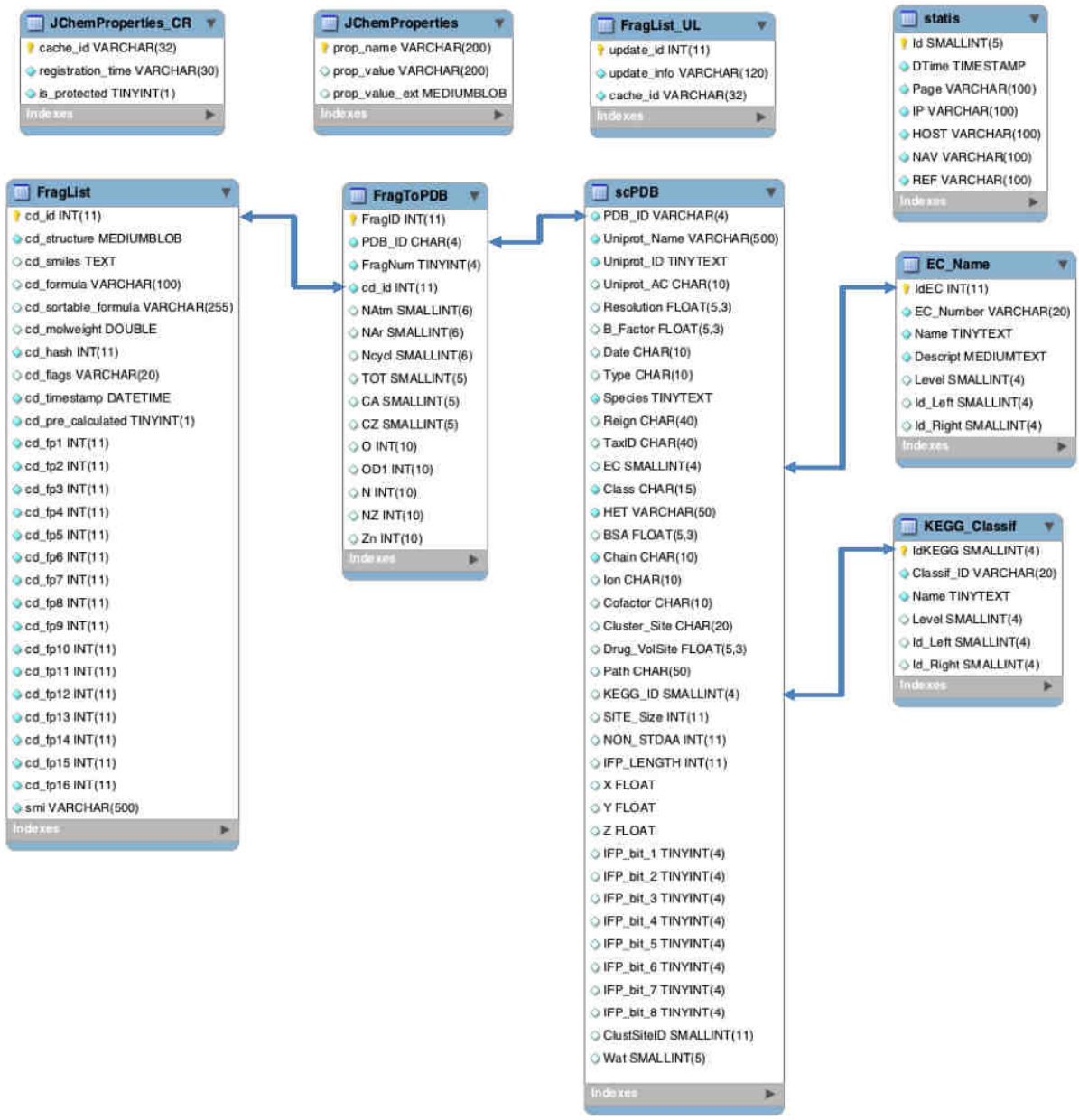


Figure 3 - Représentation schématique de la base de données scPDB-Frag. Chaque table de la base est décrite par la liste de ses éléments. Les liens en bleu correspondent aux contraintes d'intégrité permettant d'assurer une cohérence dans les données.

Pour des raisons de gestion et d'efficacité, les mesures de similarité tant structurale que des modes d'interactions n'ont pas été enregistrées dans la base de données mais sous forme de fichiers. Ainsi, pour tous les 19002 fragments possédant un nombre suffisant d'interactions, un fichier est généré pour chaque fragment et contient l'ensemble des valeurs de similarité des empreintes ECFP4, Grim, TIFP et TIFPPol par rapport aux autres fragments.

2.3.2 TECHNOLOGIE EMPLOYEE

Dans sa version actuelle, l'interface tourne sur Apache version 2.0 et utilise MySQL 5, Apache Tomcat 6, HyperText Preprocessor version 5.3 (PHP), Java, Javascript. L'interface web est réalisée selon un patron modèle-vue-contrôleur via vTemplate 1.3.3 (http://vtemplate.sourceforge.net) en PHP afin de simplifier son développement et ses futures mises à jour. Tous les scripts sont écrits en PHP/HTML. L'ensemble des données concernant les fragments est géré à travers le Système de Gestion de Base de Données (SGBD) MySQL (http://www.mysql.fr/). JQuery (http://jquery.com/) et JQuery-UI (http://jqueryui.com/) fournissent un aspect dynamique à l'interface. Deux modules externes sont employés pour dessiner les fragments et visualiser l'alignement de bioisostères : MarvinSketch 6.0.4 de ChemAxon (http://www.chemaxon.com/) et OpenAstexViewer 3.0 d'Astex Therapeutics Ltd (http://openastexviewer.net/web/) respectivement. Java et Apache Tomcat sont utilisés afin de réaliser des recherches de sous-structures via MarvinSketch. Les tableaux de résultats ainsi que les options d'ordonnancement, de filtrage et d'export sont implémentés en utilisant DataTables v1.9.4 développé par Allan Jardine (http://www.datatables.net/). L'exportation des résultats au format Microsoft Excel est réalisée via une librairie collaborative PHPExcel (http://phpexcel.codeplex.com/).

2.3.3 DESCRIPTION DE L'INTERFACE

L'interface scPDB-Frag est actuellement accessible via http://bioinfo-pharma.u-strasbg.fr/scPDB-Frag. Cette base de données est interrogeable selon le principe de l'assistant (**Figure 4**). L'utilisateur part d'un fragment et souhaite trouver un bioisostère de celui-ci. Les deux premières étapes permettent à l'utilisateur de sélectionner le fragment d'intérêt. Pour cela, il possède le choix entre dessiner le fragment, le sélectionner via le code PDB de sa protéine cible, le code HET du ligand complet ou encore le nom sc-PDB dérivé du nom Uniprot. De ces critères, une liste de fragments est proposée parmi les 19002 fragments disponibles dans la base. La troisième étape consiste dans les critères de sélection des bioisostères potentiels : similarité structurale (ECFP4), des modes d'interactions globaux (GrScore/TIFP) ou polaires (TIFPPol) et de la masse moléculaire. L'utilisateur sera amené à choisir parmi la liste proposée de fragments celui qui lui convient. En dernier lieu, un alignement sera réalisé entre le mode d'interaction *Merged* du fragment de référence et celui du fragment de

comparaison en utilisant Grim. Cet alignement est ensuite appliqué aux fragments afin que les interactions et les fragments soient visualisables sous OpenAstexViewer.

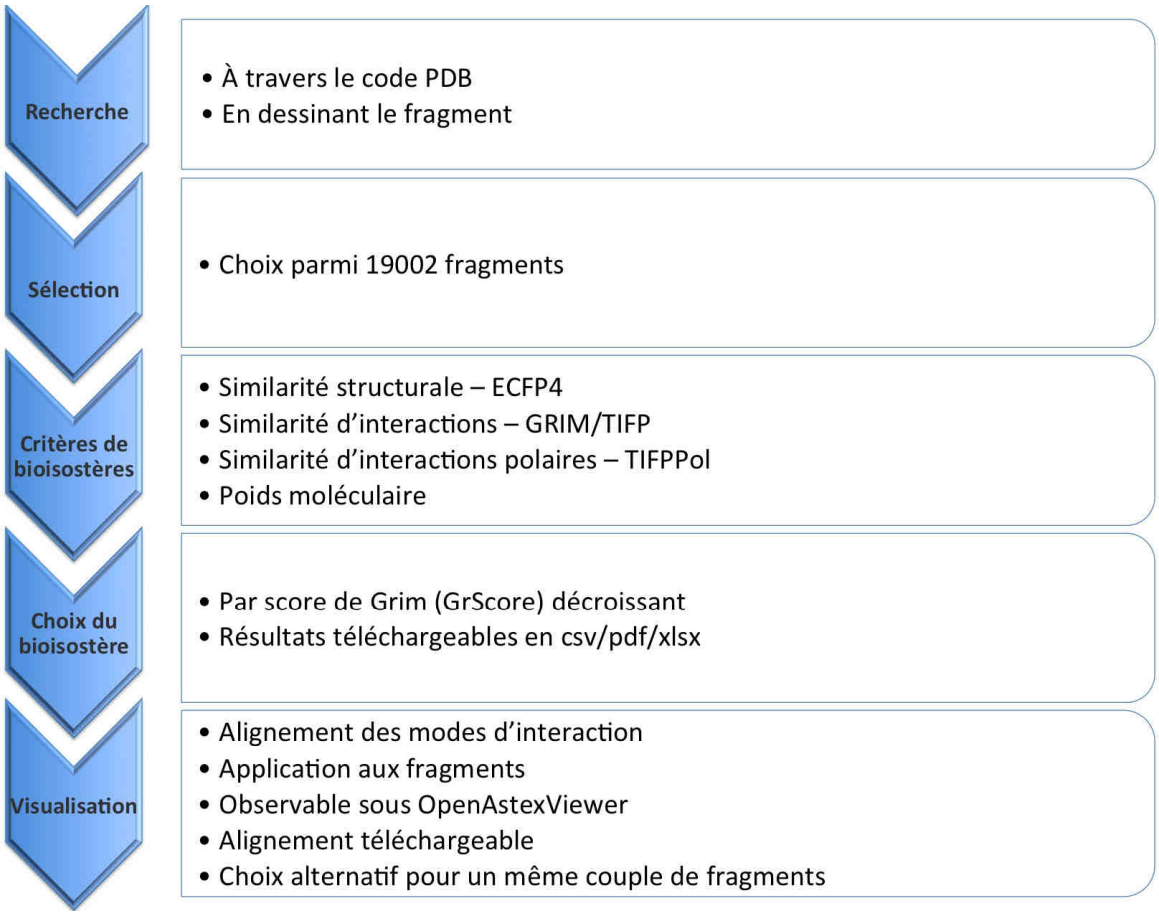


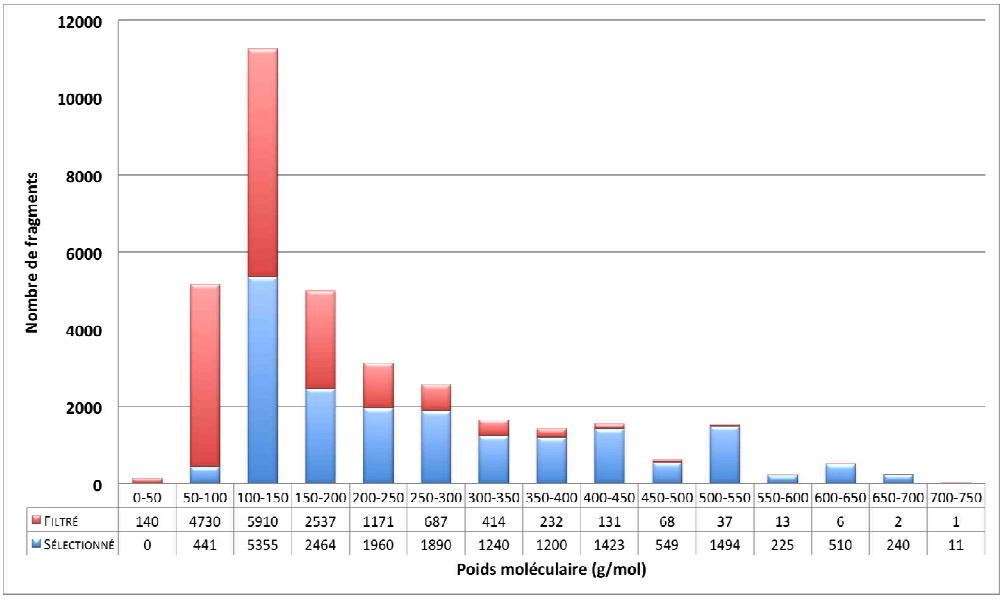
Figure 4 - Aperçu du processus côté utilisateur de l'interface scPDB-Frag. Celle-ci se compose de 5 étapes et offre diverses options suivant les cas de figure.

3. RESULTATS ET DISCUSSIONS

3.1 FRAGMENTATION

La fragmentation est un processus très conceptuel qui dépend de l'observation que l'on veut réaliser. La meilleure fragmentation possible serait de générer toutes les combinaisons possibles d'atomes en ajoutant un atome à la fois, en conservant les groupements chimiques caractéristiques. Cependant, la combinatoire tendrait très rapidement vers des valeurs exponentielles, rendant l'analyse très fastidieuse.

Dans le cas présent, nous souhaitons trouver des fragments réalisant un nombre d'interactions non-covalentes suffisant pour coder à la fois la forme générale du fragment ainsi que ses contributions dans la reconnaissance moléculaire. Pour cela, il est nécessaire de posséder des contacts hydrophobes mais aussi des interactions polaires afin d'orienter l'alignement. La **Figure 5** montre la distribution des fragments en fonction de leur masse. Le choix de se focaliser sur des cycles permet d'obtenir un large éventail de fragments, avec une gamme de masse allant de 40 à 725 g/mol, une vaste majorité se situant autour de 100 à 350 g/mol. On observe aussi que notre règle de filtrage à partir des interactions élimine la plupart des fragments de faible masse. Les fragments éliminés possédant une grande masse correspondent à des parties non enfouies dans la protéine et n'interagissant donc pas avec celle-ci.



	0-50	50-100	100-150	150-200	200-250	250-300	300-350	350-400	400-450	450-500	500-550	550-600	600-650	650-700	700-750
Filtré	140	4730	5910	2537	1171	687	414	232	131	68	37	13	6	2	1
Sélectionné	0	441	5355	2464	1960	1890	1240	1200	1423	549	1494	225	510	240	11

Figure 5 - Répartition des fragments en fonction de leur masse (g/mol). Les fragments filtrés pour manque d'interactions sont représentés en rouge, tandis que ceux sélectionnés sont en bleu.

Tandis que 5928 fragments sont des singletons, les dérivés de nucléotides (**Figure 6**) sont les plus représentés au sein de la base de données en raison d'un biais de la sc-PDB. Le fragment benzénique, pourtant présent 1239 fois, ne respecte jamais le nombre minimum d'interactions, et n'est donc jamais sélectionné.

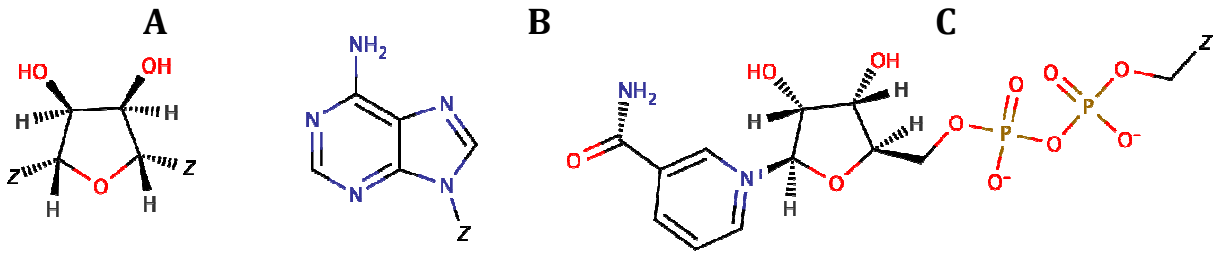


Figure 6 - Structure chimique des 3 fragments les plus présents dans la base de données. A : le (3R,4S)-oxolane-3,4-diol (n=1375) ; B : l'adénine (n=1112); C : Nicotinamide ß-D-ribose diphosphate (n=342).

3.2 DIFFICULTES LIES AUX POINTS D'ANCRAGES

Les points d'ancrage ont été particulièrement difficiles à traiter lors des divers processus de traitement. En effet, leurs caractéristiques sont bivalentes puisqu'ils ne représentent pas un atome existant dans le fragment mais décrivent un atome existant au sein de la molécule complète. Les chaines ECFP4, encodant la structure de la molécule, ne considèrent pas les atomes Z et nécessitent par conséquent leur remplacement par un autre atome. Initialement, le premier remplacement proposé était l'atome de Silicium, atome non présent biologiquement mais existant au sein du tableau périodique (en comparaison d'un atome Z). Cependant, celui-ci a entraîné lors de la conversion des fragments de 3-dimensions à 2-dimensions une très mauvaise représentation, modifiant la position de liaison double et créant des charges positives ou négatives là où elles ne devraient pas exister. Cette mauvaise définition a impliqué des valeurs de similarité structurale fausses et par conséquent une recherche de bioisostères impossible. L'une des solutions initiale était alors de corriger manuellement l'ensemble des structures mais les valeurs de similarité d'ECFP4 sont restées aberrantes. Finalement, le remplacement des points d'ancrage par des atomes de carbone s'est avéré être le meilleur compromis possible. Cette altération biaise légèrement les valeurs d'ECFP4 puisque l'on perd l'information du point d'ancrage. Si l'on considère par exemple deux fragments analogues du 3,4-diméthylphénol (**Figure 7**) où le point d'ancrage sera dans un cas en position meta et dans l'autre en position para du groupement hydroxyle, alors la conversion du point d'ancrage en carbone conduira les

deux fragments à avoir une valeur de similarité structurale de 1, ce qui n'est pas exact. Cependant, il est nécessaire de garder à l'esprit que cette approximation ne peut arriver que dans les cas où les fragments sont d'ores et déjà très similaires. Par conséquent, cette simplication n'influence pas la recherche de fragments structuralement différents tel que souhaité pour des bioisostères.

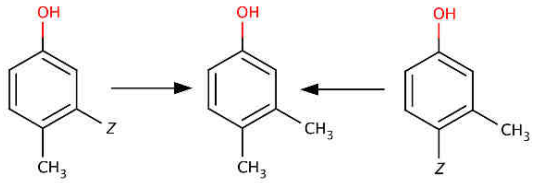


Figure 7 - Illustration de la problématique liée la conversion des points d'ancrage. Le 2-Z-3-methylphenol (à gauche) et le 2-methyl-3-Z-phenol (à droite) seront, lors de la transformation du point d'ancrage en méthyle, considérés comme le 2,3-diméthylphénol.

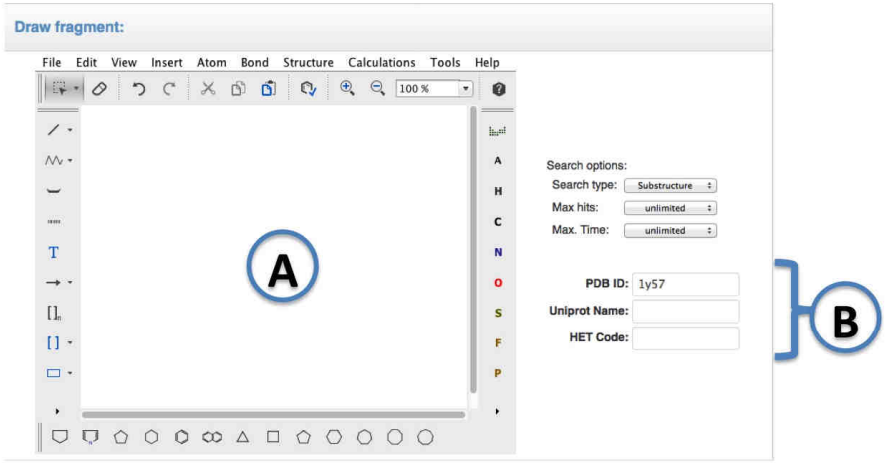
3.3 EXEMPLE DE RECHERCHE

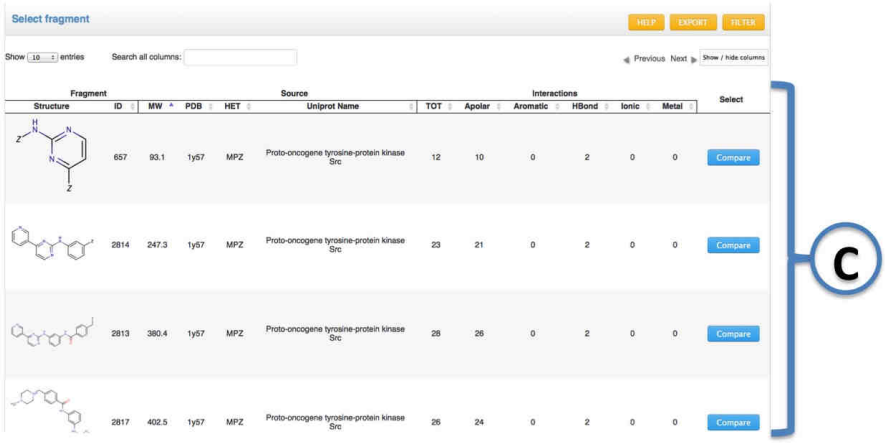
L'interface actuelle autorise l'utilisateur à réaliser deux types de recherches : focalisée ou exhaustive. Dans le premier cas, l'utilisateur connaît le fragment qu'il souhaite étudier, et dans quel contexte (protéine cible, ligand) il se trouve. Dans le second cas, l'utilisateur dessine la structure de la molécule qu'il souhaite trouver comme fragment dans la base de données et effectue ensuite sa recherche à partir de celui-ci.

3.3.1 RECHERCHE FOCALISEE

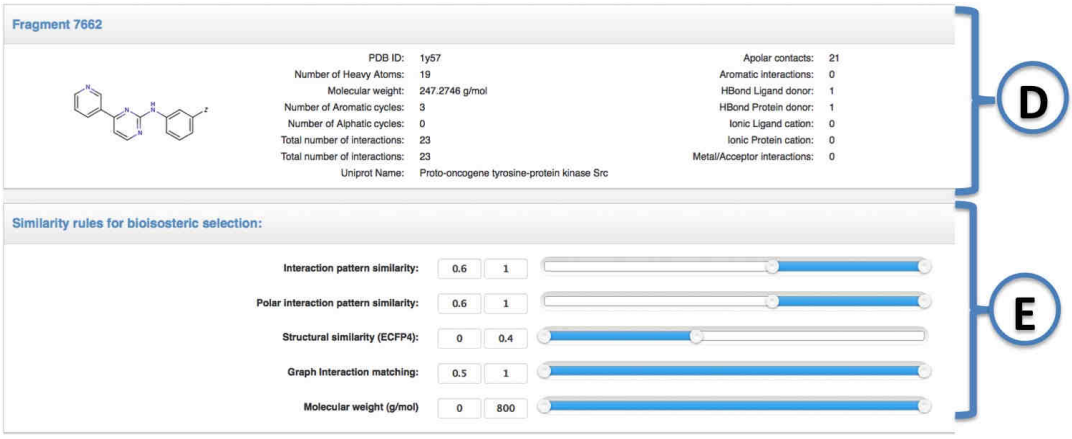
Partant d'une structure, d'un code PDB ou HET du fragment désiré, l'utilisateur est amené à naviguer dans une série de 6 étapes afin de sélectionner son fragment, les fragments considérés comme bioisostériques, et visualiser leur alignement (**Figure 8**). L'entrée PDB proposée ici est la même que celle du troisième chapitre et concerne la Proto-oncogène tyrosine-protéine kinase Src (Code PDB : 1y57). Le protocole de fragmentation ayant été modifié depuis la publication, les fragments de référence ne sont pas identiques. Cependant les résultats proposés restent sensiblement identiques si ce n'est meilleurs. Sur les 12 fragments générés pour cette molécule, seuls 4 réalisent un nombre d'interactions suffisant avec la protéine pour être considérés (**Figure 8**C). Une fois le fragment sélectionné, l'interface propose de définir un ensemble de critères de similarité pour rechercher des fragments bioisostériques (**Figure 8.E**).

SÉLECTION DU FRAGMENT





SÉLECTION DU BIOISOSTÈRE



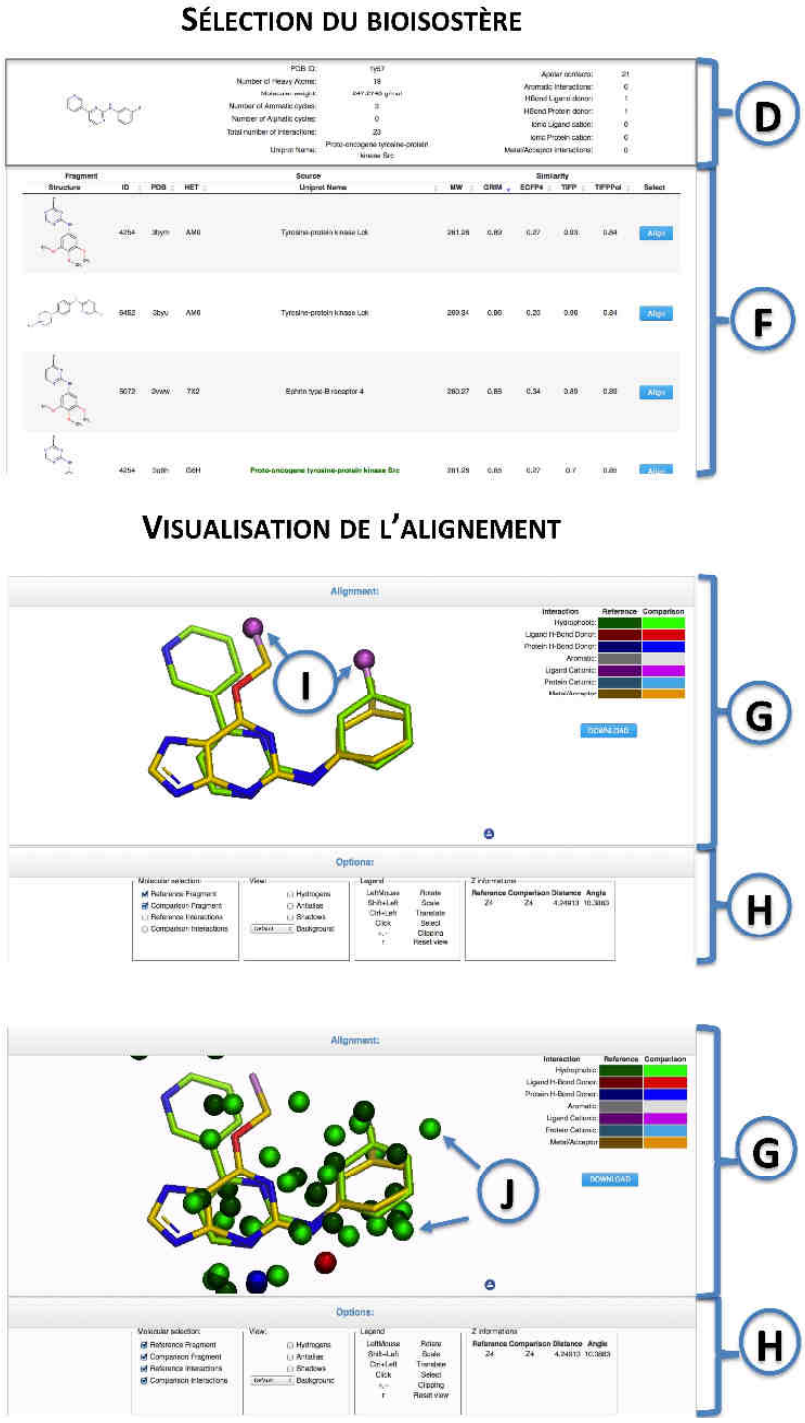


Figure 8 - Détail de la procédure de recherche de bioisostères pour le ligand d'un complexe de la PDB. En premier lieu, l'utilisateur est amené à sélectionner le fragment qui l'intéresse. Il peut soit dessiner le fragment de son choix (A) ou directement rentrer le code PDB d'intérêt (B), ici la Proto-oncogène tyrosine-protéine kinase Src (Code PDB : 1y57). L'interface fournie l'ensemble des fragments associé au ligand du code PDB fourni (C). En second lieu, l'utilisateur définit les critères de similarité du bioisostère voulu. Les propriétés du fragment initialement choisi comme référence sont récapitulées dans l'encart (D). En fonction des propriétés, il peut définir des règles de similarité structurale, des modes d'interactions et de la masse moléculaire (E) pour choisir le bioisostère. Une liste de fragments potentiels est alors proposée (F) et l'utilisateur est invité à choisir le fragment de comparaison, ici un fragment d'un ligand en complexe avec un kinase cycline-dépendante 2 (cdk2, code PDB : 1h1r). Enfin, l'alignement des fragments est réalisé à travers celui des modes d'interactions (G). De nombreuses options de visualisation sont disponibles (H). D'autre part, les points d'ancrage sont visualisables en violet (I) tandis que les points d'interaction sont coloriés suivant un code couleur propre au type d'interaction et à son appartenance : référence/comparaison (J).

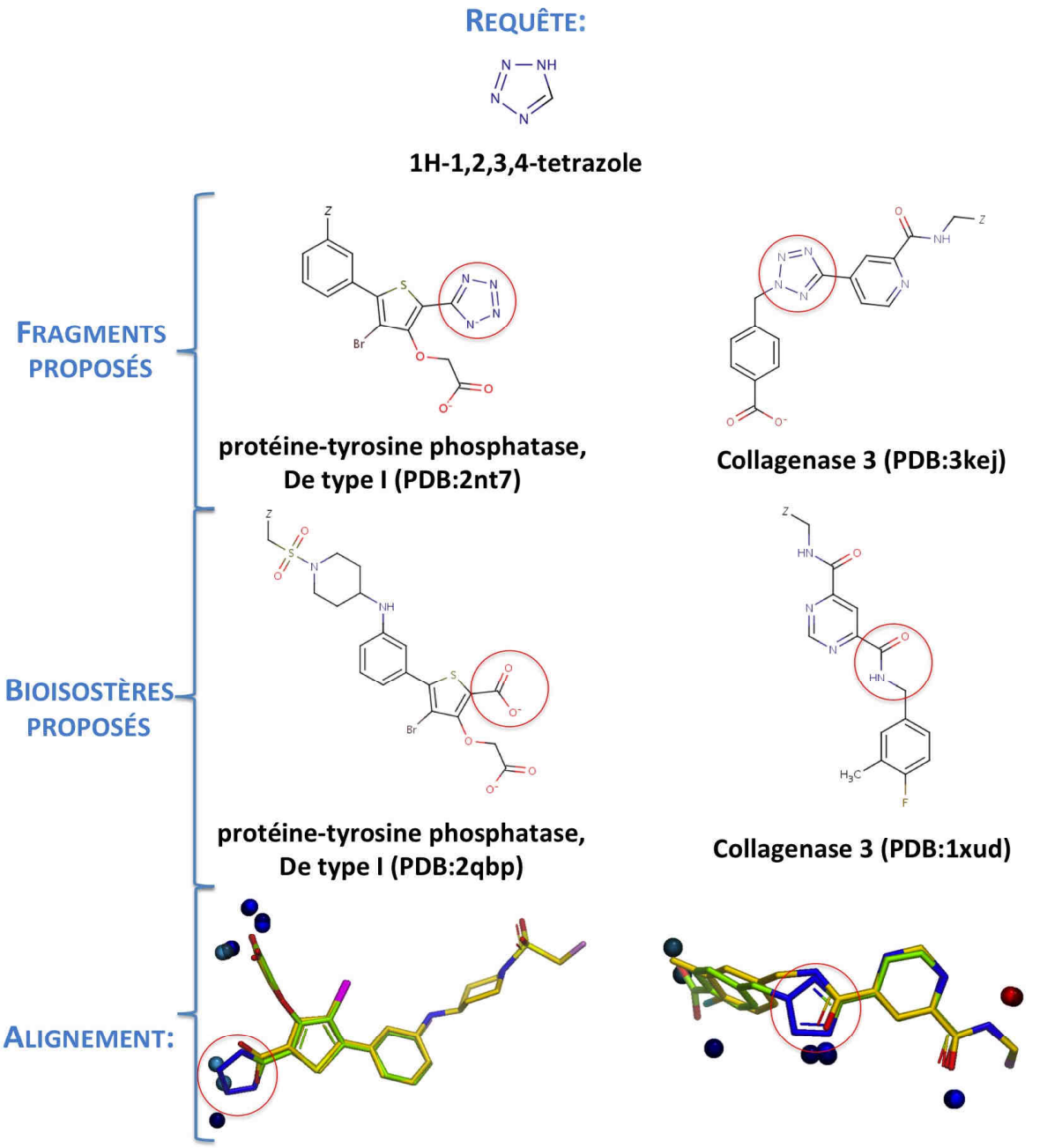
Les valeurs par défaut de chaque méthode sont récapitulées dans le **Tableau 2**. Les valeurs de similarité d'ECFP4 sont relativement faibles afin d'obtenir des fragments structuralement différents, tandis que les valeurs de similarité des empreintes TIFP, TIFPPol et de Grim (GrScore) sont relativement élevées afin de trouver des modes d'interaction similaires. Une nouvelle liste de fragments est alors proposée, ordonnée par GrScore décroissant. Pour chaque fragment de comparaison, les scores de similarité de TIFP, TIFPPol, ECFP4 et GrScore par rapport au fragment de référence sont affichés. De plus, la masse moléculaire, la structure et le code PDB associés aux fragments sont fournis. L'alignement des modes d'interaction des deux fragments sélectionnés est appliqué aux fragments eux-mêmes afin d'être visualisables facilement à travers l'interface graphique (**Figure 8.J**)

Tableau 2 - Récapitulatif des bornes par défaut pour les valeurs de similarité de chaque méthode utilisée dans scPDB-Frag. Les seuils de similarité des empreintes TIFP et GrScore correspondent à la valeur de similarité permettant d'obtenir la meilleure classification possible.

Méthode	TIFP	TIFPPol	GrScore	ECFP4
Valeur minimale	0,4	0,4	0,594	0
Valeur maximale	1	1	1,5	0,4
Seuil de similarité	0.318	/	0,594	0,4

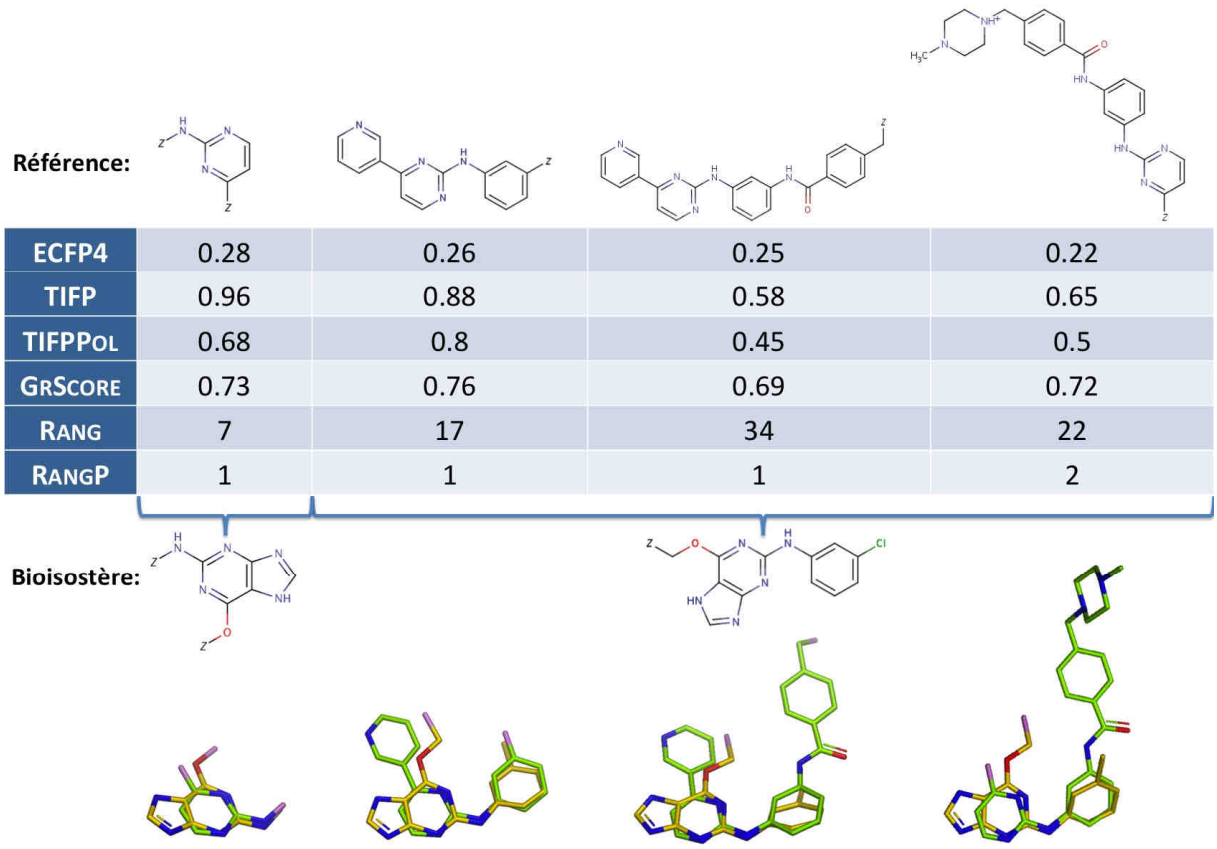
3.3.2 Recherche exhaustive

Une autre possibilité d'utilisation de la base de données scPDB-Frag est de réaliser une recherche exhaustive des remplacements possibles d'un fragment particulier, tel que le fragment 1H-1,2,3,4-tetrazole par exemple (**Figure 9**). 28 fragments de références sont proposés par l'interface web, dont la plupart ciblent les ß-lactamases, les collagenases, les thrombines, prothrombines ou les facteurs de coagulation X. Ce choix laissé à l'utilisateur permet de diversifier le nombre de bioisostères possibles. Ainsi, lorsque l'on choisit comme référence un fragment de la protéine-tyrosine phosphatase (Code PDB : 2nt7), on trouve comme meilleur bioisostère un acide carboxylique parfaitement aligné (Code PDB : 2qbp). Dans ce cas précis, il s'avère que les molécules originelles ne diffèrent que par cette modification, permettant de passer d'une constante d'inhibition de 300 nM avec le groupement tetrazole à une valeur de 4 nM avec le groupement carboxylique. Mais lorsque l'on cible un fragment



TIFP	0,6	0,6
TIFPPol	0,42	0,49
ECFP4	0,39	0,34
GrScore	0,86	0,89

Figure 9 - Exemple de résultats multiples pour une même requête, ici le 1H-1,2,3,4-tetrazole. L'utilisateur a le choix de la référence parmi 28 fragments proposés, ici des fragments de molécules ciblant la protéine-tyrosine phosphatase de type I et la collagénase 3. Parmi les bioisostères proposés en rang 1, leur protéine cible est identique à la protéine cible des référence et l'alignement fourni réalise une superposition correcte des fragments. Seuls les points d'interaction polaires « Centered » sont affichés ici afin d'améliorer la visibilité. Les sous-structures entourées en rouge représentent la requête ainsi que son potentiel bioisostère.



ECFP4	0.28	0.26	0.25	0.22
TIFP	0.96	0.88	0.58	0.65
TIFPPol	0.68	0.8	0.45	0.5
GrScore	0.73	0.76	0.69	0.72
Rang	7	17	34	22
RangP	1	1	1	2

Figure 10 - Exemple de l'incidence de la sélection du fragment sur les scores de similarité et l'alignement : cas de la Proto-oncogène tyrosine-protéine kinase Src (code PDB : 1y57). La structure des 4 fragments de référence est affichée, ainsi que leur score de similarité d'ECFP4, d'empreintes TIFP et TIFPPol, et GrScore par rapport au fragment d'un ligand en complexe avec une kinase cycline-dépendante 2 (cdk2, code PDB : 1h1r). Le rang correspond à la position du fragment bioisostère dans la liste des résultats, tandis que rangP correspond au rang par rapport aux fragments de la protéine kinase cycline-dépendante 2.

issu d'un complexe de la collagenase 3 (Code PDB : 3kej), le bioisostère obtenu s'avère ici être un groupement amide (Code PDB : 1xud). Ici la différence structurale des deux ligands est trop importante pour inférer une quelconque relation structure/activité. Cependant, il est bon de noter que dans le cas du fragment de référence, la constante d'inhibition est de 75 nM contre 8 nM pour le ligand de comparaison. Il faut cependant faire attention à cette relation entre activité et bioisostères. En effet, les valeurs de similarité sont ici symétriques, ce qui implique que l'obtention d'un gain en terme de constante d'inhibition peut aussi être une perte dans le sens inverse. De plus, ces considérations ne se limitent pas aux seules structures et ne doivent pas être généralisées. Nous avons ainsi pu mettre en évidence la possibilité de rechercher de nouveaux bioisostères à partir d'une simple requête.

3.3.3 INDEPENDANCE DE LA MASSE DU FRAGMENT

L'une des questions que l'on peut se poser à la vue de la distribution de la masse des fragments est la dépendance de ces derniers tant sur la sélection des fragments que sur l'alignement fourni. Pour répondre à cette question, reprenons l'exemple de la Proto-oncogène tyrosine-protéine kinase Src en tant que référence et de son potentiel bioisostère issu d'un ligand en complexe avec une kinase cycline-dépendante 2 (**Figure 10**). La référence et la comparaison possèdent respectivement 4 et 2 fragments différents. En réalisant la recherche pour chacun des fragments de référence avec les paramètres par défaut, au moins un fragment du ligand est retrouvé. De plus, les alignements fournis pour les 4 fragments de référence sont sensiblement identiques puisque les atomes d'azote réalisant une liaison hydrogène avec la partie hinge de la kinase sont correctement superposés. D'un point de vue des scores de similarité, les valeurs de similarité d'empreintes ECFP4 tendent à diminuer lorsque la masse de la référence augmente, ce qui est parfaitement normal. Le score de Grim (GrScore), varie légèrement mais reste au dessus de seuil de précision à 100% (0.659). Les valeurs de similarité des empreintes TIFP et TIFPPol sont plus difficiles à interpréter car plus sensibles aux effets de bord ainsi qu'au nombre de points d'interactions dans les modes d'interaction des fragments de référence et de comparaison. Cependant, leurs valeurs restent toutes au-delà des seuils de similarité définit dans le **Tableau 2** et les alignements restent sensiblement équivalent. Il est important de souligner que bien que le rang des fragments semble augmenter en fonction de la masse, le rang au sein de la protéine, ici cycline-dépendante 2, reste entre 1 et 2.

Le problème inhérent à la méthode d'alignement de Grim est le nombre de points d'interactions dans la référence et la comparaison. En effet, pour réaliser un alignement en trois dimensions, il est nécessaire de posséder au moins trois paires de points d'interactions non colinéaires. Cependant, ce type d'alignement est généralement faux ou mal ajusté, en particulier lorsque les points représentent les mêmes interactions. Plus leur nombre est important, plus le nombre d'appariements possibles est grand et par conséquent, plus fin sera l'alignement. Dans ces conditions, il serait préférable d'utiliser lorsque cela est possible les fragments possédant le plus d'atomes pour la paire de fragments observée. Malgré tout, cette hypothèse reste encore à confirmer sur des cas plus génériques que ceux obtenus à travers une analyse visuelle des plus basiques.

3.4 PERSPECTIVES

Les résultats présentés ici sont le fruit d'une analyse préliminaire. L'une des premières perspectives est de réaliser la même procédure en utilisant le protocole de fragmentation RECAP.[31] La fragmentation des molécules est réalisée en clivant certaines liaisons caractéristiques de réactions organiques connues. Une telle procédure engrangerait des fragments de plus faible masse et permettrait ainsi de les comparer avec notre méthode de fragmentation, tant d'un point de vue structural que bioisostérique. De plus, une analyse plus poussée de la distribution des scores de similarité de modes d'interaction au sein de fragments similaires ou de fragments en complexe avec une même famille de protéines serait intéressante à approfondir pour affiner les choix proposés à des vrais bioisostères.

Lorsque l'utilisateur choisit un fragment comme référence dans l'interface web, celui-ci est obligatoirement associé à un ligand d'une entrée dans la base de données. Cependant, il n'est pas rare de voir de multiples occurrences du même fragment au sein de la base pour lesquelles les protéines sont différentes et par conséquent les modes d'interaction sont différents. Il serait utile d'ajouter en option la possibilité de réaliser une recherche sur toutes les occurrences d'un fragment et non sur un seul afin d'améliorer les résultats. Cependant, cette option nécessite des temps de calculs plus importants et reste encore à être étudiée.

Un autre point intéressant serait d'ajouter une option permettant de sélectionner des fragments potentiellement bioisostériques ciblant soit la même famille de protéine, soit issus des protéines différentes, soit une famille précise. Comme on a pu l'observer dans la **Figure 10**, il est possible de trouver des bioisostères issus de protéine différentes.

L'une des problématiques actuelles concerne les points d'ancrages. En effet, lors de l'alignement de deux fragments, ces points peuvent ne pas être superposés et rendre ainsi plus difficile un remplacement d'un fragment par un autre. Une alternative profitable serait, une fois l'alignement des fragments réalisé, de reconstruire leur ligand respectif afin de trouver les atomes hors fragment qui se superposent le mieux afin d'offrir un remplacement viable. Cette reconstruction doit alors être réalisée en prenant la géométrie des atomes, au moyen de vecteurs par exemple, et de maximiser la superposition de ces vecteurs.

4. CONCLUSION

Nous avons présenté ici une nouvelle méthode de recherche de fragments dérivés de complexes cristallographiques protéine/ligand afin de découvrir de potentiels bioisostères. Pour cela, nous nous sommes basés sur une représentation des fragments à travers les interactions qu'ils effectuent avec leur protéine cible. Contrairement à d'autres méthodes ne se focalisant que sur la conservation des propriétés physico-chimiques clés des fragments pour trouver des bioisostères, cette nouvelle méthode permet de comparer des modes d'interactions de façon totalement indépendante de la structure initiale. Ainsi, la comparaison et l'alignement de ces modes d'interactions autorise la visualisation des fragments superposés et offrent un moyen simple et efficace de juger la qualité de potentiels bioisostères. Nous avons pu observer au travers de deux cas les possibilités que propose la base de données scPDB-Frag. La mise en place d'une interface web permet une interrogation facile de la base, et permet une visualisation ainsi qu'un export des fragments alignés. Cependant, ce projet nécessite une analyse plus poussée des résultats afin de mieux définir les critères de similarité, en particulier en fonction de la masse des fragments.

5. BIBLIOGRAPHIE

(1) Friedman, H. L. *Washington DC Natl Acad. Sci* **1951**, *206*, 295.

(2) Bemis, G. W.; Murcko, M. a *J. Med. Chem.* **1996**, *39*, 2887–93.

(3) Bemis, G. W.; Murcko, M. a *J. Med. Chem.* **1999**, *42*, 5095–5099.

(4) Birchall, K.; Gillet, V. J.; Willett, P.; Ducrot, P.; Luttmann, C. *J. Chem. Inf. Model.* **2009**, *49*, 1330–46.

(5) Ujváry, I. *Pestic. Sci.* **1997**, *51*, 92–95.

(6) Wagener, M.; Lommerse, J. P. M. *J. Chem. Inf. Model.* **2006**, *46*, 677–85.

(7) Schuffenhauer, A.; Ertl, P.; Roggo, S.; Wetzel, S.; Koch, M. a; Waldmann, H. *J. Chem. Inf. Model.* **2007**, *47*, 47–58.

(8) Holliday, J. D.; Jelfs, S. P.; Willett, P.; Gedeck, P. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 406–11.

(9) Ertl, P. *J. Mol. Graph. Model.* **1998**, *16*, 11–3, 36.

(10) Ertl, P. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 374–80.

(11) Ertl, P. *Bioorg. Med. Chem.* **2012**, *20*, 5436–42.

(12) Ertl, P.; Lewis, R. *J. Comput. Aided Mol. Des.***2012**, *26*, 1207–15.

(13) Griffen, E.; Leach, A. G.; Robb, G. R.; Warner, D. J. *J. Med. Chem.* **2011**, *54*, 7739–50.

(14) Wirth, M.; Zoete, V.; Michielin, O.; Sauer, W. H. B. *Nucl. Acids Res.* **2013**, *41*, D1137–43.

(15) Weber, J.; Achenbach, J.; Moser, D.; Proschak, E. *J. Med. Chem.* **2013**, *56*, 5203–7.

(16) Wood, D. J.; de Vlieg, J.; Wagener, M.; Ritschel, T. *J. Chem. Inf. Model.* **2012**, *52*, 2031–43.

(17) Kennewell, E. a; Willett, P.; Ducrot, P.; Luttmann, C. *J. Comput. Aided Mol. Des.***2006**, *20*, 385–94.

(18) Meslamani, J.; Rognan, D.; Kellenberger, E. *Bioinformatics* **2011**, *27*, 1324–6.

(19) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucl. Acids Res.* **2000**, *28*, 235–42.

(20) Weininger, D. *J. Chem. Inf. Model.* **1988**, *28*, 31–36.

(21) Tripos SybylX **2011**.

(22) Software, O. S. Filter.

(23) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. *Proteins* **2003**, *52*, 609–23.

(24) Brenk, R.; Schipani, A.; James, D.; Krasowski, A.; Gilbert, I. H.; Frearson, J.; Wyatt, P. G. *ChemMedChem* **2008**, *3*, 435–44.

(25) Hanser, T.; Jauffret, P.; Kaufmann, G. *J. Chem. Inf. Model.* **1996**, *36*, 1146–1152.

(26) Desaphy, J.; Raimbaud, E.; Ducrot, P.; Rognan, D. *J. Chem. Inf. Model.* **2013**, *53*, 623–637.

(27) Molecular Networks GmbH Corina.

(28) Rogers, D.; Hahn, M. *J. Chem. Inf. Model.* **2010**, *50*, 742–54.

(29) Accelrys Software Inc PipelinePilot **2012**.

(30) Willett, P.; Barnard, J. M.; Downs, G. M. *J. Chem. Inf. Model.* **1998**, *38*, 983–996.

(31) Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. *J. Chem. Inf. Model.* **1998**, *38*, 511–522.

Conclusion

Ce travail de thèse a porté sur l'analyse des interactions protéine/ligand au sein de structures cristallographiques issues de la Protein Data Bank. Pour cela, différents outils et points de vue ont été mis en place afin de mieux comprendre le concept de reconnaissance moléculaire. L'avantage d'utiliser l'ensemble des structures protéine/ligand est de pouvoir inférer une évaluation statistique de grande ampleur, tout en étant capable de regarder des effets plus locaux, tel qu'observable au sein d'une famille de protéines ou d'un même ligand.

Initialement, cette thèse a débuté par le développement de VolSite, une méthodologie permettant de décrire les cavités de sites de liaison des protéines. La représentation simple de ces cavités est réalisée au travers d'une discrétisation de celles-ci au moyen d'un ensemble de points caractéristiques encodant la forme mais aussi la propriété physico-chimique complémentaire aux propriétés des atomes de protéine les plus proches. Ainsi nous souhaitions assigner pour toutes les positions de l'espace la propriété physico-chimique attendue pour un atome de ligand. Cette description est très semblable aux pharmacophores mais diffère dans le sens où l'on ne se limite pas aux seules interactions des ligands connus pour une cible. Son indépendance aux coordonnées initiales, et sa faible variabilité des résultats en font une méthode de choix pour de nombreux cas d'études. L'un des objectifs premiers était d'être capable de retrouver la fonction de la protéine observée en comparant sa cavité à l'ensemble des cavités de la base de données sc-PDB. La capacité de VolSite/Shaper à mesurer la similarité et à discriminer des cavités de paires de protéines similaires et de paires de protéines différentes s'est avérée très efficace. De là, nous sommes allés plus loin en tentant, et en réussissant, à prédire la droguabilité d'un site de liaison. A terme, cette prédiction nous permettra de se focaliser sur des sites de liaison plus propices à accueillir une potentielle molécule active. Cette méthodologie reste encore jeune et nécessite certaines améliorations pour pouvoir atteindre son plein potentiel. Ainsi, l'une des modifications en cours d'études est de modifier VolSite pour le transformer en un logiciel de détection de cavités protéiques que l'on ordonnerait par valeur de droguabilité décroissante.

Dans un second temps, nous nous sommes concentrés sur les modes d'interaction entre une protéine et un ligand. En caractérisant chaque interaction par un ensemble de points nominatifs, nous avons réduit l'information d'un complexe protéine/ligand aux seules interactions. Etonnamment, cette représentation suffit amplement pour pouvoir comparer et aligner deux complexes de façon précise et efficace. De plus, la prise en compte des seules interactions covalentes clés a suffi pour caractériser, décrire, comparer et utiliser ces modes d'interactions dans divers cas de figure. Ainsi, cette méthode est tout aussi capable d'inférer la fonction d'une protéine que VolSite, à ceci près qu'elle nécessite beaucoup moins de points pour décrire un complexe. Par ailleurs, il s'avère que la comparaison des modes d'interactions est un outil potentiellement très avantageux dans le cas de criblage virtuel afin de trouver des molécules actives pour une cible. Bien que cela nécessite de posséder un certain nombre de molécules actives, elle reste une méthode viable, efficace et rapide pour réduire l'attrition, c'est à dire le nombre de molécules expérimentalement inactives. L'une des surprises de cette thèse concerne les matrices « tout-contre-tout ». Nous étions au tout début très sceptique face à la quantité de données à analyser. Cependant, nous n'avions pas imaginé trouver une corrélation si nette entre la similarité de paires de sites de liaison et les paires de modes d'interaction. Cette hypothèse jusqu'alors supposée que les modes d'interactions sont dépendants de la protéine et non du ligand a été prouvée au travers de l'analyse de près de 300 millions de mesure de similarité. De la même façon que pour VolSite, de nombreuses améliorations sont possibles telles que la prise en compte des interactions non covalentes plus occasionnelles, l'affinement du post-traitement dans le cadre de l'arrimage moléculaire ou encore l'appariement possible de différents types d'interactions.

Dans un dernier temps, nous nous sommes intéressés aux interactions protéine/ligand, mais cette fois-ci à un niveau beaucoup plus local. Ainsi, nous souhaitions observer des conversions locales de modes d'interactions et regarder la structure des fragments de ligands associés. Pour cela, tous les ligands de la base de données sc-PDB ont été fragmentés, puis chaque fragment est associé à son mode d'interaction. La comparaison de ses modes d'interactions d'une façon relativement indépendante de la structure initiale des fragments permet de rechercher des

bioisostères, c'est à dire des sous-structures de ligands structuralement différents mais réalisant le même mode d'interaction. Nous avons ainsi montré au travers de divers exemples la capacité de cette nouvelle méthodologie. De plus, une interface web permet d'interroger cette base de données, offrant ainsi aux chimistes médicinaux la possibilité de trouver de nouveaux fragments de molécules dans le cadre de l'optimisation de molécules actives. Bien sur, cette technique reste encore à être étudiée sur plus de cas que les quelques exemples présentés ici.

Au delà de ces aspects méthodologiques et techniques existe un aspect humain qu'il est important de noter. Ces trois années passées au laboratoire ont été à mes yeux une magnifique expérience, tant professionnelle qu'humaine. En effet, j'ai eu l'occasion de travailler avec une excellente équipe, toujours souriante et passionnée qui m'a beaucoup appris et avec laquelle j'ai pu évoluer.

Jérémy DESAPHY

L'analyse structurale de complexes protéine/ligand et ses applications en chémogénomique

Résumé

Comprendre les interactions réalisées entre un candidat médicament et sa protéine cible est un enjeu crucial pour orienter la recherche de nouvelles molécules. En effet, ce processus implique de nombreux paramètres qu'il est nécessaire d'analyser séparément pour mieux comprendre leurs effets.

Nous proposons ici deux nouvelles approches observant les relations protéine/ligand. La première se concentre sur la comparaison de cavités formées par les sites de liaison pouvant accueillir une molécule. Cette méthode permet d'inférer la fonction d'une protéine mais surtout de prédire « l'accessibilité » d'un site de liaison pour un médicament. La seconde tactique se focalise sur la comparaison des interactions non-covalentes réalisées entre la protéine et le ligand afin d'améliorer la sélection de molécules potentiellement actives lors de criblages virtuels, et de rechercher de nouveaux fragments moléculaires, structuralement différents mais partageant le même mode d'interaction.

Mots Clés : chémoinformatique, chémogénomique, Bioinformatique, Profilage, Criblage virtuel, Site de liaison, Interactions protéine/ligand, Bioisostères, Interactions non-covalentes.

Résumé en anglais

Understanding the interactions between a drug and its target protein is crucial in order to guide drug discovery. Indeed, this process involves many parameters that need to be analyzed separately to better understand their effects.

We propose two new approaches to observe protein/ligand relationships. The first focuses on the comparison of cavities formed by binding sites that can accommodate a small molecule. This method allows to infer the function of a protein but also to predict the accessibility of a binding site for a drug. The second method focuses on the comparison of non-covalent interactions made between the protein and the ligand to improve the selection of potentially active molecules in virtual screening, and to find new molecular fragments, structurally different but sharing the same mode of interaction.

Keywords : chemoinformatics, chemogenomics, bioinformatics, virtual screening, binding sites, binding modes, bioisosteres, non-covalent interactions.