Nancy-Université
Université Henri Poincaré

Université Henri Poincaré – Nancy I
U.F.R. Sciences et techniques de la matière et des procédés
École doctorale lorraine de chimie et physique moléculaires

Thèse

présentée pour l'obtention du titre de

Docteur de l'Université Henri Poincaré

en Chimie Informatique et Théorique

par **Alexandre Beautrait**

DÉVELOPPEMENT ET VALIDATION DE LA PLATEFORME DE CRIBLAGE VIRTUEL VSM-G ET ÉTUDE DU DOMAINE FAT DE LA KINASE D'ADHÉRENCE FOCALE FAK

Soutenue à huis clos le 15 janvier 2008

Membres du Jury :

Rapporteurs :	M. Nicolas Moitessier	Professeur Assistant, CR CNRS, Mc Gill University, Montréal, Canada
	M. Luc Morin-Allory	Professeur, Université d'Orléans, Orléans
Examinateurs :	M. Daniel Canet	Professeur, Université H. Poincaré, Nancy
	Mme Marie-Dominique Devignes	CR CNRS, LORIA, Nancy
	M. Nicolas Floquet	CR CNRS, Faculté de Pharmacie, Montpellier
	M. Bernard Maigret	DR CNRS, LORIA, Nancy (Directeur de thèse)
	M. Michel Souchet	Docteur, Fournier Pharma / Solvay, Daix

Équipe ORPAILLEUR, Laboratoire Lorrain de recherche en Informatique et ses Applications (LORIA)
UMR 7503 - Campus Scientifique - BP 239 - 54506 Vandœuvre-lès-Nancy Cedex

Nancy-Université
Université Henri Poincaré

Université Henri Poincaré – Nancy I
U.F.R. Sciences et techniques de la matière et des procédés
École doctorale lorraine de chimie et physique moléculaires

Thèse

présentée pour l'obtention du titre de

Docteur de l'Université Henri Poincaré

en Chimie Informatique et Théorique

par **Alexandre Beautrait**

DÉVELOPPEMENT ET VALIDATION DE LA PLATEFORME DE CRIBLAGE VIRTUEL VSM-G ET ÉTUDE DU DOMAINE FAT DE LA KINASE D'ADHÉRENCE FOCALE FAK

Soutenue à huis clos le 15 janvier 2008

Membres du Jury :

Rapporteurs :	M. Nicolas Moitessier	Professeur Assistant, CR CNRS, Mc Gill University, Montréal, Canada
	M. Luc Morin-Allory	Professeur, Université d'Orléans, Orléans
Examinateurs :	M. Daniel Canet	Professeur, Université H. Poincaré, Nancy
	Mme Marie-Dominique Devignes	CR CNRS, LORIA, Nancy
	M. Nicolas Floquet	CR CNRS, Faculté de Pharmacie, Montpellier
	M. Bernard Maigret	DR CNRS, LORIA, Nancy (Directeur de thèse)
	M. Michel Souchet	Docteur, Fournier Pharma / Solvay, Daix

Équipe ORPAILLEUR, Laboratoire Lorrain de recherche en Informatique et ses Applications (LORIA)
UMR 7503 - Campus Scientifique - BP 239 - 54506 Vandœuvre-lès-Nancy Cedex

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DES SIGLES ET ABRÉVIATIONS

ATP	Adénosine triphosphate
CCK9	Cholécystokinine 9
ERK	Extracellular signal-Regulated Kinases (kinases régulées par des signaux extracellulaires)
FAK	Focal Adhesion Kinase
FAT	Focal Adhesion Targeting domain
FERM	erythrocyte band. Four.1-Ezrin-Radixin-Moesin
Fyn	Tyrosine kinase Fyn
GFDL	GNU Free Documentation License
Graf	GTPase regulator associated with FAK
Grb2,Grb7	Growth factor receptor-bound protein 2, 7
GSK3	Glycogen Synthase Kinase 3
JNK	Jun N-terminal Kinase
LXRβ	Liver X Receptor beta
MAPK	Mitogen Activated Protein Kinase
MEK	Mitogen ERK Kinase
MM/PBSA	Molecular Mechanics / Poisson-Boltzmann Surface Area
MMPs	Matrix metalloproteinases
P130Cas	Protéine de 130 kDa associée à Crk (Crk-associated substrate)
PDB	Protein Data Bank
PI3K	Phosphatidylinositol-3-kinase
PKC	Protein Kinase C
PLCγ	Phospholipase Cγ
PP1	Protein Phosphatase 1
Pyk2	Proline-rich tyrosine kinase 2
Ras	Rat sarcoma virus
Rho	Ras homology protein
RMN	Résonance Magnétique Nucléaire
RMSD	Root Mean Square Deviation (en français : écart quadratique moyen)
RX	Diffraction des Rayons-X
SH2,SH3	Domaine d'homologie à Src de type 2 ou 3
Sos	Son of sevenless protein
Src	Tyrosine kinase Src
VEGF	Vascular Endothelial Growth Factor
VEGFR	Vascular Endothelial Growth Factor Receptor

AVANT-PROPOS

La recherche de nouveaux médicaments s'inscrit dans le cadre plus général des sciences du Vivant. On entend par là un ensemble de disciplines scientifiques issues des progrès réalisés au XXe siècle autour de la biologie et dont les applications fondent désormais une très grande part des découvertes en pharmacologie et en médecine. Ces dernières années, les avancées conceptuelles initiées par la recherche en génétique ont considérablement élargi l'étendue de ces connaissances. Nous sommes ainsi passés d'une vision centrée sur le génome à celle reposant sur l'ensemble des protéines (le protéome), pour enfin considérer l'ensemble des interactions entre biomolécules (l'interactome). Protéome et interactome sont les concepts de représentation des organismes vivants dans lesquels ce travail se positionne.

On peut donc déjà constater que si ce travail repose fondamentalement sur l'étude de biomolécules par des modèles théoriques, il ne saurait s'y limiter. De plus, si la modélisation moléculaire de systèmes biologiques est d'abord fortement liée à certaines techniques expérimentales, elle progresse surtout au travers des méthodes modernes de calcul informatique. Et, de façon analogue à l'évolution des concepts sur le Vivant, les concepts de calcul numérique se globalisent, de l'ordinateur à la grappe, de la grappe au réseau formant une grille.

La variété des visions comme des approches, des outils comme des connaissances, forme ainsi pour les sciences du Vivant un large spectre au milieu duquel nous nous situons et dont le dynamisme est avant tout source de motivation mais auquel il est nécessaire de s'adapter. Ainsi, l'étude de systèmes biologiques n'est pas une science exclusive pouvant être pratiquée de façon autonome mais, au contraire, une discipline ouverte dont les plus belles perspectives reposent sur les interactions interdisciplinaires. Cette caractéristique est particulièrement marquée au niveau de la recherche pharmaceutique.

Dans ce contexte, ma contribution associe étroitement théorie et application pratique et, pour l'une comme pour l'autre, profite de multiples collaborations. Ce travail est ainsi constitué de deux parties distinctes, l'une centrée principalement sur le développement d'une plateforme logicielle, et l'autre intégrée concrètement à des actions de recherche pharmaceutique. Je souhaite que le lecteur de ce manuscrit sache percevoir le lien permanent établi entre ces deux aspects de mes travaux de recherche. J'espère aussi qu'il pourra ressentir l'envie qui les a alimentés et qui a justifié mes efforts.

REMERCIEMENTS

Le travail présenté ici a été effectué au sein de l'Université Henri Poincaré (équipe eDAM), puis au LORIA, sous la responsabilité du docteur Bernard Maigret à qui je souhaite exprimer en premier lieu mes remerciements les plus chaleureux. Je lui suis reconnaissant pour la liberté d'action qu'il m'a accordée, ainsi que pour les moyens qu'il a su mettre à ma disposition pour la réalisation de ces travaux.

Mes remerciements s'adressent aussi aux professeurs Nicolas Moitessier et Luc Morin-Allory pour avoir accepté de prendre le temps de rédiger un rapport sur mon mémoire. Je remercie les autres membres du jury : le professeur Daniel Canet et les docteurs Marie-Dominique Devignes, Nicolas Floquet et Michel Souchet pour avoir également accepté de juger mes travaux.

Que les collaborateurs avec qui j'ai eu le plus grand plaisir à travailler : Mercedes, Michel, Sinan, Peter, Wensheng, Gilles et Jean-Antoine, soient assurés de ma gratitude.

Je voudrais remercier tous ceux qui ont contribué à l'élaboration de cette thèse de près ou de loin : mes acolytes Matthieu avec qui j'ai eu le plaisir de partager le bureau, Léo, Yesmine et Naziha, sans oublier les "vétérans" : Jean-Paul et Jérôme pour leurs conseils et petites astuces, et enfin toute personne ayant participé à la relecture de ce mémoire. Un merci particulier à Vincent qui a lu et relu ce manuscrit et qui m'a souvent prêté main forte durant cette thèse.

Je ne saurais oublier le soutien inestimable, même à distance, dont j'ai bénéficié de la part de ma famille et de mes amis durant cette expérience unique en son genre.

Je dédicace ce manuscrit à Christelle.

PARTIE 1 - INTRODUCTION

I - ÉTUDE THÉORIQUE DU VIVANT À L'ÉCHELLE MOLÉCULAIRE

I-1. Structure des biomolécules

I-1.1. Des acides nucléiques à l'information génétique

Le matériel héréditaire est codé par les acides nucléiques qui sont le support de l'information génétique de la cellule. Les acides nucléiques s'observent sous deux formes polymériques : l'acide désoxyribonucléique (ADN) et l'acide ribonucléique (ARN), formes dans lesquelles des groupes fonctionnels appelés nucléotides sont greffés sur un squelette polymère pentose-phosphate. Quatre nucléotides différents sont rencontrés dans la structure de l'ADN : adénine (A), thymine (T), cytosine (C) et guanine (G). La structure de l'ARN est très voisine de celle de l'ADN : les thymines sont mutées en uraciles (U) et les désoxyriboses du squelette sont remplacés par des riboses.

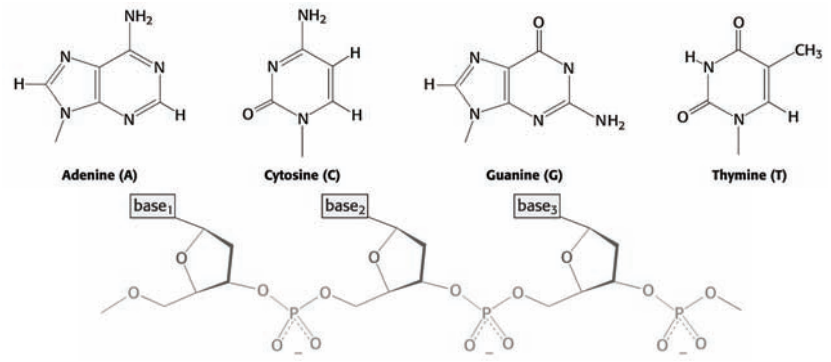


FIG. 1 - Le squelette pentose-phosphate de l'ADN et les 4 bases possibles (A, C, G, T) correspondantes.

L'ADN et l'ARN ont des fonctions différentes par leur structure. La structure de l'ADN fut déterminée par Watson, Crick et Franklin en 1953 [1, 2] et a pour particularité le positionnement des nucléotides à l'intérieur d'une double hélice (ou double brin) formée par deux chaînes pentose-phosphate antisymétriques. Ce positionnement dans un espace si confiné est rendu possible par un appariement spécifique des nucléotides : adénine et thymine d'une part, cytosine et guanine d'autre part, sont reliées d'un brin à l'autre par des liaisons hydrogène. L'information génétique est ainsi doublée et stabilisée, caractérisant le rôle biologique de conservation tenu par l'ADN.

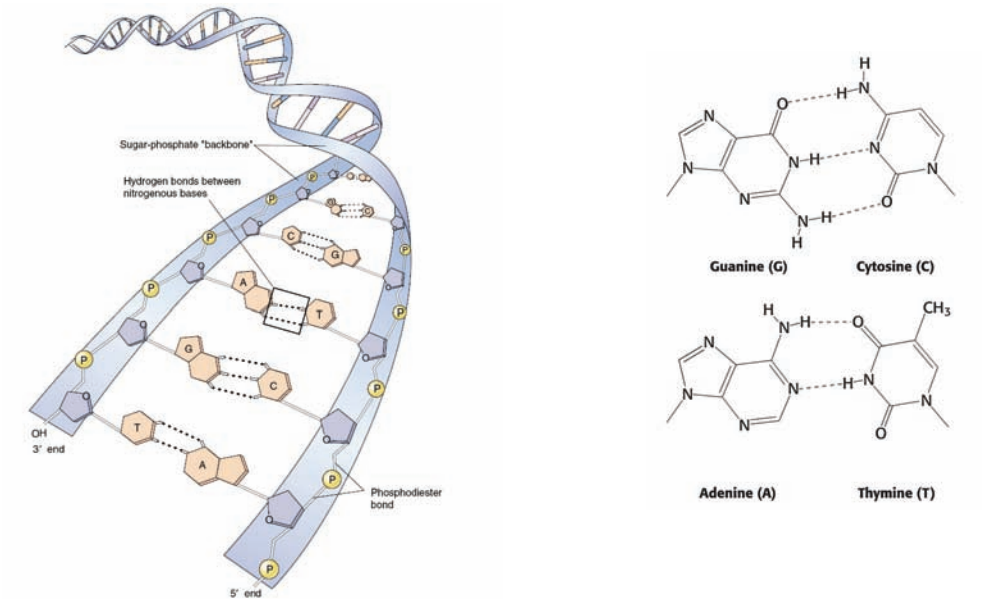
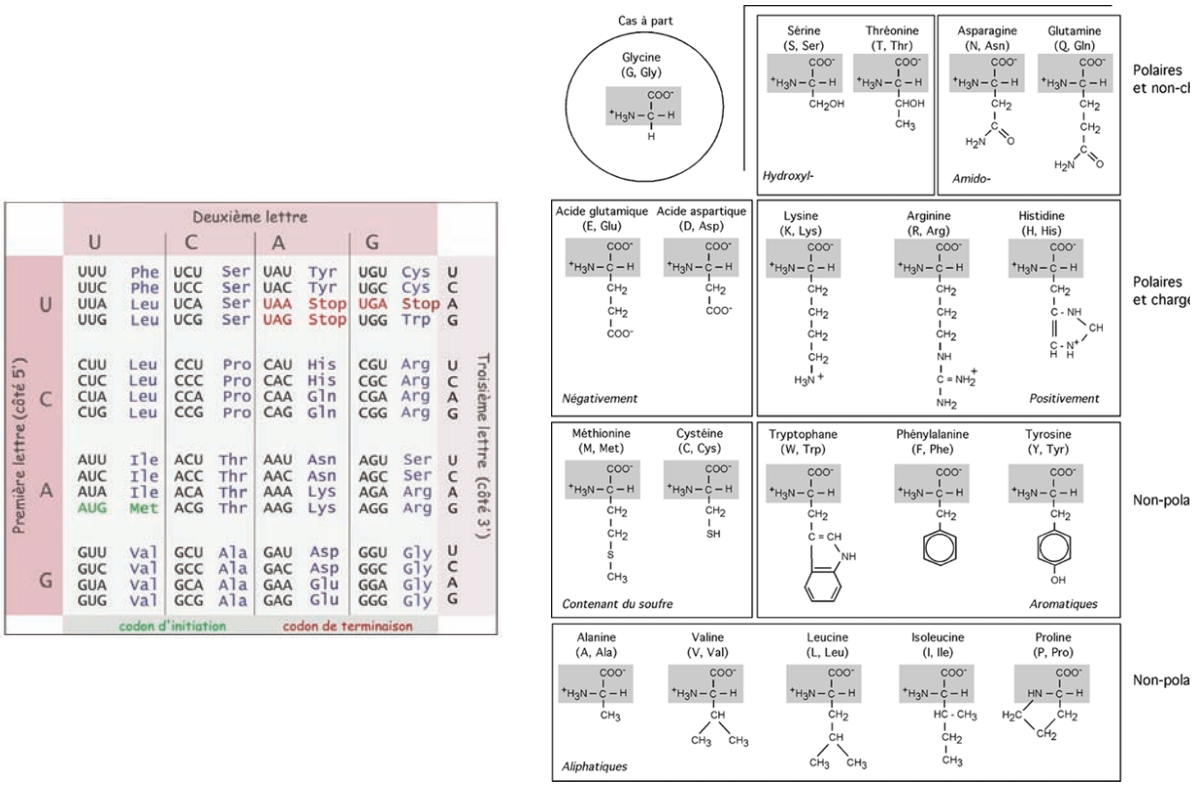


FIG. 2 - Structure de l'ADN en double hélice (d'après Raven and Johnson, Biology 6th Edition) et détails de l'appariement des nucléotides.

Dans le noyau de la cellule, le gène codant une protéine est transcrit de l'ADN en ARN. Celui-ci, dont la structure est en simple brin, est exporté du noyau vers le cytoplasme où son message sera ensuite déchiffré pour synthétiser une future protéine.

I-1.2. Du génome au protéome

Les protéines sont également des polymères, constitués à partir des vingt acides aminés naturels (Figure 3b). La signification du code génétique [3] permet de faire le lien entre les séquences nucléiques et peptidiques : les séquences codantes d'un gène codent les séquences protéiques [4]. Le code est incarné par la succession des nucléotides trois par trois, qui forment des triplets, aussi appelés codons (Figure 3a). Ce code contient un message dont la teneur est traduite par un système de lecture, le ribosome [5]. Pour s'exprimer, le message doit d'abord être copié sous la forme d'une molécule dite ARN messager (ARNm). La lecture des codons sur l'ARNm est alors assurée par le ribosome, qui utilise de petits ARN de transfert (ARNt) pour associer de manière spécifique les différents triplets aux acides aminés (Figure 4). A chaque codon de la séquence codante du gène correspond un et un seul des vingt acides aminés.



(a) (b)

FIG. 3 - (a) Le code génétique permettant la traduction de chaque codon en acide aminé.[*] (b) Les 20 acides aminés naturels, classés par propriétés.[†]

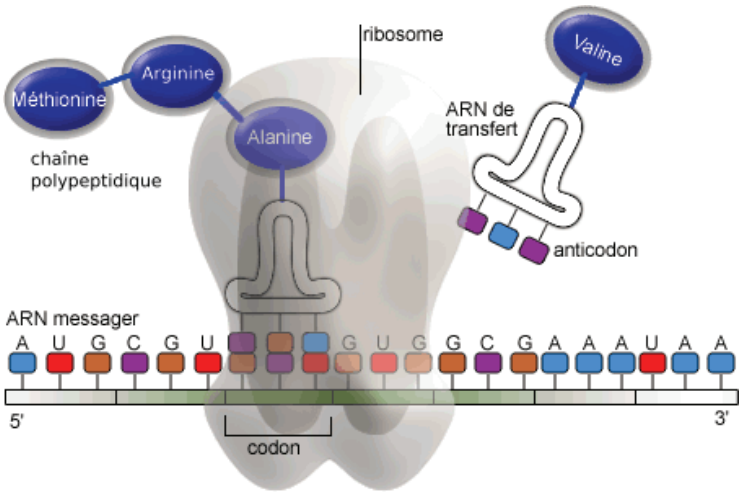


FIG. 4 - Traduction de l'ARN messager par le ribosome qui, par l'intermédiaire de l'ARN de transfert, assemble la séquence d'acides aminés formant la chaîne polypeptidique. Image sous licence GFDL.

[*] http://www.ulysse.u-bordeaux.fr/atelier/ikramer/biocell_diffusion/gbb.cel.fa.106.b3/content/access.htm

[†] http://pages.usherbrooke.ca/bcm-514-bl/1a.html

La protéine naissante fabriquée par le ribosome en sort sous la forme d'une chaîne linéaire d'acides aminés. Dans le milieu biologique où la protéine exerce son activité, cette chaîne se replie sur elle-même afin de minimiser son énergie interne. A l'issue de ce processus, une structure spécifique stable est atteinte, qui détermine la fonction biologique de la protéine[*]. Depuis les années 1970, le repliement des protéines a été le sujet de nombreuses études et certains principes de base de son mécanisme ont été dégagés et des résultats statistiques obtenus [7-9]. Le but ultime de ces recherches était de déchiffrer cet autre "code" qui aurait permis de prédire la structure tridimensionnelle d'une protéine à partir de sa séquence. Malgré tous les efforts consentis, le problème du repliement des protéines n'est toujours pas résolu à ce jour et constitue de ce fait un véritable défi de la biochimie [10]. La recherche dans ce domaine reste très active, comme en témoigne l'engouement pour le concours CASP[†] (*Critical Assessment of Structure Prediction*) qui a un réel effet stimulant dans l'amélioration des techniques bioinformatiques utilisées dans la prédiction du repliement des protéines [10, 11].

Alors que l'étude du génome correspond fondamentalement à l'exploration d'un espace séquentiel (sur 4 bases), celui du protéome, en plus de l'extension d'une partie de cet espace (cette fois codé sur 20 bases), lui ajoute un espace conformationnel[‡] beaucoup plus vaste [12]. À moins qu'une structure expérimentale[§] de la protéine soit disponible (ce qui fut le cas pour ce travail), il est difficile d'étudier théoriquement cet espace conformationnel (p. ex. en utilisant les techniques de la modélisation moléculaire). Ainsi, l'exploration du protéome reste aujourd'hui une tâche particulièrement délicate.

[*] Dans la plupart des cas, les protéines synthétisées atteignent leur repliement fonctionnel grâce à l'aide d'autres protéines chaperonnes (p. ex. le complexe protéique GroEL/GroES)[6].

[†] Site web du centre organisant les concours CASP : http://predictioncenter.org/

[‡] On passe alors de la structure primaire des protéines qui est linéaire (séquence) à la structure tertiaire (repliement de la chaîne polypeptidique dans l'espace : structure 3D). Pour plus de détails sur les 4 niveaux de structuration des protéines : voir en annexe.

[§] En grande majorité, les structures tridimensionnelles disponibles sont résolues par diffraction aux rayons X (RX) ou par résonance magnétique nucléaire (RMN). Pour plus d'information sur les principales techniques utilisées pour la détermination des structures des protéines : se référer à (*Liu et al., 2005*) [13].

I-2. Vue globale : du génome à l'interactome

Afin de mieux comprendre les liens entre les acteurs des processus biologiques, les données sur les interactions entre macromolécules au sein d'un organisme appelé *interactome*, sont identifiées. Une fois collectées, elles sont regroupées au sein de voies de signalisation et représentées schématiquement sous la forme de graphes d'interaction [14, 15].

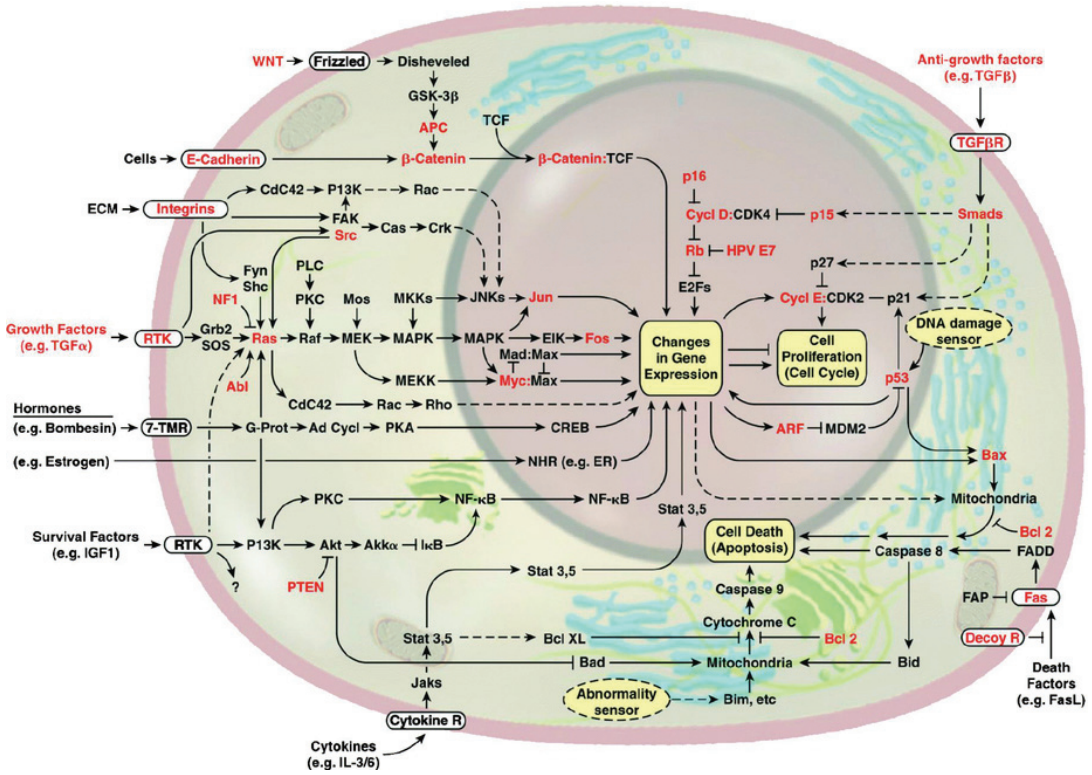


FIG. 5 - Exemple schématique de chemins de signalisation au sein de la cellule. Ici, sont représentées les principales voies connues de croissance et de régulation de la population de cellules d'un organisme (donc particulièrement impliquées dans le cas de cancers). D'après (Hanahan et al., 2006) [16].

L'étude sur le plan moléculaire du fonctionnement d'un organisme vivant peut ainsi s'effectuer à différents niveaux conceptuels successifs. Le *génome* repose sur l'espace des séquences de nucléotides ; le *protéome* y ajoute l'espace géométrique des protéines correspondantes et l'*interactome* la liste et la nature des interactions possibles qui en découlent.

Ces dernières années ont connu une accélération des découvertes des interactions entre les protéines dans divers organismes grâce aux recherches systématiques, à grande échelle, ayant recours à des techniques rapides et accessibles [17-19]. On peut citer en exemple l'identification complète de l'interactome de la levure *Saccharomyces cerevisae* [20]. Ce type d'étude devrait permettre une meilleure compréhension de l'interactome chez l'homme et, notamment, l'identification de nouvelles protéines impliquées dans le développement de pathologies.

Comme mentionné précédemment, pour modéliser les mécanismes impliqués dans les processus cellulaires et leurs dysfonctionnements, il peut s'avérer utile de replacer les biomolécules dans un contexte tridimensionnel. Ainsi, dans une optique thérapeutique de conception de médicaments, il est de grand intérêt de connaître des détails structuraux des complexes protéine-ligand ou l'interface entre macromolécules interagissant entre elles pour, par exemple, accentuer ou empêcher leur reconnaissance mutuelle. Dans ce contexte, les avancées techniques des méthodes de détermination structurale s'avèrent cruciales[*] [13, 21, 22]. À l'inverse, et en dépit des récents progrès pour l'enrichir, la cartographie de l'interactome humain n'en est qu'à ses débuts [23, 24].

I-3. L'ère de la post-génomique

Le génome humain est maintenant séquencé et annoté [25-28]. Un des grands espoirs lié à cet accomplissement est la découverte de nouvelles protéines à potentiel thérapeutique. Parmi les 20 000 à 25 000 gènes humains codant pour des protéines, il a été estimé que 3 000 d'entre eux codent pour des *cibles* thérapeutiques : des protéines qui, à la fois, sont liées à certaines pathologies ("*disease genes"*) et peuvent aussi être ciblées par des petites molécules ayant des propriétés caractérisant un médicament ("*druggable genome"*) [29, 30]. Le nombre de cibles thérapeutiques actuellement exploitées par l'industrie pharmaceutique ne représente qu'une partie mineure de cet espace pharmacologique [31]. En effet, une étude récente synthétisant toutes les précédentes estimations fixe le nombre des cibles visées par les médicaments du marché à 324 [32]. L'exploration de l'espace pharmacologique de toutes les cibles thérapeutiques potentielles n'est donc pas terminée et constitue un des objectifs de la recherche pharmaceutique dans l'ère post-génomique.

Un autre des bénéfices attendus du séquençage complet du génome humain sur le plan médical est de permettre d'identifier la source de nombreuses pathologies, tout en fournissant de précieuses

[*] Le nombre des protéines dont la structure tridimensionnelle est résolue et disponible croît exponentiellement, voir : http://www.rcsb.org/pdb/statistics/contentGrowthChart.do?content=total&seqid=100
Au 06 novembre 2007, la base PDB (*Protein Data Bank*) totalise 47 136 structures de biomolécules.

indications pour mettre en oeuvre des traitements pharmaceutiques individualisés. Ainsi, la discipline récente connue sous le nom de *pharmacogénomique* [33] (ou *pharmacogénétique*), permet d'établir un lien entre le polymorphisme de la structure génique (génotype de chaque patient) et la variabilité de la réponse à l'effet d'un médicament. Elle porte donc les espoirs d'une médecine préventive qui guide le traitement de chaque individu (choix de la molécule, posologie,…) et qui peut considérablement réduire la probabilité d'effets non désirés [34].

Mais il est à présent évident que l'étude du génome ne peut pas être la "recette miracle" qui révolutionnera la compréhension du vivant. En particulier, l'hypothèse réductionniste « un gène = une fonction biologique » s'est avérée fausse. Pour un grand nombre de gènes, la fonction biologique correspondante est inconnue et il n'est même pas possible de déterminer s'il y en a une. Pour ces raisons, on estime souvent que l'effort investi dans la génétique doit à présent être étendu en direction de l'interactome. En effet, si la génétique permet d'identifier de nombreuses pathologies, elle reste souvent impuissante dès lors qu'il s'agit de les corriger.

Si la thérapie génique [35] est un domaine de recherche connu, la mise au point de médicaments ciblant l'interactome est également un axe prometteur en particulier sur le plan pharmaceutique. Le travail de cette thèse se situe dans ce dernier domaine.

II - ÉTAPES DE MISE AU POINT D'UN MÉDICAMENT

L'effet thérapeutique principal d'un médicament est caractérisé au niveau des voies de signalisation, de régulation et de biosynthèse. Une telle activité peut se simplifier au niveau moléculaire en ne considérant que la liaison entre la molécule que constitue le principe actif du médicament et un site actif sur une macromolécule biologique bien définie. Ce site actif est la cible du médicament. La liaison correspondante implique une reconnaissance mutuelle des deux partenaires, c'est-à-dire une affinité de l'un pour l'autre, et modifie les propriétés de la cible moléculaire. Il en résulte une réponse biologique qui peut être de diverse nature (ex : sécrétoire, métabolique). Cette modification du fonctionnement de l'organisme doit répondre à la pathologie qui doit être traitée.

Le processus de développement d'un médicament s'organise en plusieurs étapes, suivant un protocole rigoureux : mise au point, essais pré-cliniques, essais cliniques (phase I à III), puis suivi post-commercialisation (phase IV).

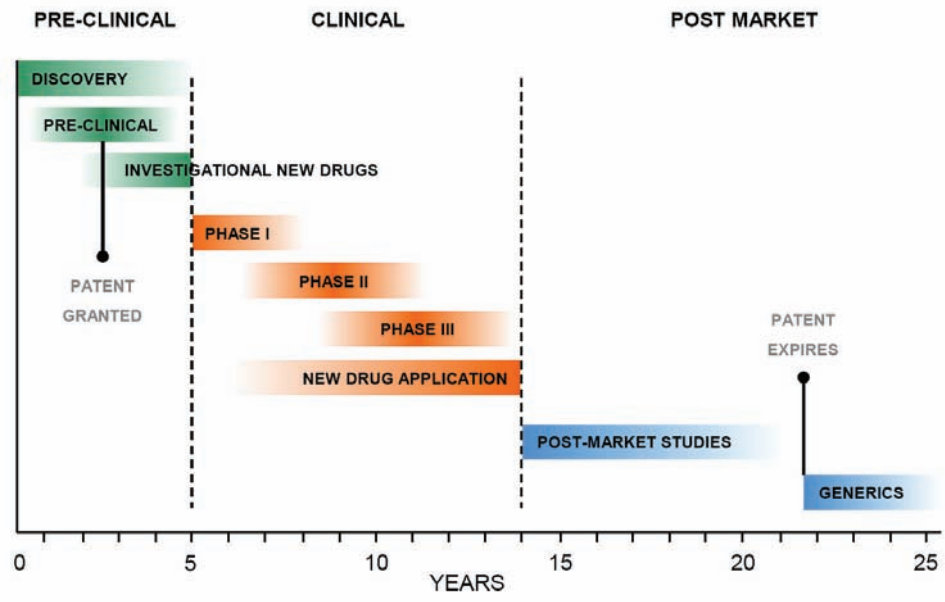


FIG. 6 - Les étapes du développement d'un médicament (reproduit de Wellcome Trust ©).

II-1. Identification et validation des cibles thérapeutiques

La notion de médicament est indissociable de celle de la cible sur laquelle il agit pour engendrer son action pharmacologique. La découverte de nouveaux médicaments nécessite donc de bien connaître la cible moléculaire correspondante, le fonctionnement de celle-ci et les voies de signalisation dans lesquelles elle est impliquée. La caractérisation précise d'une nouvelle cible thérapeutique est, par conséquent, la première étape nécessaire de la découverte de nouveaux médicaments [36]. Ce processus peut se faire par différents moyens, de façon classique par les méthodes de la biologie cellulaire et moléculaire [37], mais aussi grâce aux techniques issues de la génomique et de la bioinformatique [38].

Une cible peut être caractérisée de différentes façons : une structure, un mécanisme biologique, une séquence nucléique ou protéique, *etc*. Chacune de ces caractérisations ouvre des possibilités d'étude dans l'optique de la découverte de nouvelles molécules actives. Par exemple, dans le cadre de mon travail de thèse, il m'était essentiel de disposer d'une structure tridimensionnelle fiable de la cible afin d'accéder aux techniques *in silico* basées sur celle-ci (dites de "*structure-based drug design"* [39]).
L'accessibilité expérimentale *in vitro* et *in vivo* de la cible est également un paramètre crucial pour pouvoir mesurer l'activité biologique d'une molécule donnée sur la cible. Si le coût d'un tel test est important, cela limitera considérablement le nombre de molécules que l'on pourra envisager de tester expérimentalement par la suite.

II-2. Identification des composés prometteurs (touches)

Une fois la cible pharmaceutique identifiée et validée, plusieurs options sont possibles dont celle du criblage haut-débit robotisé pouvant être employée comme point de départ d'un programme de recherche de nouveaux médicaments. Cette approche permet l'évaluation, dans un temps très court, des propriétés pharmacologiques de l'ensemble des molécules contenues dans une chimiothèque sur une ou des cible(s) d'intérêt pharmacologique donnée(s).

Cette technique permet la recherche à grande échelle de substances actives sur une cible caractérisée (quelques milliers à quelques millions de composés testés en parallèle sur celle-ci). Initialement développé dans l'industrie pharmaceutique, le criblage haut-débit devient progressivement accessible aux laboratoires académiques grâce à l'émergence de plateformes techniques, actuellement encore en nombre limité.

L'obtention de résultats positifs et leur confirmation conduisent à l'identification de *touches* (ou "*hits*"). Il s'agit de composés interagissant significativement plus que la moyenne des autres composés testés sur la cible visée. Toutes les touches identifiées ne présentent pas obligatoirement les critères pour devenir un candidat médicament (p. ex. : compétitif vis-à-vis d'une molécule de référence, actif *in vivo*…) [40].

FIG. 7 - Aperçu d'une plateforme de criblage haut-débit robotisé et d'une plaque à 96 puits.

Le criblage à haut débit, par le coût de ses tests (estimé à ~1 € par molécule), ne permet pas de tester des millions de composés sans disposer de moyens financiers conséquents. Pour cette raison, les compagnies pharmaceutiques peuvent avoir recours aux techniques de criblage virtuel [41, 42] en complément du criblage robotisé [43]. En milieu académique, pour cause de moyens plus limités, cette approche est davantage privilégiée.

Ces techniques *in silico* sont basées sur les connaissances accumulées à propos du système étudié et qui peuvent être : la structure de ligands de référence ("*ligand-based drug design*" [43]) ou la structure tridimensionnelle de la cible ("*structure-based drug design*" [39]). Ces techniques se révèlent être très utiles pour réduire les temps de recherche d'identification des nouvelles touches et, par conséquent, les coûts qui y sont associés. En effet, elles permettent d'explorer de façon exhaustive l'espace chimique[*] [44, 45] et de proposer au final une liste raisonnable (financièrement parlant) de molécules à tester expérimentalement et/ou à synthétiser. Cette rationalisation de l'identification de nouvelles touches diffère de l'approche aléatoire du criblage haut-débit exécuté "à l'aveugle", mais permet ainsi, en la complétant, de l'étendre aux cibles de plus en plus complexes de la recherche pharmaceutique.

La diversité moléculaire des chimiothèques criblées, expérimentalement ou virtuellement, est un critère important pour mieux couvrir l'espace chimique et donc pour augmenter les chances

[*] L'avantage du criblage virtuel est de pouvoir tester *in silico* aussi bien des composés déjà synthétisés que d'autres qui ne le sont pas encore et qui peuvent avoir été générés par un programme informatique.

d'identifier de nouvelles touches. Cette diversité peut, entre autres, être obtenue par les techniques de synthèse combinatoire ou par l'accès aux substances naturelles issues de la biodiversité. Les molécules doivent également avoir les propriétés physico-chimiques caractérisant un médicament. Pour cela, des règles élémentaires permettent de filtrer rapidement hors des librairies criblées celles ne répondant pas à ces règles ; les plus connues sont celles de Lipinski [46].

Le travail de recherche effectué au cours de ma thèse ayant porté, en grande partie, sur le criblage virtuel, son concept et ses principales techniques seront détaillés ultérieurement dans ce mémoire.

II-3. Mise au point et optimisation de composés spécifiques (têtes de série)

Parmi les touches obtenues à l'issue d'une campagne de criblage, il convient d'identifier celles qui représentent des *têtes de série* (ou "*leads*") pertinentes vis-à-vis de la pathologie à traiter. Ces composés présentent une activité pour la cible très supérieure à la moyenne des touches et sont également sélectifs pour celle-ci lors d'un test expérimental.

De façon générale, à partir d'un nombre important de molécules de départ, on effectue au moins trois filtrages successifs : le premier sert à identifier les touches, le second à sélectionner les têtes de série, enfin le troisième niveau de filtrage correspond à la sélection éventuelle, après optimisation des têtes de série, d'un ou plusieurs composés candidats pour les tests cliniques. Des études SAR (*Structure Activity Relationship*) sont effectuées au cours des deux dernières étapes de filtrage. De la découverte de touches à la validation de candidats pré-cliniques, le coût des techniques employées, ainsi que l'expertise humaine nécessaire, augmentent considérablement. Seules la découverte de touches et, dans une moindre mesure, la sélection de têtes de série, peuvent correspondre à des protocoles de filtrage plus ou moins automatisés.

II-4. Essais pré-cliniques et cliniques

Les essais pré-cliniques ont pour objectif d'évaluer, avant l'étude chez l'homme, l'innocuité de la molécule (toxicité, mutagenèse, cancérogenèse,…), son action sur les organes cibles, ainsi que son cycle de vie dans l'organisme (absorption, propagation, élimination).

La phase I des essais cliniques correspond à la première administration à l'homme, effectuée sur des centaines de volontaires sains durant 6 à 18 mois, afin de d'évaluer la toxicité potentielle du produit.

La phase II, effectuée en général en milieu hospitalier, sur un groupe de malades durant 2 à 3 ans, a pour but de déterminer les conditions optimales d'administration (dose et posologie) conduisant au minimum d'effets secondaires. La phase III, pouvant être étendue à plusieurs milliers de patients, mesure l'efficacité du médicament (rapport traitement de la pathologie / effets secondaires) dans les conditions d'utilisation préconisées, en concurrence avec un placebo et d'éventuels médicaments de référence. Une autorisation de mise sur le marché est délivrée en cas de succès et le suivi post-commercialisation constitue la phase IV.

II-5. Bilan financier

La mise au point d'un médicament est une activité sur le long terme et particulièrement onéreuse. En effet, l'industrie pharmaceutique consacre, à l'heure actuelle, jusqu'à 15 ans pour amener un médicament du stade de concept à sa commercialisation, pour un coût total estimé à environ 1 milliard de dollars [47]. On explique souvent ce rendement faible par le fait que, sur 10 000 molécules testées, 10 seulement parviennent au stade des essais clinique et qu'une seule satisfait finalement tous les tests cliniques et parvient plus tard au stade de la mise sur le marché en tant que médicament.

Les méthodes informatiques jouent un rôle crucial pendant les différentes phases préliminaires du développement d'un médicament (cf. FIG. 8 ci-dessous), car elles permettent souvent une réduction des coûts et un traitement plus approfondi des programmes de recherche, en complément des approches expérimentales [48]. Il est également reconnu que seule une collaboration étroite entre les disciplines différentes mais complémentaires (en particulier, nous concernant, entre les expérimentateurs et les théoriciens) peut garantir une efficacité maximale pour la recherche et le développement de nouveaux médicaments. Une telle approche est considérée comme cruciale dès les premières étapes du processus, car celles-ci conditionnent directement les étapes ultérieures dont le coût est de plus en plus important.

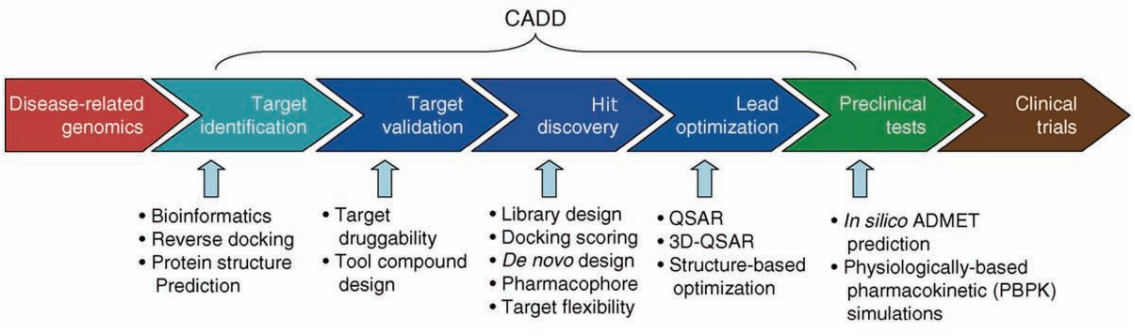


FIG. 8 - Pipeline de développement d'un médicament sur lequel sont indiqués les endroits où les méthodes in silico peuvent intervenir (CADD : Computer-Aided Drug Design). D'après (Tang et al., 2006) [49].

III- RÔLE DES MÉTHODES INFORMATIQUES DANS L'ÉTUDE DU VIVANT ET CONTRIBUTION DE LA MODÉLISATION MOLÉCULAIRE

III-1. L'essor de l'informatique et l'avènement des simulations

L'informatique a été une révolution, car elle a permis à la communauté scientifique de pouvoir étudier des systèmes de plus en plus complexes. Elle a offert la possibilité de conduire des calculs étendus et automatisés sur des systèmes que leur taille et leur complexité ne rendent pas traitables par d'autres moyens. Les techniques informatiques permettent ainsi de se rapprocher de la réalité physique, validant ou invalidant les modèles utilisés avec plus de rigueur.

L'apparition de moyens de calcul informatique de plus en plus importants[*] à partir des années 1950 a altéré le binôme théorie-expérience de la recherche scientifique en insérant une méthode intermédiaire : la simulation numérique. Fondamentalement, l'aspect théorique aussi bien que l'aspect expérimental de la recherche sont intégrés dans l'outil informatique, le premier à travers des programmes et des algorithmes et le second à travers des simulations et des données. Ainsi, la conduite d'une simulation peut être analogue, selon les circonstances, à celle d'une expérience, ou à la mise en place d'une théorie, rendant l'une et l'autre plus accessibles et renforçant leur lien.

Un autre avantage considérable de l'informatique est qu'une simulation numérique n'est soumise à aucune autre contrainte matérielle que celle du temps de calcul nécessaire à son exécution. Elle permet ainsi de pouvoir conduire des expériences dans des conditions expérimentales inaccessibles ou trop dangereuses (p. ex. conditions de température et de pression délicates), ou encore de repousser artificiellement certaines contraintes physiques, afin de mieux appréhender le comportement du système d'étude. Ainsi, les simulations numériques peuvent, non seulement consolider le champ de la recherche, mais aussi l'étendre.

[*] Progression qui suit la loi de Moore, selon laquelle la puissance de calcul des microprocesseurs double approximativement tous les 18 mois (hypothèse vérifiée et toujours d'actualité).

III-2. La modélisation moléculaire et ses applications

Le terme de modélisation moléculaire *in silico* est assez général et, sous cette étiquette, on trouve un certain nombre de techniques : mécanique moléculaire, chimie quantique, simulation de dynamique moléculaire, construction par homologie, criblage virtuel, chemoinformatique, *etc.* Ces méthodes théoriques et/ou empiriques appartiennent toutes à la famille des techniques informatiques, par opposition aux méthodes expérimentales.

Depuis ces dix dernières années, la modélisation moléculaire connaît un intérêt tout particulier dans l'ère de la post-génomique grâce à son large champ d'applications possibles. Elle peut, en particulier, s'avérer précieuse lors de la mise au point d'inhibiteurs pour une cible thérapeutique par le biais du criblage virtuel [41, 42], souvent par l'utilisation des algorithmes d'arrimage moléculaire (ou "*docking*") [50]. Le grand intérêt d'utiliser les méthodes dites de "*structure-based drug design*" (*SBDD*) [39] peut être illustré par des succès tels que la mise au point d'inhibiteurs de l'HIV protéase [51] ou d'autres cibles [52-55].

Les méthodes de SBDD requièrent une structure tridimensionnelle de la cible étudiée. Cependant, diverses techniques, comme la modélisation par homologie, peuvent permettre de pallier l'absence de structures complètes dans le cas de certaines cibles d'intérêt majeur, telles que les récepteurs couplés aux protéines-G [56, 57]. La simulation par dynamique moléculaire [58], alliée à la puissance de calcul actuelle, est une autre technique de la modélisation moléculaire qui permet, entre autres, l'étude des propriétés dynamiques de systèmes biologiques de grande taille [59].

La modélisation profite directement des progrès dans une grande variété de domaines. En premier lieu, on peut citer les progrès de l'informatique, aussi bien sur le plan matériel (puissance de calcul brute, architectures parallèles et distribuées) que logiciel (outils spécialisés, techniques de programmation). La disponibilité croissante de structures expérimentales de bonne qualité caractérisant des cibles potentielles est également très profitable ; les progrès technologiques des appareillages ont donc un impact déterminant pour les modélisateurs. Il en va de même pour les avancées purement théoriques, ainsi que les nouvelles découvertes en biologie et biochimie, qui peuvent permettre d'améliorer les modèles au niveau conceptuel. Le travail du modélisateur se fait toujours dans un contexte interdisciplinaire, à l'interface des disciplines que sont la chimie, la biologie et l'informatique.

V- PRÉSENTATION DES TRAVAUX DE RECHERCHE

Le lecteur trouvera dans le présent manuscrit les principaux résultats issus de la thèse que j'ai effectuée sous la direction de Bernard Maigret, d'abord au sein de l'université Henri Poincaré / Nancy I (UMR 7565, équipe eDAM), puis au LORIA (UMR 7503, équipe ORPAILLEUR). Ce travail a deux aspects : il rassemble des développements méthodologiques et des applications à travers l'étude d'un système biologique particulier. Le point commun de ces deux aspects est la recherche de nouveaux médicaments par le biais de techniques informatiques.

La première partie des résultats (section 3.I) présente la plateforme logicielle VSM-G pour laquelle j'ai été le principal développeur jusqu'à ce jour. VSM-G repose sur plusieurs concepts novateurs dont la pertinence doit être validée sur le plan scientifique en parallèle à leur implémentation. On trouvera d'abord la description d'une des composantes de VSM-G, suivie d'une présentation générale avec preuve de concept de la plateforme. L'étude des caractéristiques d'un système-cible viendra ensuite justifier certains choix effectués pour l'application à un problème pharmaceutique concret. Nous présenterons enfin brièvement les résultats de la campagne de calculs correspondante. Le projet VSM-G s'inscrit dans un contexte de multiples collaborations, dont les principales ont impliqué l'équipe de Wensheng Cai (Université de Nankai, R.P.Chine), Peter Bladon (Interprobe, Royaume-Uni) et Gilles Moreau pour l'aspect développement d'une part ; Michel Souchet et Sinan Karaboga (Fournier Pharma / Solvay, Daix) pour la campagne d'application/validation d'autre part.

La seconde partie des résultats (section 3.II) présente les principaux résultats de notre contribution au projet ANR multidisciplinaire "Tyrosines kinases de la famille de FAK : bases structurales de la régulation et de la localisation intracellulaire". Mon travail s'est concentré sur le domaine FAT, une des composantes de la protéine FAK, cible de grand intérêt pour la lutte anti-cancer. Par l'étude de la dynamique de ce système, j'ai d'abord cherché à mettre en évidence les caractères structuraux de son activité biologique. Une première phase de mise au point de molécules visant à réguler cette activité a ensuite été entreprise. Dans le cadre du projet FAK, j'ai collaboré en particulier avec Jean-Antoine Girault (Université Pierre et Marie Curie / Paris VI), Stephan Arold (Centre de Biochimie Structurale de Montpellier) et Mercedes Martinez (Institut de Chimie Médicinale de Madrid).

Avant de présenter ces résultats, nous passerons brièvement en revue, dans la partie méthodologie qui suit, les connaissances actuelles relatives aux méthodes utilisées pour ce travail. Nous nous focaliserons ainsi sur la mécanique moléculaire puis sur les techniques informatiques de recherche de médicaments.

PARTIE 2 - MÉTHODOLOGIE

I- LA MÉCANIQUE MOLÉCULAIRE

Dans le domaine de la modélisation moléculaire, les méthodes reposant sur les lois de la mécanique quantique permettent de fournir une description satisfaisante, sur le plan théorique, de l'énergie et de l'état d'un système. Les techniques correspondantes sont basées sur une résolution mathématique approchée[*] de l'équation de Schrödinger. Toutefois, la complexité associée à de tels calculs est si importante que ces simulations dites *ab initio* sont limitées à des modèles ne comportant qu'un petit nombre d'atomes (<< 1000). En contrepartie, ces calculs traitent explicitement les électrons d'un système moléculaire et permettent ainsi l'étude précise de sa réactivité chimique (création ou destruction de liaisons).

La modélisation de systèmes moléculaires repose sur l'approximation de Born-Oppenheimer [60] qui considère les noyaux atomiques fixes par rapport aux électrons[†]. Les techniques *ab initio* se concentrent ainsi sur la caractérisation des seuls électrons. La simulation de systèmes de grande taille tels que les biomolécules nécessite d'avoir recours à des modèles bien plus simplifiés. Ainsi, en mécanique moléculaire, on ne traite plus les électrons mais les atomes représentés comme des sphères. La première conséquence de ce choix est l'impossibilité de reproduire rigoureusement tout phénomène électronique (p. ex. la conductivité, les réactions chimiques).

En raison de leurs approximations, les méthodes de mécanique moléculaire ne sont donc pas adaptées à la description des propriétés électroniques des systèmes étudiés, mais elles permettent, sous certaines conditions, de déterminer de façon appropriée leurs propriétés structurales et thermodynamiques. La plupart du temps, ces méthodes, à condition d'être correctement paramétrisées, mènent à un résultat analogue à ce qui aurait été obtenu par la mécanique quantique et ceci dans un intervalle de temps de calcul bien inférieur. La mécanique moléculaire offre des techniques de choix (p. ex. dynamique moléculaire) pour simuler des systèmes biologiques de taille importante (de 10^5 à 10^6 atomes) et permet ainsi d'obtenir des informations détaillées reliant la structure d'un système biologique à sa fonction.

[*] L'équation de Schrödinger ne peut pas mathématiquement être résolue de façon exacte pour les systèmes moléculaires.

[†] En effet, la masse des noyaux étant ~ 1 000 fois plus importante que celle des électrons, on peut supposer que les électrons se déplaceront beaucoup plus vite que les noyaux et donc considérer ces derniers comme fixes.

I-1. Notions de potentiels et de champ de force

La mécanique moléculaire s'appuie sur les lois de la mécanique Newtonienne pour modéliser le comportement des systèmes moléculaires. Elle assimile les atomes d'un système à un ensemble de masses ponctuelles chargées interagissant entre elles à travers une fonction d'énergie potentielle prédéfinie. Cette fonction d'énergie potentielle, associée à un jeu de paramètres qui lui est propre, constituent ce qu'on appelle un champ de force. Celui-ci définit le lien entre la structure géométrique des atomes du système d'étude et leur énergie potentielle au cours de la simulation [61].
La fonction d'énergie potentielle est la clé de voûte des calculs moléculaires, car son rôle est de reproduire les interactions intra- et intermoléculaires du système étudié aussi fidèlement que possible. Elle peut être décomposée par la somme des potentiels régissant l'interaction entre atomes liés d'une part (liaisons covalentes, angles de valence, dièdres, *etc.*) et ceux régissant l'interaction entre atomes non-liés (électrostatique, van der Waals) d'autre part :

$$V_{total} = V_{liés} + V_{non-liés} \qquad (1)$$

I-1.1. Potentiel entre atomes liés

Le potentiel entre atomes liés au sein d'une structure moléculaire peut se décomposer en interactions entre séries de deux, trois, quatre, ... atomes liés. Si l'on suppose que ces différents termes sont découplés (cas des champs de force de classe I tels que CHARMM [62-65], AMBER [66-72], GROMOS [73, 74], OPLS [75, 76]), et que l'on néglige les termes d'ordres supérieurs à 5, on obtient l'expression générale la plus courante du potentiel entre atomes liés :

$$V_{liés} = V_{liaisons} + V_{angles} + V_{dièdres} + V_{\substack{dièdres \\ impropres}} \qquad (2)$$

$\mathbf{V}_{liaisons}$ décrit l'interaction formée à partir d'une liaison covalente entre 2 atomes ; elle est caractérisée au niveau de la topologie par les longueurs des liaisons r. $\mathbf{V}_{angles}$ est l'interaction pour une suite de 3 atomes liés consécutivement ; il s'agit de l'angle de valence θ. $\mathbf{V}_{dièdres}$ et $\mathbf{V}_{dièdres\ impropres}$ s'appliquent à une suite de 4 atomes et correspondent aux angles dièdres ϕ et ω respectivement.

Pour les groupes d'atomes auxquels ils s'appliquent, chacun des termes est minimal lorsque les valeurs des paramètres géométriques associés correspondent aux valeurs d'équilibre

thermodynamique. L'éloignement par rapport à ces valeurs de référence correspond à une différence de potentiel qui peut être assimilée à une pénalité.

Le formalisme le plus fréquent emploie des fonctions harmoniques :

$$V_{liaisons} = \sum_{liaisons} k_r (r - r_0)^2 \qquad (3)$$

$$V_{angles} = \sum_{angles} k_\theta (\theta - \theta_0)^2 \qquad (4)$$

$$V_{dièdres} = \sum_{dièdres} k_\phi (1 + \cos(n\phi - \gamma)) \qquad (5)$$

$$V_{\substack{dièdres \\ impropres}} = \sum_{dièdres'} k_\omega (\omega - \omega_0)^2 \qquad (6)$$

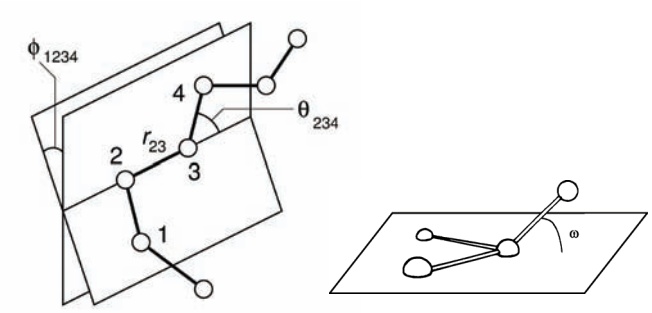


FIG. 9 - Représentation schématique des 4 contributions clés d'un champ de force pour les atomes liés.

I-1.2. Potentiel entre atomes non-liés

De façon analogue à $\mathbf{V}_{liés}$, la partie $\mathbf{V}_{non\text{-}liés}$ de la fonction d'énergie potentielle peut se développer en somme de termes à n corps (n > 1). On se limite en général aux termes à deux corps, d'une part parce que les suivants ont une amplitude bien plus faible, d'autre part pour des raisons de complexité[*]. Dans l'expression des champs de force, on entend par atomes *non-liés* aussi bien les atomes de deux molécules distinctes que ceux d'une même molécule mais qui sont séparés par trois liaisons covalentes au minimum[†].

[*] Contrairement aux termes $V_{liés}$ qui ne s'appliquent qu'aux atomes liés (ce qui en limite le nombre), les termes $V_{non\text{-}liés}$ s'appliquent au moins à toute paire d'atomes du système, d'où une complexité en $O(n^2)$ en l'absence d'optimisation.

[†] L'interaction entre atomes liés par moins de 3 liaisons étant déjà prise en charge dans l'expression de $V_{liés}$.

Dans sa formulation par paires, on distingue du potentiel entre atomes *i* et *j* non-liés les interactions électrostatiques des interactions dites de van der Waals :

$$V_{non-liés}(\vec{r}_1,\vec{r}_2,...,\vec{r}_n) \approx \sum_i \sum_j V_2(\vec{r}_i,\vec{r}_j) \qquad (7)$$

$$V_2(r_{ij},q_i,q_j) = V_{électrostatique}(r_{ij},q_i,q_j) + V_{vdW}(r_{ij}) \qquad (8)$$

FIG. 10 - Représentations schématiques d'interaction électrostatique (gauche) et d'interaction de van der Waals (droite) entre atomes non-liés.

Les interactions électrostatiques se définissent par le potentiel de Coulomb classique qui s'établit entre deux atomes *i* et *j* séparés par une distance r_{ij} et portant respectivement les charges ponctuelles[*] q_i et q_j. Cette interaction dépend également de la constante diélectrique du milieu ε_0 :

$$V_{électrostatique}(r_{ij},q_i,q_j) = \frac{1}{4\pi\varepsilon_0}\frac{q_i q_j}{r_{ij}} \qquad (9)$$

En ce qui concerne le terme de van der Waals, une formulation simple, qui traduit la nature des interactions de type répulsion et dispersion, est généralement suffisante. Celle de Lennard-Jones [77] est l'une des plus employées :

$$V_{vdW}^{L-J}(r_{ij}) = 4\varepsilon\left[\left(\frac{\sigma}{r_{ij}}\right)^6 - \left(\frac{\sigma}{r_{ij}}\right)^{12}\right] \qquad (10)$$

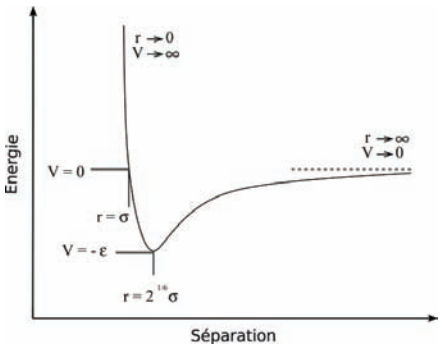


FIG. 11 - Le potentiel de Lennard-Jones.

[*] La répartition globale des charges sur une molécule est simplifiée à des charges partielles localisées sur les atomes et dont la valeur dépend de leur électronégativité. Ces charges partielles peuvent être déterminées par des calculs *ab initio* sur des petites molécules. Les monopoles atomiques sont ensuite optimisés pour reproduire au mieux la distribution du potentiel électrostatique autour des molécules étudiées.

Le paramètre ε est la valeur absolue du minimum énergétique. σ correspond au point d'équivalence entre la composante répulsive à courte distance en $1/r^{12}$ et la composante attractive à longue distance en $1/r^6$, c'est-à-dire la distance interatomique à partir de laquelle le potentiel devient attractif. Le terme $1/r^6$ traduit le couplage dipôle-dipôle à l'origine de l'interaction van der Waals.

Le tableau ci-dessous récapitule les variables et les paramètres de la fonction énergie potentielle. Les variables (en bleu), caractérisant l'état géométrique du système à un instant *t*, sont toutes reliées aux coordonnées (*x*,*y*,*z*) de l'ensemble des atomes. Les paramètres (en vert) caractérisent la topologie constante du système et sont définis pour chaque ensemble d'atomes par le champ de force.

<table>
<tr><th>Force</th><th>Nombre d'atomes impliqués</th><th>Variables</th><th>Constantes de force</th><th>Termes d'équilibre</th><th>Autres paramètres</th></tr>
<tr><td>van der Waals</td><td rowspan="2">2 non-liés</td><td rowspan="2">r</td><td rowspan="2">-</td><td>σ</td><td>ε</td></tr>
<tr><td>Electrostatique</td><td>-</td><td>q_1, q_2, ε_0</td></tr>
<tr><td>Liaisons</td><td>2 liés</td><td>k_r</td><td>r_0</td><td rowspan="4">-</td></tr>
<tr><td>Angles</td><td rowspan="2">3 liés</td><td>θ</td><td>k_θ</td><td>θ_0</td></tr>
<tr><td>Dièdres</td><td>φ</td><td>k_φ</td><td>n, γ</td></tr>
<tr><td>Dièdres impropres</td><td>4 liés</td><td>ω</td><td>k_ω</td><td>ω_0</td></tr>
</table>

TAB. 1 - Tableau récapitulatif des variables et des paramètres de la fonction d'énergie potentielle.

Notons que certains champs de force incluent, dans l'expression du potentiel entre atomes non-liés, un terme décrivant les liaisons hydrogène qui, dans sa forme la plus simple, est du type :

$$V_{HB} = \sum_{paires} \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{10}} \qquad (11)$$

D'autres encore contiennent également des termes dus au couplage entre les variations de différentes coordonnées internes. C'est le cas des champs de force dits de classe II tels que, par exemple, UFF [78], CFF [79] et MM3 [80].

I-2. Paramétrisation du champ de force et son application aux biomolécules

Avant toute simulation de mécanique moléculaire, il est nécessaire de procéder à l'application du champ de force. Ce processus relie la structure du système (liste des atomes, de leur nature, de leurs coordonnées, des liaisons qu'ils forment et de la nature de celles-ci) à sa topologie (liste des groupes d'atomes que chaque terme du potentiel doit considérer). Celle-ci est ensuite associée aux paramètres prédéfinis dans le champ de force. Les paramètres sont, au cours d'une simulation de mécanique moléculaire, des constantes qui doivent être définies rigoureusement avant toute simulation. La détermination des paramètres du champ de force s'effectue sur la base de résultats expérimentaux (données spectroscopiques, calorimétriques, structurales,…), ou lorsque ceux-ci ne sont pas disponibles (ou trop peu précis), à partir de résultats de calculs de mécanique quantique *ab initio* [81]. Cela correspond à différentes procédures itératives complexes [66, 82-84].

La transférabilité des paramètres d'une molécule vers un groupement structuralement similaire d'une autre molécule est l'une des hypothèses fondamentales de la mécanique moléculaire. Bien que cette approche puisse paraître rudimentaire, les multiples applications effectuées depuis son introduction ont permis de démontrer son efficacité. La validation d'un champ de force correspond à la reproduction, sur la base de celui-ci, de résultats expérimentaux. Notons que chaque champ de force constitue un compromis entre simplicité conceptuelle et précision. L'amélioration d'un champ de force est délicate, par exemple à cause de la détermination de la corrélation entre paramètres et résultats expérimentaux, ou encore du fait que différents paramètres peuvent être couplés [85]. Toutefois, l'augmentation des performances informatiques (loi de Moore) permet d'y incorporer des termes de plus en plus sophistiqués sans que cela se fasse trop sentir sur les temps de calcul.

Pour conduire une simulation, le choix du champ de force est à faire sur la base des résultats déjà obtenus dans la littérature concernant son application aux systèmes moléculaires analogues au système étudié. En effet, l'emploi d'un champ de force paramétrisé pour décrire une certaine catégorie de molécules (p. ex. les protéines) se limite à une classe de molécules ayant des ressemblances structurales et fonctionnelles. Les champs de force couramment utilisés pour la modélisation de biomolécules [86], tels que CHARMM [62-65], AMBER [66-72], GROMOS [73, 74] et OPLS [75, 76], fruits de nombreuses années de recherches spécialisées, peuvent être considérés comme matures. Leur paramétrisation est spécifiquement orientée vers la reproduction correcte du comportement des acides aminés et des acides nucléiques. Ils permettent donc, couplés à des programmes appropriés, tels que NAMD [87-89], GROMOS [90], CHARMM [91, 92] et AMBER [93, 94], de conduire des simulations de dynamique moléculaire considérées comme fidèles a priori pour des systèmes biomoléculaires aussi classiques que des complexes protéiques.

I-3. Principal axe de développement des champs de force : la polarisation

Le développement de champs de force polarisables - qui soient utilisables de façon courante pour conduire des simulations de dynamique moléculaire sur les biomolécules - est un enjeu majeur de la nouvelle génération de champs de force. Un certain nombre de revues récentes sur le sujet témoignent de son attractivité et des efforts nécessaires à leur développement [61, 86, 95]. La plupart des champs de force non polarisables actuels (cités dans la précédente section) utilisent un modèle de charge ponctuelle fixe pour décrire la charge électrique de chaque atome d'un système et traitent la polarisation de façon moyennée et implicite dans leurs paramètres d'interaction entre atomes non-liés [86, 95]. Cette charge est invariante au cours d'une simulation et ne rend donc pas compte explicitement des phénomènes de polarisation, c'est-à-dire de l'influence de l'environnement électrostatique local qui peut changer la distribution de charge d'un système.

Les champs de force non polarisables sont donc limités dès lors qu'il s'agit de traiter des systèmes pour lesquels les effets de polarisation électronique sont importants [96]. Par exemple, le site actif d'une enzyme peut fortement se polariser à l'approche d'un ligand polaire ou chargé. Une application directe des champs de force polarisables dans le domaine du *drug design* serait de les utiliser de façon courante pour évaluer très précisément l'énergie d'interaction d'une protéine avec un ligand [86].

La prise en compte des effets de polarisation par un champ de forces nécessite, au minimum, d'ajouter un terme d'interactions supplémentaire entre atomes *non-liés* dans la formulation de la fonction d'énergie potentielle. D'un point de vue technique, les méthodes les plus courantes comprennent les modèles de dipôles induits, les modèles de charges fluctuantes ou encore les oscillateurs de Drude pour traiter ces effets [61, 97, 98].

Bien qu'il soit reconnu que l'inclusion de la polarisation dans les champs de force est essentielle pour en améliorer la description des phénomènes de reconnaissance moléculaire, les temps de calculs associés sont en général considérables au point que le modélisateur préfère parfois l'ignorer au profit de plus longues simulations. Enfin, leur développement restant délicat et leur traitement efficace en dynamique moléculaire n'étant pas encore établi, l'utilisation de tels champs de force pour simuler des biomolécules polarisables n'est pas encore courante. Cela explique pourquoi les principaux travaux publiés utilisant les champs de force polarisables sont issus des équipes qui les développent. Parmi les champs de force polarisables, on peut citer de façon non exhaustive : SIBFA [99-101], PFF [102, 103], SDFF [104] et AMOEBA [105].

I-4. Méthodes d'exploration de l'hypersurface d'énergie d'un système moléculaire

I-4.1. Minimisation de l'énergie

La minimisation de l'énergie potentielle est l'une des méthodes les plus couramment employées pour optimiser la structure des biomolécules. Cette opération peut être utile afin de relaxer la structure initiale de la molécule (en éliminant les mauvais contacts atomiques) et de rechercher une géométrie de basse énergie, correspondant à un état stable du système. La minimisation de l'énergie potentielle trouve ainsi son application dans les protocoles de raffinement de structures moléculaires obtenues expérimentalement (RMN ou RX) et dans l'optimisation de la structure de biomolécules en préalable à une simulation de dynamique moléculaire.

Le nom d'"hypersurface" est employé pour représenter l'énergie potentielle du système étudié par un espace dans lequel les positions atomiques sont des variables. Le problème consiste à trouver les coordonnées des atomes qui permettent de minimiser la fonction d'énergie potentielle (1) du système étudié. La solution à ce problème est loin d'être triviale car le nombre important de degrés de liberté fait fluctuer le système autour d'un ensemble de conformations stables correspondant à différents minima sur l'hypersurface d'énergie potentielle. L'état de plus basse énergie, sur l'ensemble de l'hypersurface de potentiel, est appelé minimum *global*, par opposition aux minima *locaux*.

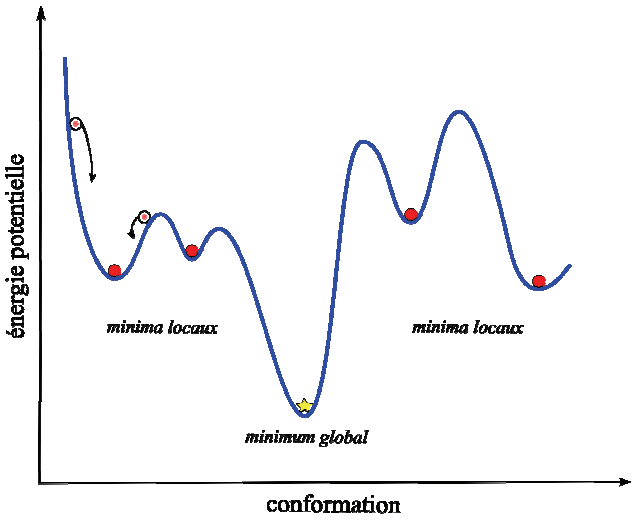


FIG. 12 - Représentation schématique d'une hypersurface d'énergie potentielle pour un seul degré de liberté.

Il n'existe pas de méthode mathématique générale qui permette de trouver ce minimum global. Cependant, des méthodes d'analyse numérique, telles que les algorithmes de minimisation, peuvent être employées afin de trouver des minima locaux. L'approche la plus souvent utilisée consiste à descendre de manière itérative le long de l'hypersurface dans la direction correspondant à une diminution de l'énergie ; ce qui peut se résumer en :

$$r_{n+1} = r_n + \delta_n \qquad (12)$$

où r_i correspond aux coordonnées des atomes du système étudié, n est le numéro de l'itération et δ_n le énième déplacement dans l'espace des configurations.

Les algorithmes de minimisation les plus courants [106] diffèrent tous dans la manière de se déplacer sur l'hypersurface, c'est-à-dire dans l'évaluation des δ_n. Celles de la plus grande pente et des gradients conjugués utilisent uniquement les dérivées premières de la fonction d'énergie potentielle, alors que les méthodes dites de Newton-Raphson requièrent le calcul des dérivées secondes de cette fonction.

La méthode de la plus grande pente s'avère efficace lorsque la conformation de départ est loin du minimum ; cependant elle converge lentement et peut adopter un comportement oscillatoire au voisinage de celui-ci. La méthode du gradient conjugué converge mieux, car elle présente l'intérêt d'échapper aux comportements oscillatoires autour du minimum tout en accélérant la convergence, mais elle n'est pas adaptée aux structures présentant de mauvais contacts ou pas très bien relaxées. Au vu des avantages et inconvénients de chacune des deux méthodes, on a généralement recours à la combinaison des deux algorithmes de minimisation l'un après l'autre.

De leur côté, les méthodes de Newton-Raphson utilisent la dérivée seconde de l'énergie en calculant de manière explicite le hessien H_i de l'énergie potentielle (matrice des dérivées secondes) pour rechercher l'extremum le plus proche. Ces méthodes permettent de converger très rapidement vers le minimum local, à condition d'en être déjà assez proche au préalable. Le temps de calcul requis par les méthodes de Newton-Raphson, ainsi que l'espace mémoire nécessaire, les rendent non adaptées à l'étude de gros systèmes comme les biomolécules.

I-4.2. Méthodes de recherche conformationnelle

Les algorithmes de minimisation permettent d'obtenir une structure stable de la molécule d'étude. Cependant, pour éviter d'être piégé dans le minimum le plus proche, il est conseillé d'effectuer un échantillonnage de l'espace conformationnel. Cette recherche a pour but de déterminer par la simulation un ensemble représentatif de conformations stables de la molécule à l'équilibre thermodynamique, c'est-à-dire susceptibles d'exister dans les conditions expérimentales.

Outre la recherche systématique, les méthodes de recherche conformationnelle peuvent être divisées en deux grandes classes : les méthodes stochastiques et les méthodes déterministes. Nous évoquerons brièvement les méthodes stochastiques à travers la méthode Monte Carlo. Les méthodes déterministes seront illustrées en décrivant la dynamique moléculaire, technique dont l'utilisation fut centrale pour ce travail.

Recherche systématique

L'approche d'exploration d'une hypersurface la plus évidente consiste à en évaluer tous les états possibles. Dans le cas d'un système moléculaire, il s'agit de calculer l'énergie de chaque conformère en faisant varier les angles dièdres considérés par pas de n degrés[*]. Le nombre de conformères théoriques est alors de P^i, où P est le nombre d'états possibles par torsion pour chaque angle dièdre, et i est le nombre d'angles dièdres contenus dans la molécule

Par exemple, une recherche systématique sur une molécule possédant 4 angles dièdres avec un pas de n = 30°, générerait 20 736 conformères (P = 360/30 = 12 → 12^4 = 20 736).
Une telle approche s'avère inutilisable, même sur des molécules relativement simples, car trop coûteuse en temps de calcul du fait de sa complexité exponentielle par rapport au nombre de dièdres. Les méthodes stochastiques ou déterministes sont alors privilégiées.

[*] On suppose ici que les angles de valence et les longueurs de liaison sont fixés aux valeurs d'équilibre. Cette approximation est acceptable car les variations par rapport à l'état d'équilibre des termes correspondants sont bien plus importantes que celles des autres termes du potentiel décrits dans les équations (3) et (4).

Recherche stochastique – exemple du Monte Carlo

Contrairement à la recherche systématique, la recherche stochastique fait varier les degrés de liberté du système de manière aléatoire. La méthode de Monte Carlo [106] est basée sur la génération aléatoire d'un ensemble de conformations et utilise certains critères de sélection pour conserver, ou non, les nouvelles conformations créées. Ces critères assurent que la probabilité d'obtenir une conformation donnée est égale à son facteur de Boltzmann[*] $e^{-\frac{V(r_i)}{k_B T}}$, où k_B est la constante de Boltzmann, T la température et $V(r_i)$ est calculé par la fonction d'énergie potentielle. La plupart des simulations de Monte Carlo se réfèrent à l'utilisation de l'algorithme de Metropolis [107].

Dans une simulation de Monte Carlo, chaque nouvelle conformation générée résulte du mouvement aléatoire d'un ou plusieurs atome(s) ou de la variation d'un ou plusieurs angle(s) dièdre(s). L'énergie de la conformation obtenue est alors évaluée. Si la nouvelle conformation est de plus basse énergie que la précédente, elle est acceptée. En revanche, si elle est plus haute en énergie, alors le facteur de Boltzmann de la différence d'énergie entre les deux conformations est calculé et constitue la probabilité d'acceptation de la nouvelle conformation. Cette approche permet de visiter des états de plus haute énergie que la configuration actuelle ; cependant, les conformations de plus basse énergie restent préférentiellement échantillonnées. Les méthodes stochastiques ne garantissent pas l'obtention du minimum global de l'énergie, mais elles permettent d'obtenir un ensemble de conformères de faible énergie constituant un échantillonnage supposé être représentatif.

Recherche déterministe – exemple de la dynamique moléculaire

Alors qu'au cours d'une simulation de Monte Carlo les conformations sont générées de manière aléatoire et discrète (dépendant de la probabilité de transition entre l'état courant et le suivant), celles issues d'une simulation par dynamique moléculaire sont, elles, connectées dans le temps et forment une trajectoire continue. La dynamique moléculaire est dite déterministe puisque tous les états[†] futurs du système peuvent être prédits à partir de l'état actuel. Cette technique de simulation, que nous avons utilisée pour ce travail, est détaillée dans une prochaine section.

[*] Le facteur de Boltzmann caractérise la probabilité relative d'un état *i* dans une situation d'équilibre thermique à la température *T*.

[†] Pour la dynamique moléculaire, un état du système ne contient pas seulement l'ensemble des coordonnées spatiales de chaque atome mais aussi les valeurs des vecteurs quantité de mouvement.

I-5. Les ensembles thermodynamiques

Lors d'une simulation, il est possible de travailler dans différents ensembles thermodynamiques. Les plus communément utilisés sont : l'ensemble microcanonique NVE (nombre d'atomes, volume et énergie constants), l'ensemble canonique NVT (nombre d'atomes, volume et température constants), et l'ensemble isobare-isotherme NPT (nombre d'atomes, pression et température constants).
Les simulations de Monte Carlo s'effectuent traditionnellement dans l'ensemble NVT alors que les simulations de dynamique moléculaire sont généralement associées à l'ensemble NVE. Cependant, par l'insertion de contraintes dans leurs implémentations respectives, ces deux techniques peuvent s'appliquer à d'autres ensembles thermodynamiques. Dans le cadre de ce travail, toutes nos simulations de dynamique moléculaire ont été réalisées dans l'ensemble isobare-isotherme NPT, où la pression et la température du système sont maintenues constantes par couplage à un barostat et à un thermostat respectivement [108, 109]. Cet ensemble est, en effet, celui qui se rapproche le plus des conditions réelles d'expérience de laboratoire.

I-6. La dynamique moléculaire

I-6.1. Principe

La dynamique moléculaire [85, 106] (DM) est une technique couramment utilisée pour la simulation de biomolécules [58]. Son but est d'étudier l'évolution d'un système moléculaire (défini par mécanique moléculaire, voir section I.1) au cours du temps en intégrant les équations de Newton relatives au système :

$$m_i \vec{a}_i = \vec{F}_i \qquad (13)$$

où m_i est la masse d'un atome *i*, $\vec{a}_i$ son accélération et $\vec{F}_i$ la somme des forces qui lui sont appliquées du fait de son interaction avec les autres atomes et l'environnement.

Lors de la simulation de DM, le système subit des changements conformationnels et cinétiques qui permettent d'explorer l'espace des phases espace-temps accessible par le système. À chaque particule, en tout temps *t*, on associe un couple (position $\vec{r}_i(t)$, vitesse $\vec{v}_i(t)$). L'ensemble des coordonnées sur la totalité de l'espace temporel exploré constitue ce qu'on appelle une trajectoire. Suivant l'hypothèse

ergodique, l'étude d'une trajectoire infiniment longue d'un système par DM revient à échantillonner tout l'espace des phases de ce système. Il est alors possible d'accéder à des grandeurs thermodynamiques (coefficients de diffusion, fonctions de distributions radiales, énergie libre, *etc.*) afin de relier la simulation à l'échelle microscopique aux expérimentations à l'échelle macroscopique.

Les équations du mouvement de Newton peuvent s'écrire de la manière suivante :

$$m_i \frac{d\vec{v}_i(t)}{dt} = -\frac{d\vec{V}(\vec{r}_1, \vec{r}_2, ..., \vec{r}_N)}{d\vec{r}_i} = \vec{F}_i(t) \qquad (14)$$

où m_i est la masse d'un atome i, $\vec{v}_i(t)$ sa vitesse à l'instant t, et $\vec{r}_i(t)$ sa position dans l'espace. $\vec{V}(\vec{r}_1, \vec{r}_2, ..., \vec{r}_N)$ est le potentiel d'interaction (champ de force) entre les atomes du système.

À partir d'un système dont les conditions initiales ont été fixées, la DM consiste en la répétition de deux opérations : tout d'abord évaluer la force agissant sur chaque atome au temps t, puis déterminer les coordonnées et les vitesses des atomes au temps $t + \Delta t$ en fonction des forces subies par chacun d'entre eux où Δt correspond au pas d'intégration.

I-6.2. Intégration des trajectoires

Il n'existe pas de solution analytique exacte aux équations du mouvement[*]. Différents algorithmes ont été développés afin de les résoudre numériquement. Pour cela, une approximation usuelle consiste à diviser l'évolution du système en intervalles de temps (discrétisation temporelle). A l'issue de chacun de ces pas de temps, le potentiel de chaque particule est recalculé. L'erreur induite par cette approximation est négligeable quand les intervalles de temps utilisés sont suffisamment petits. Dans la pratique, le pas d'intégration Δt, adapté pour la simulation de systèmes biologiques, est de l'ordre de la femtoseconde (10^{-15}s).

Différents intégrateurs sont disponibles, chacun se caractérisant par un rapport spécifique entre précision et efficacité. L'algorithme de Verlet [110] est parmi les plus utilisés. Dans cet algorithme, les coordonnées sont développées en séries de Taylor au troisième ordre, aux temps $t + \Delta t$ et $t - \Delta t$:

[*] Leur résolution analytique peut être fastidieuse, voire impossible, en particulier si les mouvements des particules sont couplés (problème à N-corps).

$$\vec{r}_i(t+\Delta t) = \vec{r}_i(t) + \vec{v}_i(t)\Delta t + \frac{\vec{F}_i(t)}{m_i}\frac{\Delta t^2}{2!} + \frac{d\vec{F}_i(t)}{m_i dt}\frac{\Delta t^3}{3!} \qquad (15)$$

$$\vec{r}_i(t-\Delta t) = \vec{r}_i(t) - \vec{v}_i(t)\Delta t + \frac{\vec{F}_i(t)}{m_i}\frac{\Delta t^2}{2!} - \frac{d\vec{F}_i(t)}{m_i dt}\frac{\Delta t^3}{3!} \qquad (16)$$

La différence de ces deux séries permet d'obtenir les vitesses des atomes :

$$\vec{v}_i(t) = \frac{\vec{r}_i(t+\Delta t) - \vec{r}_i(t-\Delta t)}{2\Delta t} \qquad (17)$$

Des variantes optimisées de l'algorithme de Verlet existent. On peut citer par exemple : les algorithmes *leap frog* [111], de Beeman [112] et la plus utilisée, le *velocity Verlet* [113].

La modélisation d'un système macromoléculaire dans des conditions de solvatation explicite requiert la gestion d'un nombre de degrés de liberté tel, que souvent il s'avère nécessaire de réduire artificiellement ce nombre. Pour accélérer le calcul, on peut également avoir recours à l'algorithme SHAKE [114-116] qui permet d'augmenter modérément le pas d'intégration (p.ex. Δt = 2 fs au lieu de 1 fs) en corrigeant les oscillations trop importantes qui peuvent alors apparaître sur les degrés de liberté les plus "rapides" du système.

Pour améliorer l'efficacité de l'intégration, des intégrateurs à pas multiple, tels que l'algorithme r-RESPA (*reversible Reference System Propagator Algorithm*) [117-119] peuvent être employés afin de réduire substantiellement le temps de la simulation (d'environ un ordre de grandeur), sans perte de précision notable. Une telle optimisation diffère ici de la contrainte de type SHAKE indiquée précédemment car ce sont les degrés de liberté les plus lents qui sont périodiquement figés[*], tandis que le pas de calcul reste invariant.

[*] Les forces qui varient le plus lentement (p. ex. l'électrostatique à longue distance) sont recalculées moins fréquemment que les plus rapides (p. ex. interactions entre atomes liés).

I-7. Méthodes de description de l'environnement

I-7.1. Représentation du solvant

Les premières applications de simulation de protéines, il y a 30 ans, ont été réalisées dans le vide [120]. Bien que les résultats obtenus aient permis de donner un aperçu de la flexibilité de la macromolécule d'étude, ils ne pouvaient pas rendre compte précisément des propriétés dynamiques de celle-ci en milieu solvaté [121]. Il est pourtant primordial de tenir compte des effets de solvant lors de l'étude des biomolécules car ils jouent un rôle essentiel dans la stabilisation de ces dernières. Pour considérer les effets de solvant lors d'une simulation, deux approches sont possibles : la solvatation implicite ou la solvatation explicite.

Le traitement implicite du solvant repose sur une forme de potentiel qui traite le solvant comme un continuum diélectrique (p. ex. via le modèle de Born généralisé) [65]. Les divers modèles de solvatation implicite se révèlent utiles pour réduire de façon significative le temps de calcul des simulations. Toutefois, bien qu'ils reproduisent fidèlement l'effet global du solvant, ils ne peuvent pas, par définition, reproduire les interactions des molécules de solvant au niveau local (p. ex. à l'interface eau/biomolécule) [61].

Le traitement explicite, lui, correspond à la modélisation d'un nombre suffisant de molécules d'eau autour des systèmes biomoléculaires d'étude au sein d'une cellule qui constitue une boîte d'eau placée dans les conditions périodiques. Bien que ce traitement s'avère très coûteux en temps de calcul (car il augmente considérablement le nombre total d'atomes à considérer), il permet de modéliser de façon plus réaliste le milieu physiologique dans lequel les biomolécules évoluent. Dans le cadre de ce travail, les simulations ont été conduites dans les conditions de solvatation explicite.

I-7.2. Conditions périodiques aux limites

Une méthode particulièrement adaptée pour conduire une simulation dans les conditions de solvatation explicite, tout en réduisant les effets de bord, est celle des conditions de limites périodiques [107]. Cela consiste à répliquer implicitement l'ensemble fini de particules du système d'étude, réparties dans une boîte centrale (en général cubique ou parallélépipédique) selon les trois directions de l'espace.

Les atomes dans les cellules images reproduisent les mouvements des atomes correspondants dans la cellule centrale (voir figure ci-après). La simulation par DM s'effectue pour les atomes de la cellule centrale uniquement en tenant compte de la présence des cellules images lors du calcul du potentiel. Le caractère pseudo-infini du système ainsi simulé contraint à effectuer certaines approximations concernant le traitement des interactions entre molécules, en particulier celle dite de "l'image minimale" qui suppose que chaque particule *i* de la cellule centrale n'interagit qu'avec l'image la plus proche de toutes les autres particules *j*.

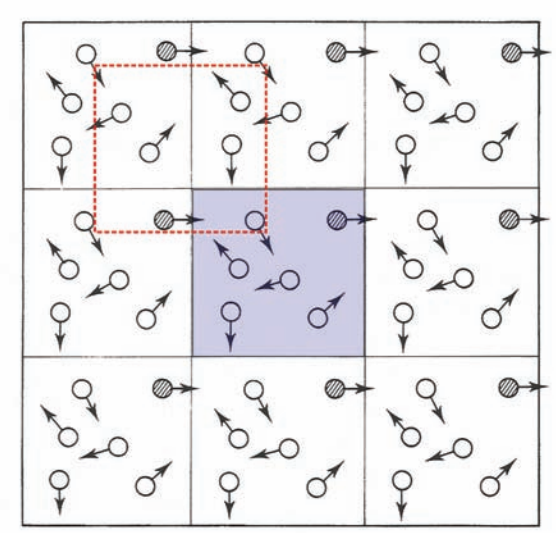


FIG. 13 - Conditions périodiques illustrées en 2 dimensions avec une boîte cubique. La boîte en bleu représente la cellule centrale répliquée ; la boîte en pointillés rouges symbolise l'image minimale.

I-7.3. Prise en compte des interactions entre atomes non-liés

La complexité des algorithmes de DM, en l'absence d'optimisation, est en O(n^2) ; cela limite rapidement l'intervalle d'espace et de temps accessible des simulations. En effet, alors qu'à un ensemble de particules liées ne correspond qu'un nombre limité d'interactions à prendre en compte dans le calcul (d'où une complexité en *n* du terme $V_{liés}$), le nombre de paires de particules non-liées croît, lui, en O(n^2). L'optimisation du calcul des interactions entre atomes non-liés est par conséquent centrale afin d'améliorer l'efficacité des algorithmes de DM.

Afin de limiter le temps de calcul nécessaire au calcul des potentiels entre les paires d'atomes non-liés s'exerçant à longue distance, il est courant de limiter la prise en compte de ces interactions pour un atome donné à ses voisins les plus proches, c'est-à-dire à ceux inclus dans une sphère dite de troncature ("*cut-off*"), centrée sur l'atome en question.

Dans le cas de l'utilisation des conditions périodiques aux limites, le rayon de troncature doit être inférieur ou égal à la moitié du plus petit côté de la cellule centrale, afin qu'il n'y ait pas plus d'une image de chaque atome prise en compte. Ainsi, la complexité du calcul n'est plus que de O(*n*) car le nombre total d'atomes au sein de la cellule centrale est borné. Mais, si l'emploi d'un cut-off de l'ordre de 8-10 Å est acceptable pour les interactions de van der Waals, ce n'est pas le cas de l'interaction Coulombienne en 1/r, même dans le cas de faibles charges atomiques partielles. Ainsi, le problème des interactions entre atomes non-liés se réduit principalement à celui du traitement des interactions électrostatiques [122].

I-7.4. Calcul des interactions électrostatiques

L'emploi d'un cut-off abrupt sur les interactions non-liées provoque des artefacts importants dans les calculs de DM [123]. Ceci peut être atténué par l'emploi d'un cut-off progressif, défini en réalité par deux seuils de cut-off, délimitant ainsi l'intervalle de prise en charge des interactions électrostatiques de façon décroissante de 100% à 0%. Dans les conditions périodiques, la technique de sommation d'Ewald [124] permet un calcul exact, mais au prix d'une complexité en $O(n^{3/2})$[*]. Pour pallier cela, une des solutions les plus utilisées de nos jours est la variante SMPE (*Smooth Particle Mesh Ewald*) [125] (complexité en O(*n*log(*n*))), qui est d'ailleurs implémentée dans le programme NAMD [89] utilisé pour conduire les simulations de ce travail.

I-8. Particularité des systèmes biomoléculaires par dynamique moléculaire

La DM est une technique de choix pour étudier la structure et la dynamique des macrobiomolécules au niveau atomique et aider ainsi à la compréhension de leur fonction biologique. Cependant, la résolution des équations du mouvement, nécessitant l'utilisation d'un pas de temps extrêmement petit, explique pourquoi le temps réel simulé dépasse rarement quelques dizaines voire centaines de nanosecondes (ns) pour des systèmes complexes de plusieurs dizaines de milliers d'atomes (p. ex. protéines dans solvant explicite ou insérées dans une membrane,…). Ainsi, l'étude de biomolécules relativement complexes, telles que les acides nucléiques ou les complexes de protéines par DM, est maintenant devenue possible.

[*] Cette technique introduit comme contrainte supplémentaire la neutralité électrique de l'ensemble du système. Si nécessaire, cette neutralité est obtenue artificiellement par l'emploi de contre-ions placés sur les bords de la cellule centrale.

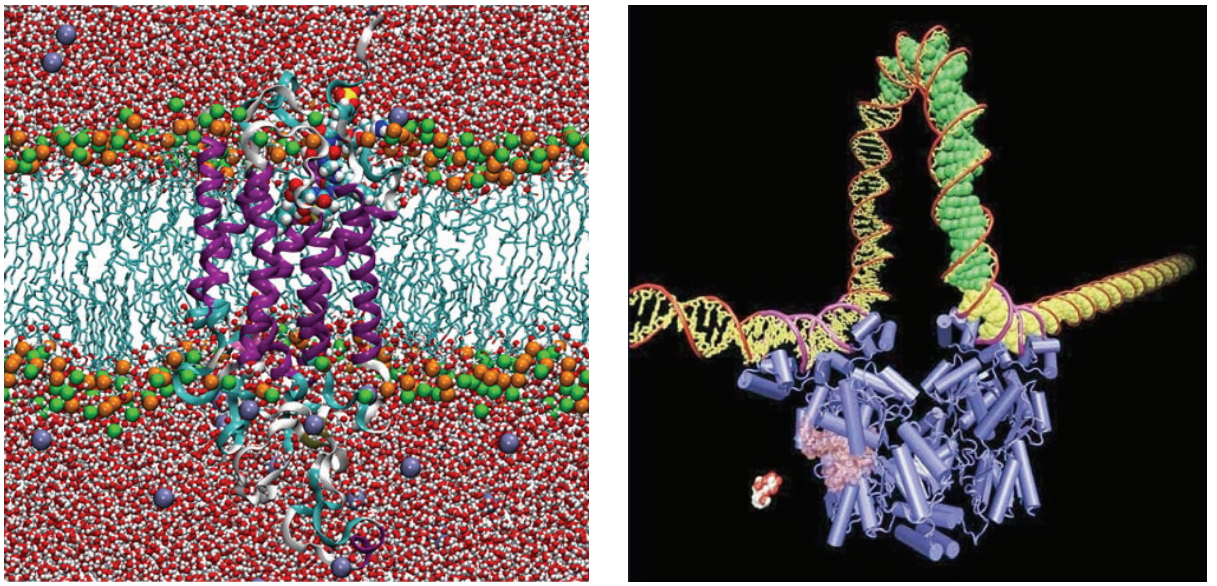


FIG. 14 – Exemples de systèmes biomoléculaires simulés par dynamique moléculaire. A gauche : récepteur humain de la cholécystokinine avec son ligand CCK9, au sein d'une bicouche lipidique hydratée [57]. *A droite : inhibition de l'expression du gène codant pour la production de lactose par la protéine lac repressor, liée à l'ADN* [126].

I-8.1. La dynamique moléculaire portée sur des architectures parallèles

Les ressources d'un seul ordinateur étant limitées pour conduire une simulation de DM dans un temps raisonnable, l'usage actuel et quasi-systématique consiste à recourir à la puissance de calcul de grappes d'ordinateurs, appelés *clusters*. Les calculs de DM se prêtant bien à la parallélisation [127], ils peuvent donc être déployés sur de telles architectures afin de répartir la charge de calcul sur les différents processeurs. Lors d'une simulation de DM, chaque pas de calcul dépend du résultat du précédent ; par conséquent, la parallélisation ne peut se faire sur le plan temporel. En revanche, cette manœuvre peut se faire sur le plan spatial car chaque particule du système a une évolution propre au cours d'un pas de calcul et ne dépend que d'un certain nombre d'atomes du voisinage suivant le terme du potentiel considéré. La cellule périodique est alors divisée en sous-ensembles homogènes pour les différents termes du potentiel. Les programmes de DM modernes les plus efficaces tels que NAMD (utilisé pour ce travail) possèdent des routines de parallélisation spécifiques et optimisées, permettant la *scalabilité parallèle*[*] du calcul [89] ; ainsi, la simulation de systèmes biologiques sur des intervalles de temps intéressants est facilitée.

[*] Efficacité du comportement du calcul lorsque le nombre de processeurs augmente pour un système simulé de taille donnée.

I-8.2. Vers la simulation de systèmes de plus en plus complexes

Le champ d'application de la DM va en s'élargissant au fur et à mesure que le rapport performance/prix des moyens de calcul intensif augmente. Les progrès techniques et méthodologiques des simulations de DM (p. ex. répartition de charge sur les serveurs de calcul, optimisation du calcul des forces électrostatiques), alliés à l'accès aux architectures massivement parallèles (p. ex. larges clusters ou grilles de calcul) [128], ouvrent la voie à l'investigation de systèmes biologiques de plus en plus complexes. En est l'illustration, l'étude de systèmes macromoléculaires d'intérêt biologique de tailles considérables sur des échelles de temps significatives. Parmi les dernières prouesses récemment publiées, on peut citer celle du virus satellite de la mosaïque du tabac dans son intégralité [59] (1 million d'atomes sur 50 ns), ou encore celle du ribosome [129] (2,64 millions d'atomes sur 20 ns). Enfin, si d'autres barrières symboliques ont été franchies, comme passer le cap de la microseconde pour la simulation d'une protéine en solution [130], la cellule, élément central de la biologie, est encore un objet trop complexe pour être simulé de façon directe [131].

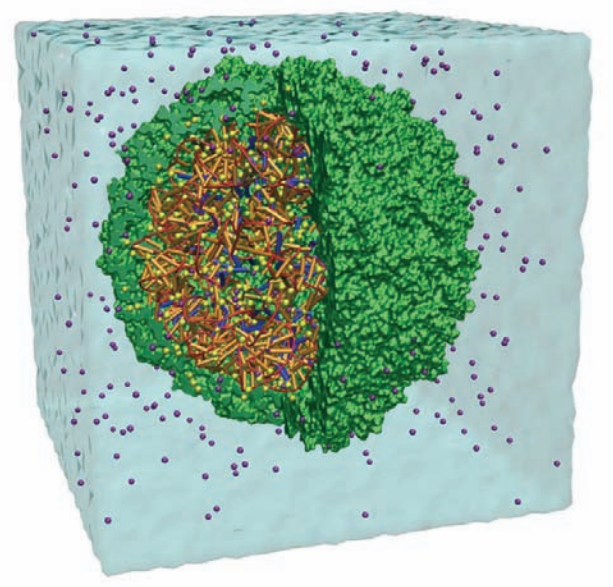


FIG. 15 - Représentation schématique du virus satellite de la mosaïque du tabac complet (capside + ARN) simulé en milieu aqueux explicite (le système contient 1 million d'atomes au total). [59]

II- LES TECHNIQUES INFORMATIQUES DE LA RECHERCHE DE MÉDICAMENTS

La conception de molécules d'intérêt thérapeutique a bénéficié ces dernières décennies des progrès issus de diverses disciplines scientifiques telles que la biologie, la pharmacochimie et l'informatique. Ainsi la recherche, qui consistait autrefois à synthétiser et tester les composés sélectionnés sur la base de l'intuition et de l'expérience du chimiste médicinal, a radicalement évolué. L'essor de l'outil informatique a particulièrement changé la donne, en conduisant à l'émergence d'une nouvelle discipline pouvant participer aux étapes initiales de la recherche pharmaceutique en complément des méthodes expérimentales déjà reconnues. On parle alors de conception de médicaments *in silico* – c'est-à-dire assistée par ordinateur – qui correspond à un ensemble de techniques informatiques spécifiques souvent désigné par l'acronyme CADD [48, 49] (pour "*Computer-Aided Drug Design"*).
Bien que ces outils aient un large champ d'application dans le processus de recherche de nouveaux médicaments (cf. FIG. 8 – partie Introduction), nous nous limiterons à la description des méthodes utilisées pour ce travail, à savoir le criblage virtuel [41, 42] et, plus particulièrement, une de ses sous-composantes : l'arrimage moléculaire, ou "*docking"* [50, 132].

II-1. Le principe du criblage virtuel

II-1.1. Utilisation dans le cadre de la recherche de nouveaux médicaments

Le criblage virtuel est la stratégie *in silico* la plus utilisée pour l'identification de *touches* ("*hits"*) dans le cadre de la recherche de nouveaux médicaments [42, 133, 134]. Celui-ci fait désormais partie intégrante de la plupart des programmes de recherche de composés bioactifs, que ceux-ci se déroulent en milieu académique ou industriel, car il constitue un complément essentiel au criblage biologique haut-débit [43, 135, 136].

Le criblage virtuel permet l'exploration de larges chimiothèques (> 10^6 molécules) à la recherche de composés actifs vis-à-vis d'une cible thérapeutique donnée. Ce processus vise à réduire de façon significative la chimiothèque de départ à une liste limitée de composés jugés les plus prometteurs. Cette approche conduit souvent à une nette amélioration de la "concentration" de molécules actives pour la cible ("*hit rate"*), tandis qu'une sélection aléatoire de molécules de la chimiothèque ne saurait fournir un tel *enrichissement*. Ainsi, le temps aussi bien que les coûts de l'identification de nouvelles

touches peuvent être réduits de façon significative [43]. Plus précisément, le recours au criblage *in silico,* en préalable à un criblage biologique à plus petite échelle, permet d'ajuster au mieux le nombre de tests expérimentaux en fonction des contraintes budgétaires et temporelles. Quand les conditions le permettent, le criblage biologique peut être employé en parallèle au criblage virtuel, afin d'évaluer l'efficacité de ce dernier et de pouvoir améliorer les paramètres des programmes informatiques utilisés [137].

La pertinence de la chimiothèque employée est la première condition pour le succès d'un criblage virtuel, bien avant celle des algorithmes utilisés pour la recherche de touches au sein de la chimiothèque [40]. En effet, seule une librairie de composés suffisamment diverse [138, 139] peut garantir une exploration satisfaisante de l'espace chimique [44, 45], maximisant ainsi les chances de découvrir de nouvelles touches.
Par ailleurs, pour éviter de perdre du temps avec des molécules possédant des caractéristiques incompatibles avec celles de composés d'intérêt pharmaceutique, le processus de criblage comporte généralement une étape préliminaire de filtrage. Cette tâche, qui peut être prise en charge par des programmes spécialisés [140], consiste à exclure les composés toxiques ou supposés tels [141, 142] ou comportant des groupements jugés trop réactifs [143, 144]. Ensuite, ne sont retenus que les composés obéissant à des définitions empiriques simples du profil de molécule active (caractère dit "*drug-like"*), telle que la populaire ”règle des 5” de Lipinski [46] et ses extensions [145-147].
Durant leur phase d'optimisation, la complexité, le poids moléculaire et le caractère lipophile des touches ont généralement tendance à augmenter [148]. Par conséquent, des filtres physico-chimiques additionnels peuvent être employés pour restreindre la chimiothèque aux composés ayant un profil de tête de série, dits "*lead-like"* [149]. On estime alors que les risques d'échec aux tests cliniques sont minimisés, ce qui est souvent un impératif économique. Toutefois, appliquées de façon stricte, ces règles de pré-filtrage peuvent assez facilement conduire à l'exclusion de molécules d'intérêt. Par exemple, de nombreux médicaments commercialisés ne respectent pas les règles de Lipinski [144] ; c'est pourquoi ces règles empiriques sont souvent employées de façon plus ou moins permissive.

II-1.2. Les différentes stratégies du criblage virtuel

Suivant la nature de l'information expérimentale disponible, on distingue deux approches distinctes pour le criblage virtuel. La première se base sur la structure de la cible et est connue sous le nom de "*structure-based virtual screening"* [55, 150]. La seconde, reposant sur la connaissance d'un nombre suffisant d'informations concernant une ou plusieurs molécules actives de référence, est appelée "*ligand-based virtual screening"* [43, 151]. Bien que ces deux approches soient surtout utilisées de

manière exclusive (souvent parce que la nature des données de départ ne laisse qu'un seul choix possible), leur combinaison lors d'une campagne de criblage permet de maximiser les chances de succès pour identifier de nouvelles touches [152].

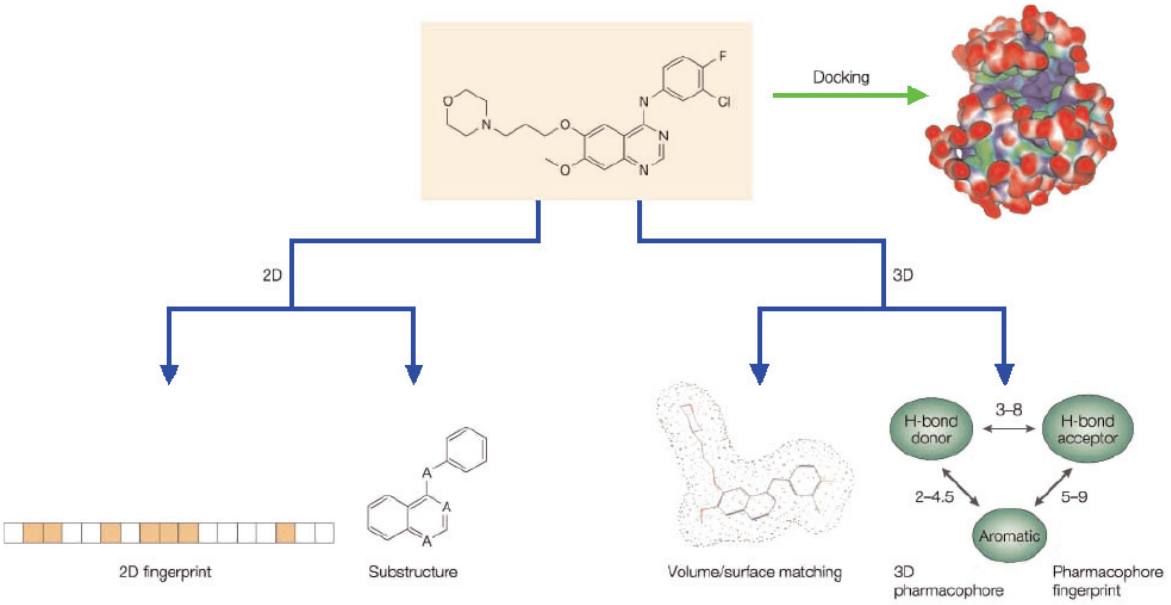


FIG. 16 - Illustration des deux approches classiques du criblage virtuel. Modifié de (Bajorath et al., 2002) [43].
Flèche verte : criblage basé sur la structure de la cible (docking).
Flèches bleues : criblage basé sur des règles empiriques établies à partir d'informations sur des ligands connus, exprimées avec des descripteurs topologiques ou structuraux.

Les méthodes "*ligand-based*" consistent en premier lieu à classifier la chimiothèque testée à travers divers descripteurs (motifs pharmacophoriques [153] et autres [154]) décrivant les caractéristiques physico-chimiques et structurales des molécules. Ensuite, il s'agit d'identifier les molécules comportant une similarité avec les ligands connus pour être actifs sur la cible étudiée, ainsi qu'une dissimilarité avec les molécules connues pour être inactives (ou présentant des caractéristiques indésirables). Cette approche, rapide et relativement simple à mettre en œuvre, présente comme inconvénient majeur son interdépendance envers les informations de référence utilisées pour construire le modèle de prédiction d'affinité. Par conséquent, les résultats ne sont souvent pas satisfaisants en termes de diversité. En particulier, il est *a priori* impossible d'identifier de nouvelles classes de molécules actives dont les propriétés diffèreraient de celles des ligands actifs connus.

De son côté, l'approche "*structure-based*" se rapporte souvent aux algorithmes de docking protéine-ligand et quelquefois aux recherches basées sur un motif pharmacophorique du site actif de la protéine cible. Ces méthodes ne consistent pas à rechercher des touches comportant une similarité avec des composés connus, mais à estimer la complémentarité structurale de chaque molécule criblée avec le site actif considéré. Ainsi, contrairement à l'approche basée sur des ligands de référence, cette

approche peut potentiellement identifier de nouvelles classes de molécules actives. En revanche, ces méthodes sont généralement plus coûteuses en puissance de calcul et leur emploi requiert souvent une expertise plus importante.

La figure ci-dessous présente le scénario classique d'un criblage virtuel par docking.

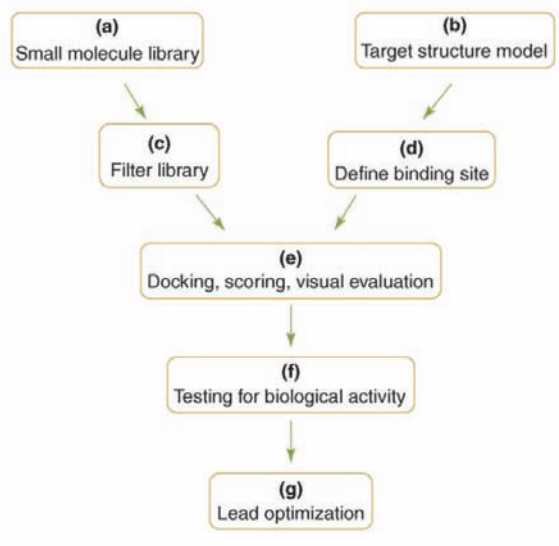


FIG. 17 - Déroulement d'une expérience de criblage virtuel par docking. D'après (Ghosh et al., 2006) [55].

Dans un premier temps, la chimiothèque à cribler *(a)* est préparée[*], puis pré-filtrée *(c)* comme décrit précédemment. Concernant la cible, il faut disposer d'une structure tridimensionnelle fiable de la protéine ciblée (en particulier au niveau de son site actif) *(b)*, qui peut être issue de techniques expérimentales telles que la diffraction aux rayons-X ou par RMN ou encore provenir d'un modèle par homologie validé [56]. Une fois la cible sélectionnée et préparée[†], le site actif de celle-ci est défini ; cela consiste à déterminer les résidus clés du site de liaison qui forment le lieu de reconnaissance moléculaire avec les ligands *(d)*.

Quand les petites molécules et la cible sont prêtes, un algorithme de docking, associé à sa fonction de score, est employé afin d'identifier les composés de la chimiothèque ayant les meilleures affinités avec la cible *(e)*. Une inspection visuelle des solutions de docking des ligands les mieux classés par la fonction de score peut avoir lieu, afin de vérifier leurs modes d'interaction avec les résidus clés du site actif *(e)*. Les molécules candidates obtenant les meilleures prédictions d'affinité constituent enfin des touches potentielles qu'il faut valider en les soumettant aux tests expérimentaux *(f)*. A l'issue des ces

[*] La préparation d'une chimiothèque concerne par exemple l'uniformisation des données, l'ajout des protons à pH physiologique, la conversion en 3D, *etc.* L'accessibilité expérimentale des molécules (disponibilité immédiate chez un fournisseur, facilité de synthèse, *etc.*) pouvant s'avérer cruciale en cas d'identification de composés prometteurs, il est alors nécessaire de la garantir à ce stade.

[†] La préparation d'une cible comprend, entre autres, la vérification de la structure, sa protonation, la prise en compte de certaines molécules d'eau, *etc.*

essais, les composés démontrant une activité biologique peuvent passer en phase d'optimisation en vue de déterminer des têtes de série *(g)*.

La prochaine section de ce mémoire développe spécifiquement le docking car celui-ci constitue l'outil le plus utilisé pour nos travaux de criblage virtuel.

II-2. Le docking

Mon exposé se limitera ici à une description assez générale des techniques de docking afin de situer le contexte de ce travail. Cette section n'a donc pas vocation de s'ajouter aux nombreuses revues publiées sur les caractéristiques détaillées des différents programmes de docking ou sur l'évaluation de ceux-ci [132, 155-165].

II-2.1. Principe

Le docking est utilisé pour prédire la structure du complexe intermoléculaire résultant de l'association entre au moins deux molécules. Quand il s'agit de deux protéines, on parle de docking protéine-protéine [166], par opposition au docking protéine-ligand que nous avons utilisé pour ce travail.

Le processus du docking est itératif et chaque passe de calcul s'articule en deux étapes. Tout d'abord, une portion limitée de l'espace des conformations du complexe protéine-ligand est explorée afin d'améliorer la *pose* (orientation et/ou conformation) du ligand dans le site actif de la protéine. L'étape suivante fait intervenir une *fonction de score* qui évalue la qualité de la pose générée [167]. Cette estimation *in silico* de l'affinité du ligand pour la cible est basée sur un examen simplifié des interactions entre les deux partenaires. La répétition du cycle de recherche associé à l'estimation du score guidant l'exploration de l'espace doit assurer la convergence de l'algorithme vers un état représentatif du minimum global de l'énergie libre d'association.

Avant de décrire plus en détail les deux composantes du principe de docking protéine-ligand évoquées ici, nous présenterons les principaux types de représentation du système employés par les diverses techniques de docking. Nous verrons ensuite en quoi la nature de l'espace de recherche permet de classer les algorithmes de docking en plusieurs familles distinctes. Enfin, nous discuterons de la signification des résultats obtenus dans le contexte du criblage virtuel et des perspectives générales du docking.

II-2.2. Représentation du système

La plupart des méthodes de docking reposent sur une représentation simplifiée qui peut réduire la protéine aux atomes du site actif et du voisinage de celui-ci et dans laquelle les molécules du solvant ne sont pas représentées explicitement. Les modèles utilisés sont généralement analogues à ceux de la mécanique moléculaire (cf. Partie 2, Section I), dans lesquels chaque atome des structures moléculaires est représenté comme un point unique sur lequel ses propriétés sont projetées. La majorité des méthodes de docking utilise des types atomiques simplifiés par rapport aux définitions d'un champ de force courant. Dans certains cas, des pseudo-atomes supplémentaires sont rajoutés afin de simplifier le traitement de tous les types d'interaction entre protéine et ligand, par exemple les paires électroniques libres afin de mieux détecter la formation d'éventuelles liaisons hydrogène.

Moins fréquemment, les ligands et/ou les sites actifs peuvent être modélisés par l'intermédiaire de surfaces. Différents types de représentation peuvent être envisagés, par exemple un ensemble de points dans l'espace ou bien une série de fonctions harmoniques sphériques [168, 169].

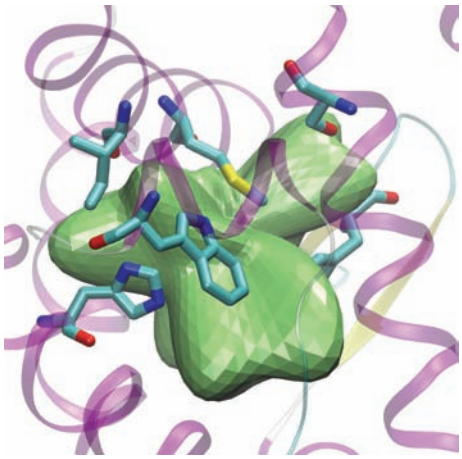


FIG. 18 – Surface du site actif d'une protéine dont la représentation est basée sur les harmoniques sphériques.

Les propriétés des ligands et/ou des sites actifs réparties dans l'espace peuvent également être représentées à travers leur projection sur une grille. Ce type de représentation est souvent employé quand un ensemble de conformations de la protéine cible (pouvant être obtenu expérimentalement ou par simulation de dynamique moléculaire par exemple) est utilisé pour considérer sa flexibilité sans l'implémenter explicitement au niveau de l'algorithme de recherche. Ces conformations peuvent être combinées en une seule pseudo-conformation de diverses façons (moyennées, unifiées, *etc.*). Les programmes DOCK [170], AutoDock [171] et FlexE [172] permettent l'emploi de ce type d'approche.

Il est aussi possible d'effectuer un docking sur chacune des multiples conformations [173, 174]. Cette approche systématique peut s'avérer coûteuse mais a l'avantage de ne pas dépendre des paramètres et

de la pertinence d'un programme spécifique. Cette stratégie a été adoptée pour les travaux de criblage utilisant la plateforme VSM-G, présentés dans ce mémoire.

II-2.3. La recherche de la pose optimale

Un système biologique réduit à la zone d'interaction protéine-ligand fait intervenir trois acteurs dans le phénomène de reconnaissance moléculaire : la protéine, le ligand et le solvant. Un algorithme de docking idéal prendrait en compte tous les paramètres associés à ces trois composantes, tels que la complète flexibilité du ligand et de la protéine, les effets de solvant, *etc.* Toutefois, une telle approche implique l'exploration d'un espace composé d'un nombre très important de degrés de liberté. La résolution du docking par une recherche exhaustive nécessiterait alors des temps de calcul considérables. Pour cette raison, la complexité du système est généralement réduite au niveau de la modélisation structurale. Les approximations les plus courantes à ce niveau consistent à considérer la protéine comme une entité rigide et à ne pas représenter explicitement les molécules du solvant.

On peut distinguer plusieurs niveaux de représentation du système dont les approximations évoluent conjointement avec le progrès des moyens informatiques au fil des ans. Au début des années 1980, quand l'approche de modélisation moléculaire par docking a été pour la première fois étudiée [175], Kuntz et ses collaborateurs ont proposé la stratégie du tout rigide dans laquelle l'exploration de l'espace des positions (discrétisé par l'emploi d'une grille) se limite au positionnement du ligand et n'exploite donc que six degrés de liberté élémentaires (rotations et translations). Depuis, l'essor des moyens de calcul a permis de considérer le système d'étude comme semi-flexible : la protéine est traitée de façon rigide, contrairement au ligand dont l'espace conformationnel est pris en compte. Cette approche reste la plus employée aujourd'hui, les algorithmes tenant compte explicitement de l'espace conformationnel du site actif étant encore en cours de développement.

Les différentes approches du docking que nous allons à présent détailler se distinguent au niveau de leurs conditions d'application et de la nature des informations qu'elles peuvent fournir. La pertinence du choix d'un programme de docking donné repose en premier lieu sur l'adéquation entre ces caractéristiques et celles du système étudié. L'efficacité de l'algorithme choisi sera par ailleurs un compromis entre rapidité d'exécution et précision des résultats.

Docking rigide

Dans le cas des méthodes de docking rigide, la recherche de la pose optimale se limite au positionnement. Cette opération consiste en la recherche exhaustive dans l'espace discrétisé des 6 degrés de liberté. Certains programmes, s'ils n'appartiennent pas à la famille des techniques de docking rigide, utilisent plusieurs étapes successives d'optimisation dont les premières peuvent s'apparenter à du docking rigide. Par exemple, le programme Glide [176] utilise initialement, dans son approche multi-étapes, une recherche systématique pour positionner le ligand de façon approchée au sein du site actif de la protéine.

Il est possible de considérer indirectement la flexibilité des ligands en utilisant des programmes de docking rigide. Pour cela, un jeu de conformères de basse énergie pour chaque molécule à tester sur le site actif ciblé peut être généré efficacement par les programmes tels qu'OMEGA [177] ou Catalyst [178]. Ces données sont calculées une fois pour toutes, réutilisables pour d'autres criblages une fois stockées. Certains programmes tels que FLOG [179], FRED [177] et EUDOC [180] travaillent sur un ensemble de conformères pré-calculés par un programme tiers ou générés à la volée par leurs algorithmes. Une telle approche est nécessaire dans le contexte du docking rigide car il est admis que la conformation complexée d'un ligand peut varier considérablement par rapport à sa conformation "libre" [181].

Docking semi-flexible

Lorsque l'espace conformationnel des ligands est exploré, le nombre de degrés de liberté de l'espace de recherche peut être conséquent dans le cas de molécules très flexibles. Dans un tel contexte, l'emploi de méthodes de recherche exhaustives apparaît souvent inapproprié car nécessitant des simplifications importantes au niveau de l'échantillonnage. D'autres algorithmes, dits de fragmentation, sont employés pour construire de façon incrémentielle le ligand au sein du site actif de la protéine. L'espace des conformations du ligand est alors restreint au voisinage d'un ensemble initial d'états simplifiés. Cette stratégie de recherche par construction, qui se présente sous diverses variantes [132], est notamment adoptée par les programmes DOCK [175, 182], FLExX [183] et Hammerhead [184].

Les programmes de docking semi-flexible considérés comme les plus efficaces emploient des méthodes de recherche aléatoires ou stochastiques. L'exploration de l'espace de recherche se fait de façon plus ou moins aléatoire et les états générés sont soit acceptés, soit rejetés, suivant des règles spécifiques. On distingue trois principales classes de méthodes aléatoires : les méthodes de Monte

Carlo (cf. Partie 2, Section I), les méthodes évolutionnaires basées sur des algorithmes génétiques [185] et les méthodes de recherche Tabou [186].

Les algorithmes basés sur la méthode de Monte Carlo génèrent aléatoirement des états du système acceptés ou rejetés sur la base probabiliste de la fonction de Boltzmann. Les principaux programmes utilisant cette méthode sont ICM [187], QXP [188] et MCDOCK [189].

Les algorithmes génétiques [185] s'inspirent des théories de l'évolution pour sélectionner les états. Une population initiale aléatoire de confirmations du ligand dans le site actif de la protéine est définie et les degrés de liberté à explorer sont assimilés à un jeu de *gènes*. L'échantillonnage de l'espace se fait ensuite par des opérations génétiques (mutations, croisements et migrations) sur la population. La sélection des individus-conformères générés est basée sur leur capacité d'adaptation à l'environnement (la fonction de score). Malgré leur efficacité algorithmique, les algorithmes génétiques appliqués au docking ont parfois tendance à sélectionner des minima locaux. Pour pallier cela, une solution consiste à répéter un même calcul plusieurs fois afin de maximiser les chances d'obtenir, à l'issue de cette procédure, au moins un résultat satisfaisant (structure proche du minimum global). Les programmes de docking les plus connus implémentant un algorithme génétique sont AutoDock [190], GOLD [191] et DARWIN [192].

Le principe de base des méthodes de recherche Tabou, utilisées par exemple dans PRO_LEADS [193], est de pouvoir prendre en compte les régions de l'espace ayant déjà été visitées (par des calculs de similarité entre ligands, p. ex. calcul de RMSD). La recherche dans les régions inexplorées est privilégiée, réduisant considérablement la taille de l'espace à explorer.

Docking flexible

De nombreuses études utilisant le docking par l'approche ligand flexible/protéine rigide ont montré que cette stratégie semi-flexible conduisait à des résultats concluants [54, 55]. Toutefois, les réussites issues de son utilisation concernent surtout l'étude de protéines relativement rigides. À la suite de leur liaison avec un ligand, de nombreuses protéines peuvent pourtant subir des réarrangements structuraux de plus ou moins grande amplitude. Négliger cet aspect, désigné sous le terme d'"*induced fit*" [194], peut avoir des conséquences fâcheuses sur la pertinence des résultats issus d'un docking [195]. Même concernant des changements conformationnels mineurs, la flexibilité du site actif de la protéine peut avoir une grande influence dans le phénomène de reconnaissance moléculaire avec un ligand [196].
La mise au point d'algorithmes de docking prenant en compte explicitement la flexibilité de la protéine est toutefois une tâche délicate car le nombre de degrés de liberté associé à une telle représentation peut s'avérer très important (>> 50). Les programmes incorporant, au moins

partiellement, la flexibilité du site actif adoptent des stratégies assez diversifiées dont certaines d'entre elles sont évoquées ici à titre illustratif.

Le docking rigide permissif (dit "*soft docking"*) considère la flexibilité de la protéine de manière indirecte, en atténuant certains termes de répulsion. Cela peut permettre au ligand de pénétrer légèrement la surface de la protéine en prévision des réarrangements qui auraient eu lieu lors de l'association de partenaires flexibles [197]. Une telle approche indirecte de la flexibilité de la protéine est cependant limitée pour couvrir le spectre des effets d'*induced fit* qui peuvent être observés, par exemple par dynamique moléculaire (cf. article #3 de ce travail).

Une librairie de rotamères pour un ensemble de chaînes latérales de la protéine autorisées à être flexibles peut être utilisée dans une approche plus explicite [198, 199]. ROSETTALIGAND [200] et Glide [176] sont deux exemples de programmes utilisant cette stratégie.

Enfin, de nouveaux programmes de docking qui traitent la flexibilité du site actif en employant la même stratégie que celle des ligands (chaque rotamère constitue un degré de liberté), sont actuellement en développement [201, 202]. Certains reposent sur des algorithmes innovants tels que la simulation de colonies de fourmis ou l'optimisation par essaims de particules [203].

II-2.4. L'évaluation du score des poses

L'évaluation et le classement des différentes conformations du ligand positionné dans le site actif de la protéine constituent le second aspect crucial du docking, après la nature de la recherche conformationnelle. Une fonction de score doit théoriquement pouvoir distinguer, parmi les différentes poses générées pour un ligand, celles qui correspondent aux modes de liaison les plus représentatifs de la reconnaissance moléculaire. Ainsi, elle doit permettre, sur la base d'une pose optimale proposée, de différencier les molécules bioactives des autres composés inactifs vis-à-vis de la protéine.

Des techniques basées sur le calcul d'énergie libre ont été développées et constituent un moyen quantitatif rigoureux pour estimer l'affinité de liaison d'un ligand pour une protéine [204, 205]. Bien qu'elles soient très précises, leur coût (en temps, en contraintes et en expertise) ne les rend pas appropriées pour une application de docking dans le cadre d'un criblage virtuel. Pour représenter le meilleur compromis entre vitesse et précision, les fonctions de score implémentées dans les programmes de docking sont donc basées sur la simplification des phénomènes impliqués dans la

reconnaissance moléculaire, en particulier de ceux qui sont délicats à évaluer en dehors des calculs d'énergie libre (p. ex. l'entropie).

Les différentes fonctions de score implémentées dans les programmes de docking ont fait l'objet de nombreuses publications [167, 206-208]. Elles sont généralement classées suivant trois catégories : les fonctions de score basées sur un champ de force, les fonctions de score empiriques et celles basées sur des connaissances statistiques. Enfin, plusieurs fonctions de score de nature différente peuvent être combinées pour former des fonctions de score dites de *consensus*.

Les fonctions de score basées sur un champ de force

Les champs de force dans leur forme standard (cf. Partie 2, I. Mécanique moléculaire) évaluent la somme de deux énergies : l'énergie entre atomes liés au sein d'une molécule donnée (énergie interne) et l'énergie entre atomes non-liés. Ce deuxième terme correspond au terme principal de l'énergie d'interaction protéine-ligand dans le cas d'une application à un problème de docking. La plupart du temps, les fonctions de score basées sur un champ de force ne considèrent qu'une conformation donnée de la protéine. Ainsi, si l'on compare l'activité de deux ligands par rapport à cette même conformation, cela permet de faire abstraction du terme d'énergie interne de la protéine qui s'annule dans l'expression de la différence d'énergie libre d'interaction entre les deux ligands.

L'affinité d'un ligand donné pour le site actif, mesuré en tant qu'énergie d'interaction, correspond à la somme des énergies d'interaction de van der Waals (souvent représentée par un potentiel de Lennard-Jones) et électrostatiques (potentiel de Coulomb). On peut ajouter à la fonction de score un terme d'énergie interne du ligand. Tous ces termes peuvent s'exprimer à partir des paramètres du champ de force considéré.

De telles fonctions de score présentent certaines limitations qui s'ajoutent à celles induites par la représentation du système (p. ex. la non représentation explicite du solvant). En particulier, les effets d'entropie, qui peuvent varier d'un ligand à l'autre pour un site actif donné, aussi bien que pour un ligand donné d'un site actif à l'autre, ne sont pas pris en compte. Seule la contribution enthalpique de l'énergie libre d'interaction est ainsi prise en compte.

Les fonctions G-Score [209] (basée sur le champ de force de Tripos [209]) et celle implémentée dans AutoDock [210] (basée sur le champ de force AMBER [68]) sont des exemples de ce type de fonction de score.

Les fonctions de score empiriques

Ce type de fonction de score approxime l'énergie libre de liaison en sommant de façon pondérée différents termes d'interaction dérivés de paramètres structuraux. Les différents poids de la fonction de score sont ajustés pour reproduire en priorité des données expérimentales, telles que les constantes de liaison tirées d'un jeu d'entraînement de complexes protéine-ligand.

La plupart des programmes de docking implémentent ce type de fonction de score témoignant de leur efficacité (en terme de rapport précision/rapidité). Cependant, le principal inconvénient de ces fonctions empiriques est leur forte dépendance aux données utilisées pour les calibrer qui, en cas de mauvaise paramétrisation, peut limiter leur transférabilité sur des systèmes différents. Parmi les principales fonctions de score empiriques, on peut citer : ChemScore [211], PLP [212], et LigScore [213].

Les fonctions de score basées sur des connaissances statistiques

Ces fonctions de score sont construites à partir de règles fondées sur une analyse statistique des complexes protéine-ligand résolus expérimentalement. Elles partent du principe que les distances interatomiques les plus représentées statistiquement dans les complexes constituent des contacts énergétiques favorables et, qu'à l'inverse, les plus rares représentent des interactions moins stables. Ainsi, leur paramétrisation dépend de la quantité d'informations expérimentales disponibles et on doit leur apparition à la profusion de données structurales accessibles dans des bases de données telles que la *Protein Data Bank*. Les exemples populaires de ces fonctions de scores sont PMF [214], SMoG [215] et DrugScore [216].

Les fonctions consensus

Ces fonctions hybrides combinent les résultats issus de diverses fonctions de score. On estime qu'il est possible de compenser partiellement les faiblesses intrinsèques de chacune des fonctions de score employées, évitant leurs erreurs individuelles et ainsi d'augmenter la probabilité d'identifier des composés actifs [217]. Cependant, si les termes des différentes fonctions de score sont fortement corrélés, l'intérêt du consensus devient limité car il peut entraîner une amplification des erreurs, au lieu de les atténuer. Ces fonctions consensus ont récemment fait l'objet d'une revue [218].

Autres types de fonctions de score

Les méthodes de docking reposant sur des surfaces utilisent des fonctions de score qui sont adaptées et principalement centrées sur la complémentarité géométrique surface-surface et auxquelles il peut être adjoint une estimation d'interactions sur le modèle des fonctions de score plus conventionnelles. Par exemple, LigandFit [219] génère, par Monte Carlo, les conformations du ligand dont les formes sont ensuite comparées à celle du site actif. D'autres programmes, tels que FRED [177, 220] ou SHEF [221], comparent la forme de chacun des conformères, générés au préalable (p. ex. avec OMEGA [177]), à la forme du site actif de la protéine. Cette approche est celle utilisée dans la première étape de la stratégie de la plateforme de criblage VSM-G décrite dans ce travail.

II-2.5. Pertinence des résultats de docking et perspectives

La façon la plus simple et la plus courante d'exploiter les résultats d'un criblage virtuel par docking consiste à se limiter au classement des molécules candidates suivant la valeur de la fonction de score pour la structure optimale générée. Dans ce contexte, une fonction de score est considérée comme efficace si elle est capable de discriminer les composés actifs des non-actifs vis-à-vis de la cible et, mieux, de classer les molécules par ordre d'affinité. On considère alors qu'afin d'améliorer la pertinence des programmes de docking, il convient en priorité d'améliorer celle des fonctions de score, sans entraîner un surcoût de calcul trop important.

Une autre approche plus qualitative consiste à considérer que le but du docking est la production d'une structure géométrique pertinente du complexe protéine-ligand, laquelle peut servir de base pour des calculs ultérieurs d'optimisation. Dans ce cas, la corrélation du score avec l'énergie libre d'interaction pourrait s'avérer secondaire, du moment que la fonction de score permette à l'algorithme de recherche de converger avec une forte probabilité vers une structure pertinente. Dans une optique de criblage, il s'avérera alors nécessaire de mettre en œuvre une étape d'optimisation *post-docking* (p. ex. via un minimiseur basé sur un potentiel MM/PBSA [222]) afin de discriminer les molécules candidates de façon automatisée (et éventuellement fournir une valeur de score mieux corrélée à l'énergie libre d'interaction).

Il convient aussi de tenir compte des interactions entre fonction de score et recherche conformationnelle au sein du processus de docking. Une amélioration du score doit en effet être évitée si elle est susceptible, par exemple, de faire converger trop vite la recherche conformationnelle vers un

minimum local non pertinent du point de vue de l'interaction protéine-ligand. De telles considérations peuvent rendre le développement de techniques de docking efficaces particulièrement délicat.

Même si du point de vue de la précision certains programmes actuels se distinguent [162], il peut être difficile de déterminer dans l'absolu qu'un programme de docking donné "fonctionne mieux" qu'un autre ou de tirer des conclusions générales sur leur pouvoir prédictif ; ceci pour plusieurs raisons [223, 224]. En effet, chaque programme de docking est unique par la représentation du système, la méthode de recherche et la fonction de score employée. Cette spécificité se répercute sur sa pertinence en termes de rapport entre précision et efficacité de calcul, rapport qui dépend aussi des paramètres employés pour les calculs et de la nature de la cible étudiée. C'est pourquoi, on assiste par exemple à l'émergence de fonctions de score spécifiques à certaines classes de protéines [225]. Enfin, le jugement de la qualité d'un programme peut se faire à plusieurs niveaux, notamment sur la qualité des poses des ligands les mieux classés, sur sa prédiction de l'affinité de liaison ou encore sur son efficacité à cribler des chimiothèques.

De façon plus générale, le choix d'un programme donné pour un criblage virtuel par docking doit répondre à la question suivante : « Que veut-on obtenir, en combien de temps, et avec quelle précision ? ». Il semblerait qu'une approche combinant plusieurs programmes s'avèrera souvent plus pertinente que l'emploi d'un seul programme, particulièrement dans le cadre d'un criblage à haut débit [40, 41]. Cette remarque a été à la base du choix de développer l'approche de criblage multi-étapes implémentée dans la plateforme VSM-G.

L'utilisation des programmes de docking a conduit à de nombreux succès dans le domaine de la découverte de nouvelles molécules bioactives [54, 55] ; néanmoins, leurs algorithmes sont toujours en maturation. Les axes principaux de leur amélioration méthodologique visent essentiellement à considérer la totale flexibilité de la protéine pendant le processus de docking et à prendre en compte les effets d'entropie et de solvant dans l'évaluation de l'affinité du complexe protéine-ligand par la fonction de score. Certains travaux concernant le développement de programmes de docking innovants [201, 226, 227] sont la preuve que les limitations actuelles des algorithmes de docking et des fonctions de score pourraient être surmontées dans un avenir relativement proche. Le potentiel actuel et futur des techniques de docking est incontestablement source de motivation.

PARTIE 3 - RÉSULTATS

I. DÉVELOPPEMENT ET VALIDATION DE LA PLATEFORME DE CRIBLAGE VIRTUEL À HAUT-DÉBIT VSM-G

Cette section expose les principaux résultats se rapportant au projet VSM-G. À la suite d'une présentation succincte de la plateforme, seront décrits nos travaux en relation avec le développement, la validation et l'usage de VSM-G dans des études de criblage virtuel.

I-1. Motivations

Le conception de nouveaux médicaments est un processus long et très coûteux [47]. L'identification de nouveaux composés bioactifs par criblage expérimental sur une cible thérapeutique donnée constitue souvent le point de départ de ce processus. Or son recours systématique, sans *a priori*, n'est plus toujours pertinent (principalement pour des raisons financières), en particulier appliqué à l'intégralité de librairies chimiques pouvant désormais contenir plusieurs millions de composés [137]. Dans ce contexte et en lien avec l'essor des moyens informatiques, le criblage virtuel s'est peu à peu imposé comme véritable complément au criblage expérimental pour accélérer les premières étapes du processus de conception de nouveaux médicaments [43].

Le criblage virtuel se base, dans la majorité des cas, sur l'emploi de programmes de docking. L'efficacité des algorithmes de docking dans l'identification de nouvelles molécules bioactives n'est plus à démontrer [55, 156]. Toutefois, leur degré de sophistication les rend souvent inadaptés au criblage virtuel haut-débit où le nombre de composés à traiter avoisine le million. Il est donc nécessaire, dans pareille situation, de recourir à d'autres filtres en amont afin de pouvoir effectuer le processus de criblage dans un temps raisonnable.

C'est dans cette optique que nous avons développé la plateforme logicielle VSM-G (**V**irtual **S**creening **M**anager for computational **G**rids) afin de proposer un outil original et intégré, permettant le criblage de larges chimiothèques. Le but premier de VSM-G concerne l'accessibilité et la facilité d'utilisation à travers une interface graphique dans laquelle certains aspects les plus "techniques" de la mise en place d'un protocole de criblage sont pris en charge de façon transparente pour l'utilisateur.

I-2. Présentation de la plateforme VSM-G

Une description détaillée de la plateforme VSM-G est présentée dans l'article #2 : "*Multiple-step virtual screening using VSM-G: Overview and validation of fast geometrical matching enrichment*".

L'objectif principal de la plateforme VSM-G est de permettre, via une interface conviviale, d'effectuer des études de criblage virtuel à large échelle. La conduite d'un criblage avec VSM-G s'organise globalement en trois étapes : la préparation de la chimiothèque, la préparation de la cible (pouvant être représentée par un ensemble de conformations) et le criblage multi-étapes. Cette stratégie multi-étapes vise à optimiser le temps de calcul et peut être vue comme un entonnoir au sein duquel diverses techniques de criblage *in silico* sont organisées de façon hiérarchique. L'entonnoir de criblage constituant l'élément central de VSM-G est décrit dans le paragraphe suivant.

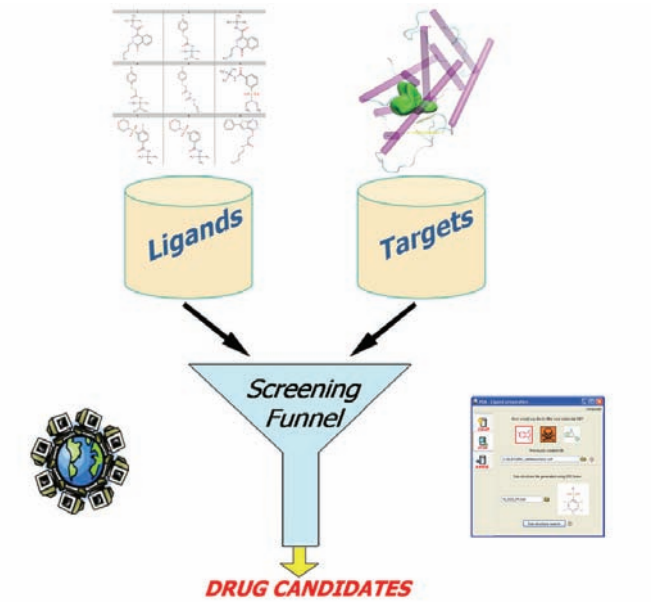


FIG. 19 – Présentation schématique de VSM-G.

I-2.1. Stratégie multi-étapes de VSM-G

Le spectre d'application des techniques *in silico* se mesure généralement selon un compromis entre rapidité et précision. Les algorithmes traditionnels de docking visent à offrir une précision pouvant permettre de classer les composés par ordre d'affinité vis-à-vis d'une cible donnée, avec un taux d'erreur acceptable. Cependant, dans le cadre d'un criblage virtuel haut-débit, leur emploi peut être

discutable à moins de disposer de moyens de calcul considérables [228]. A l'opposé, des filtres plus élémentaires peuvent aisément traiter de larges chimiothèques dans un temps réduit, mais, du fait de leur moindre précision, leur usage à taux de filtrage élevé entraîne souvent un taux de faux négatifs* inacceptable.

Les principaux avantages et inconvénients de ces diverses méthodes ont guidé la mise au point de la stratégie multi-étapes de VSM-G. Les techniques formant l'entonnoir de criblage sont organisées de façon hiérarchique, où des filtres assez rudimentaires sont placés en amont d'autres bien plus élaborés, profitant ainsi, dans l'idéal, de la rapidité des premiers et de la précision des seconds. La succession de ces filtres à complexité croissante vise à éliminer, le plus rapidement possible, les composés de la chimiothèque les moins appropriés à la cible étudiée. Le maximum de temps de calcul est alors consacré aux molécules jugées les plus prometteuses.

I-2.2. Premier filtre de l'entonnoir de criblage de VSM-G

Au niveau supérieur de la stratégie de criblage virtuel implémentée dans VSM-G, la rapidité d'exécution est recherchée. L'objectif est de pouvoir tester plusieurs millions de molécules dans des temps de calcul raisonnables et de ne retenir que quelques dizaines de milliers de composés qui seront ensuite soumis au filtre suivant. Ce type de filtre rapide peut être issu, par exemple, des techniques de recherche de similarité sur la base de ligands connus pour être actifs ou encore de la recherche de motifs pharmacophoriques [151, 153]. D'autres approches consistant à comparer des surfaces moléculaires peuvent aussi être employées à ce stade d'un criblage virtuel haut-débit [229-231].

Dans son implémentation actuelle, la première étape de l'entonnoir de VSM-G est un filtre purement géométrique qui évalue la complémentarité de formes entre celles de chaque conformation des ligands et celles de chaque conformation du site actif de la protéine. Ce programme, appelé SHEF, et dont la représentation se base sur les harmoniques sphériques [168, 169, 232], est présenté dans l'article #1 : "*SHEF: a vHTS geometrical filter using coefficients of spherical harmonics molecular surfaces*". Cet article montre le potentiel de cette approche pour son utilisation future comme technique d'enrichissement dans le criblage de larges chimiothèques.

* Molécules prédites à tort avec une faible ou non-activité. La génération de faux négatifs correspond à une perte d'information pertinente au cours du filtrage. La génération de faux positifs, inévitable au cours d'un criblage, est moins handicapante car elle ne correspond qu'à une perte de temps dans le cadre d'un protocole multi-étapes.

I-2.3. Validation du filtre géométrique au sein de l'entonnoir de VSM-G

Afin de pouvoir utiliser VSM-G dans des études de criblage virtuel haut-débit, nous avons procédé à la validation du filtre SHEF dans le contexte de la stratégie multi-étapes. Cette preuve de concept est décrite dans l'article #2 : "*Multiple-step virtual screening using VSM-G: Overview and validation of fast geometrical matching enrichment"* [*] qui, en plus de présenter la plateforme, évalue l'efficacité du module SHEF comme premier filtre de l'entonnoir de VSM-G lors d'un criblage-test sur le récepteur nucléaire LXRβ. Il est montré que SHEF permet de filtrer rapidement une chimiothèque sans compromettre l'identification de molécules candidates[†]. De plus, il s'avère que l'emploi d'un ensemble de conformations de la cible, pour prendre en compte, au moins partiellement, la flexibilité de son site actif, peut avoir un impact favorable sur les performances de SHEF.

I-2.4. Examen de la flexibilité du site actif de LXRβ

A la suite des travaux méthodologiques concernant la plateforme VSM-G, pour lesquels nous avons utilisé comme système test le récepteur LXRβ, il est apparu que ce récepteur présentait une diversité conformationnelle assez importante au niveau de son site actif. Le simple fait que ce domaine ait été co-cristallisé avec plusieurs agonistes de tailles et de formes très diverses, témoigne déjà de la grande plasticité de son site actif. Se posait donc la question de savoir si celle-ci était intrinsèque ou induite par la complexation de ligands (phénomène d'"*induced fit"*). C'est pourquoi, avant d'entreprendre la campagne de criblage haut-débit sur la cible LXRβ, nous avons analysé la flexibilité de son site actif.

Dans cette optique, l'article #3 : "*Induced fit in Liver X Receptor beta: A molecular dynamics-based investigation"* élucide, au moyen d'une série de simulations de dynamique moléculaire (DM), les mécanismes moléculaires mis en jeu dans le phénomène d'induction par le ligand rencontré dans LXRβ. Cette étude a permis d'identifier des résidus clés du site actif impliqués dans son remodelage lors de la reconnaissance moléculaire protéine-ligand. Cet article illustre enfin comment un tel protocole de DM pourrait être employé aussi bien (i) comme générateur de conformations du site actif, utiles en entrée d'un criblage (cf. article #2 et section I.3) ; que (ii) à la sortie de l'entonnoir de criblage de VSM-G pour optimiser les meilleurs candidats issus des filtres précédents.

[*] Note : dans l'article #1, le "s*core*" de SHEF reflète la complémentarité de la forme d'un ligand avec celle du site actif de la protéine. La fonction de score qu'utilise SHEF dans l'article #2 est la même mais elle est appelée "RMSD" ; dans ce cas, cela revient à évaluer la similarité entre la surface d'un ligand et celle du "ligand idéal" qui représenterait l'empreinte négative du site actif de la protéine.

[†] L'évaluation du filtrage de SHEF a été réalisée en comparant ses résultats avec ceux du programme GOLD, second filtre de l'entonnoir pour cette étude.

I-3. Article # 1

"SHEF: a vHTS geometrical filter using coefficients of spherical harmonics molecular surfaces", soumis à *Journal of Molecular Modeling*.

I-4. Article # 2

"Multiple-step virtual screening using VSM-G: Overview and validation of fast geometrical matching enrichment", *Journal of Molecular Modeling*, sous presse.

I-5. Article # 3

"Induced fit in Liver X Receptor beta: A molecular dynamics-based investigation", *Proteins: Structure, Function, and Bioinformatics*, accepté.

ARTICLE #1

"SHEF: a vHTS geometrical filter using coefficients of spherical harmonics molecular surfaces"

SHEF: A VHTS GEOMETRICAL FILTER USING COEFFICIENTS OF SPHERICAL HARMONIC MOLECULAR SURFACES

Wensheng Cai [1,*], Jiawei Xu [2], Xueguang Shao [1], Vincent Leroux [3], Alexandre Beautrait [3], Bernard Maigret [3]

[1] *Department of Chemistry, Nankai University, Tianjin 300071, P.R. China.*
[2] *Department of Chemistry, University of Science & Technology of China Hefei, Anhui 230026, P.R. China.*
[3] *Nancy Université, LORIA, Groupe ORPAILLEUR, Campus scientifique, BP 239, 54506 Vandœuvre-lès-Nancy Cedex, France.*

* Corresponding author: wscai@nankai.edu.cn

ABSTRACT

SHEF, a geometrical matching procedure constituting a preliminary step in virtual high throughput screening of large databases of small drug-like molecules, is demonstrated. This filter uses a description of both target's binding site and ligand surfaces using spherical harmonic polynomial expansions. Using this representation, which is based upon limited sets of spherical harmonic coefficients, the complexity of surface complementarity calculation is reduced considerably. As a first test, 188 known protein-ligand complexes were used and the results of docking the abstracted ligands into the bare proteins using SHEF were compared to the original X-ray structures. The ability of SHEF to retrieve known ligands "hidden" in a virtual library of 1000 randomly selected drug-like compounds is also demonstrated.

KEYWORDS

protein-ligand interactions; virtual screening; spherical harmonic expansions; molecular surfaces; surface complementarity and similarity; drug discovery; SHEF.

INTRODUCTION

Recent progress in high-through screening (HTS) and combinatorial chemistry has greatly improved the hit-rate and cost-effectiveness of drug discovery campaigns, so that it radically changed the chemist's approach to drug design. Using computers, virtual high-throughput screening (vHTS) is gaining use in drug discovery as a complementary approach to the experimental techniques [1-4].

Associated with the vHTS strategy, numerous docking algorithms have been reported in literature and several reviews are dealing with the merits of them [5-9]. These algorithms use more or less accurate physico-chemical representations of both receptor and ligand structures. These are associated with scoring functions [10] (necessarily approximate) to measure the docking efficiency. The docking processes are finally driven by search strategies that, due to the complexity of the problem, are not usually exhaustive. These "classical" docking methods can give good results in the hit discovery context [11], but the time (and cost) of computation is too great for screening millions of compounds.

Preliminary crude but fast filters are thus required in large vHTS campaigns in order to reduce the number of candidate molecules to be passed to more elaborate docking calculations. In this context, several filtering methods, using for example shape [12] or fingerprint signatures [13], have already been proposed. The main goal of such approaches is to overcome the time bottleneck of accurate docking methods in structure-based drug design strategies [14].

Here, we describe the SHEF filtering method (Spherical Harmonic coEfficient Filter) whose aim is to fulfill this objective. The core of the SHEF method is the generation of a set of spherical harmonic coefficients that converts 3D surface information into a 1D coefficient vector. A scoring function that only uses these coefficients to compare surfaces is then used. In order to test the effectiveness of the method, the following experiments have been carried out: (i) SHEF was applied to a test system consisting of 188 protein-ligand complexes selected from the PDB database [15] ; (ii) SHEF was tested for its capabilities to retrieve known active ligands hidden in a database of randomly selected compounds; (iii) SHEF computational cost has been evaluated and compared with another vHTS method.

METHODS

Representations of molecular surfaces using spherical harmonic expansions

Spherical harmonics (SH) are single-valued, continuous bounded, complex functions of the spherical coordinates (θ,ϕ), which can be considered as "standing waves on a sphere". They are characterized by two "quantum numbers", l and m, which together determine the number and spatial arrangement of nodes in each function. SH functions [16-19] are evaluated using equation (1)

$$Y_l^m(\theta,\phi) = \sqrt{\frac{2l+1}{4\pi}\frac{(l-m)!}{(l+m)!}} P_l^m(\cos\theta) e^{im\phi} \tag{1}$$

where l and m are integers (with $-l \le m \le l$), and $P_l^m(\cos\theta)$ the associated Legendre functions, that form a complete orthonormal basis set.

For a given protein-ligand complex, the molecular surfaces of a ligand and of the cavity in its binding site region can be modeled by our deflation and inflation techniques [20, 21]. Any single-valued three dimensional surfaces can be approximated by encoding the radial distance of surface points from the origin as a sum of SH functions as follows:

$$r(\theta,\phi) = \sum_{l=0}^{L} \sum_{m=-l}^{l} C_{lm} Y_l^m(\theta,\phi) \tag{2}$$

In this equation, $r(\theta,\phi)$ is the distance function of surface points from the origin inside. C_{lm} are the expansion coefficients of SH arranged by l and m ($0 \le l \le L$; $-l \le m \le l$). L is the order that determines the degree or accuracy of the representation.

Therefore, the C_{lm} set of coefficients, considered as "surface descriptors", can completely define and represent the 3D surface shape, as approximated by SH expansions. It is possible to attain any degree of accuracy by adjusting the expansion order of the coefficients. Thus, any 3D surface shape can be converted into a 1D vector and, consequently, the comparison of different 3D molecular surfaces can be achieved by matching their corresponding 1D coefficient vectors.

Surface comparison using the expansion coefficients

The molecular surfaces of a target binding site and of a ligand being represented by their expansion coefficients, a shape comparison between the two surfaces can be now achieved. For this purpose, considering the surface of the target as rigid and fixed, the coefficients of the molecule are rotated in order to obtain the minimal root-mean-square distance (RMSD) of these coefficients and those of the target. The rotation matrix used for this purpose has been described by Ritchie and Kemp [22, 23].

The difference D of coefficients [22] is applied to measure the shape similarity in this study. If vectors $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{B}'$ are the SH surface representations of the target receptor active site, the ligand and the rotated ligand, respectively (those vectors having $(L+1)^2$ SH coefficients), and $\mathbf{A}_1$, $\mathbf{B}_1$ and $\mathbf{B}'_1$ the centroid vectors restricted to $l = 1$ (thus possessing three coefficients, representing the surfaces' average orientation in the Cartesian space), then:

$$D(\mathbf{A},\mathbf{B},\mathbf{B}') = \sqrt{f(\mathbf{A},\mathbf{B}) - 2g(\mathbf{A},\mathbf{B},\mathbf{B}')} \tag{3}$$

with:

$$f(\mathbf{A},\mathbf{B}) = \|\mathbf{A}\|^2 + \|\mathbf{B}\|^2 + \frac{1}{3}\|\mathbf{A}_1 - \mathbf{B}_1\|^2$$

$$g(\mathbf{A},\mathbf{B},\mathbf{B}') = \mathbf{A} \bullet \mathbf{B}' + \frac{1}{3}(\mathbf{A}_1 - \mathbf{B}_1) \bullet (\mathbf{A}_1 - \mathbf{B}'_1)$$

A three-step optimization strategy was designed to minimize D (hence to maximize g) for one screening process. In the first step, a grid exploration is performed, using the Euler angles to rotate the coefficients of the molecule surface with a regular increment (30° was used here). In the second step, for each of the best 10 orientations found previously, its 27 neighbors (each of the three Euler angles kept or varied by ± 10°) are also calculated, and the new best 10 solutions are selected from the total 270 orientations. In the last step, the local minimizer L-BFGS [24] is applied to optimize these 10 orientations. From this procedure, an optimal set of Euler angles giving the best similarity score to the pair surfaces can be finally obtained. The final coefficient string related to the molecule surface obtained this way can be used to evaluate its similarity to the target one.

Construction of the filter

In this method, for a given target binding site surface, the flexibility of the ligand partners can be modeled by considering different conformers as separate docking candidates. The first step of our procedure is therefore to generate a set of low energy conformers for each ligand and the calculation of the associated SH surfaces. The resulting coefficients constitute our ligand-coefficients database. An analogous target-coefficients database can also be generated; this can encompass several active site conformations obtained either from diverse experimental structures or by conformational sampling methods (molecular dynamics or Monte Carlo simulations).

Note that the databases (both ligands and proteins) are reusable for further applications; the calculation of SH coefficients has to be done only once. Moreover, new molecules and receptors will be easily added. The size of such databases is reasonable compared to storing molecular structures as atomic coordinates and bond information records. More interestingly, provided the same expansion order is used for all the conformers in the database, the size of each record is constant. This allows the implementation of efficient schemes for storing and accessing data.

Matching the surface of each candidate conformer to a given target is realized by minimizing the difference function between two coefficient sets as stated in equation (3). Another scoring function is then used to evaluate the optimized pose. If $\mathbf{A}_{\bar{0}}$ and $\mathbf{B}'_{\bar{0}}$ are the $\mathbf{A}$ and $\mathbf{B}'$ vectors minus the first coefficient (for which $l = m = 0$, representing the average radius of the SH expansion volume), then:

$$Score(\mathbf{A},\mathbf{B}') = \frac{\|\mathbf{A}-\mathbf{B}'\|}{\|\mathbf{B}'\|} + w\,(1-\cos(\mathbf{A}_{\bar{0}},\mathbf{B}'_{\bar{0}})) \qquad (4)$$

with:

$$w = \frac{\max(\|\mathbf{A}\|,\|\mathbf{B}'\|)}{\min(\|\mathbf{A}\|,\|\mathbf{B}'\|)}$$

$$\cos(\mathbf{A}_{\bar{0}},\mathbf{B}'_{\bar{0}}) = \frac{\mathbf{A}_{\bar{0}} \bullet \mathbf{B}'_{\bar{0}}}{\|\mathbf{A}_{\bar{0}}\|\|\mathbf{B}'_{\bar{0}}\|}$$

It should be noted that as the radial coefficient is usually much larger than the others, the first term in equation (4) is most sensitive to the size matching between the two surfaces described by the $\mathbf{A}$ and $\mathbf{B}'$ vectors. The second term mainly delineates the shape similarity.

The *Score* value is the criterion used for screening using SHEF. For each ligand molecule the conformer providing the lowest score is retained and the relative effectiveness of the ligands are compared using these values in our vHTS procedure. The whole virtual screening process is shown in **figure 1**.

Docking protocol

A test set of 188 protein-ligand complexes has been chosen from the PDB database on the basis of their diversity and non-redundancy. For each complex the ligand was taken apart from the protein active site and redocked using SHEF. The goal here is to compare the poses obtained after the SHEF coefficient optimization procedure and those from the X-ray structures of the corresponding complexes.

The calculation time and precision obviously depend on the value of the SH expansion order *L*. In order to measure this behavior, different values of *L* were used for the docking calculations. Since for surface matching, large *L* (>10) is not necessary [22], values from 3 to 10 were tested.

Input data for evaluating filtering efficiency in the virtual screening context

Another important test is to check how good SHEF is as a screening filter. The significance and efficiency of a filter depend on how effectively it can sort out suitable compounds from the input database. An efficient filter is therefore one that can reduce the database to a manageable size for the subsequent more precise experimental measurements and/or structure-based drug discovery techniques. Methods used for such a purpose have to be able to select the most probable inhibitors from randomly chosen drug-like molecules.

A random database with 1000 compounds, randomly selected from 10000 drug-like compounds in NCI 3D database [25, 26], has been constructed. The average number of atoms per compound is about 32, and the average molecular weight is 237.6. The conformational flexibility of molecules was considered by storing multiple conformers for each of them. The corresponding structures were first generated using OMEGA [27], giving an average of 34 conformers per molecule. Next the SH expansion coefficients of each conformer were calculated (L = 5, giving a vector of 36 coefficients) and stored in the ligand-coefficients database. These data were used as the decoys in the ligands database.

Metrics for measuring filtering performance

Given a test database composed of *n* structures, divided into the actives *a* molecules (with known activity for the reference target) and the decoys *d* random molecules (with presumably no affinity for the target).

A screening also divides *n* into two groups: those that are predicted to be active (*h* hits) and those that are filtered out (*f*). With a virtual screening program such as SHEF ranking molecules using a scoring function for evaluating affinity, the *h* value is a parameter set by the user. Screening performance is related to the number of retrieved actives h_a, and inversely related to the number of false positives f_+ and false negatives f_-. Those definitions are summarized in **figure 2.a**.

From this, a number of metrics for evaluating virtual screening performance (bound from 0 to 1, and that can be expressed as percentages) can be formulated. The filtering amount *F* (taken as the screening parameter), the coverage *C*, the yield of actives *Y*, the efficiency *E* and the Güner-Henry score *GH* [28] are defined as:

$$\begin{aligned} F &= (1-\frac{h}{n}) \\ C(F) &= \frac{h_a}{a} \\ Y(F) &= \frac{h_a}{h} \\ E(F) &= \frac{Y(F)}{Y(0)} = \frac{nh_a}{ah} \\ GH(F) &= wY(F) + (1-w)C(F) \end{aligned} \tag{5}$$

In order to have a single value for a given method, only the filtering amount *F**, the maximum value giving full coverage ($h_a = a$; $f_- = 0$) was computed in our tests:

$$F^* = \max\left(F \middle/ C(F) = 1\right) \tag{6}$$

This particular case is represented on **figure 2.b**. In order to better express the screening accuracy, *w* is set to 0.75, and the GH-score is weighted using the ratio of false positives on decoys. Finally GH^*, a value derived from the GH-score expressing the screening efficiency is obtained:

$$GH^* = \frac{3Y(F^*)+1}{4}\left(1-\frac{f_+(F^*)}{d}\right) \tag{7}$$

RESULTS AND DISCUSSION

Rigid docking test of 188 complexes

After performing SHEF for the 188 ligands and the related binding sites, the atomic coordinates of each of them corresponding to the obtained minimal *D* value (see **equation 3**) was compared to the original X-ray structure. **Figure 3** shows the relationship between the docking results and the order *L*. It can be seen that the average RMSD between the experimental ligand bound conformation and the docking results for 188 complexes decreases very quickly until *L* is equal to 5, and then it changes slightly when *L* is between 5 and 10. Consequently, a value of $L = 5$ is recommended and was used in the docking tests presented here.

Docking results for $L = 5$ are given in **table 1**. The RMSDs of all 188 entries, except two, are smaller than 3.0 Å, giving an average RMSD of 0.813 Å. Two complexes, namely 1HVR and 2DBE, have much larger RMSD: 10.0 Å and 8.7 Å for 1HVR and 2DBE respectively because of a symmetry problem while, in fact, the two SHEF poses fit quite well with the X-ray ones.

It should be noted that the optimized value of *D* for each complex reflects the degree of solvent exposure of the corresponding binding cavities. Indeed, small *D* means that the ligand is entirely embedded in the much closed binding cavity, whereas, larger values indicate that the protein holds a more open binding cavity, so that a limited match exists between the ligand and the binding site. Most of the complexes in our test set have relatively closed or partly opened cavity and show therefore good complementarities. As an example, the crystal structures of three complexes, namely 1IE9, 1HVK and 1ETS, and their interface sections generated by their optimized coefficients are shown in **figure 4**.

Filtering performance in virtual screening

To perform the screening test, two representative groups of protein-ligand complexes were selected from **table 1** according to their binding cavity characteristics. They possess closed and half-closed cavities and are classified into two distinct protein families (**table 2**): vitamin D (9 complexes) and HIV-1 protease receptors (30 complexes). The corresponding ligands are considered as actives in the filtering process, and are merged in two composite databases based on the 1000 random drug-like decoys.

The vitamin D and HIV-1 databases have respectively 1009 and 1030 compounds; 33765 and 38015 conformers. In order to assess the robustness of the method, the X-ray conformations of the known ligands were removed from the database in our test experiments, leaving only OMEGA-generated conformers.

The filtering results are shown in **figure 5**. The X-axis denotes the optimized *Score* (**equation 4**) of the compounds in the composite database against the target cavity, and the solid line indicates the corresponding cutoff score (the largest one among all active compounds) used to filter the docking poses in order to retrieve all active compounds ($C = 1$; $F = F^*$). The Y-axis denotes the corresponding *Score* of the compounds against the target ligand, and the dash cutoff line is used to recover all known ligands. The low left rectangular formed by these two cutoff lines and two axes recovers all active compounds when screening against both target receptor and target ligand.

It can be seen from **figure 5** that the distribution of most of the points in each figure is mostly linear. It means that the more complementary a candidate molecule is to the receptor target, the more similar this molecule is also to the reference ligand. It can also be seen that the distribution of the points in **figure 5.b** is better than that in **figure 5.a**. It is due to the higher sensitivity of the SH procedure to complicated shapes presenting several clear lobes and holes of the ligands and the receptor binding site which were higher in HIV-1 protease complexes than in the Vitamin-D ones.

The effectiveness of the filtering has been measured using the GH^* value (see **equation 7**) – results are shown in **table 3**. The corresponding a, d, F^*, $h(F^*)$, $Y(F^*)$ and $E(F^*)$ values (**equations 5 and 6**) are also shown. In order to compare SHEF results with those of a classical rigid docking method recognized for performing exhaustive and fast calculations, virtual screening on the reference dataset was also done using FRED [27]. The results indicate clearly that SHEF is superior to FRED regarding filtering performance.

CPU time used for computing the coefficients and screening the coefficient database

In SHEF, the total computational time consists of two parts, the CPU time used for calculating the coefficients to build the ligand- and target's pocket-coefficient databases, and the CPU time used for the screening itself. The average CPU time on a computer composed of a AMD MP2200+ processor with 1 Gb memory (with a computing speed comparable to the CPUs in current low-end PC desktops) used for calculating the coefficients with $L = 5$ for one ligand conformer (with 32 atoms) is about 1 second. For one protein cavity (with 350 wall atoms) it takes about twenty seconds. Both these calculations need to be done only once.

The filter will then work from the pre-constructed coefficient databases without any atomic coordinate information. Using SHEF the average screening time for one conformer is 0.046 s, which is about 2.4 times faster than FRED on the same computers. The average time to screen a compound (*i.e.* 34 conformers) is 1.564 s, which is much faster than the new technique proposed by Putta *et al.* [29].

CONCLUSION

An efficient filter SHEF for vHTS using the SH coefficients of the molecular surfaces has been presented. Both the rigid docking and filtering performance tests of this method gave satisfactory results. The accuracy of the flexible docking depends on the pre-generated conformers in the database. The aim is to eliminate most of the compounds or conformers that do not fit to the target binding cavity, and not to identify the best binders. More accurate docking calculations based on binding energy estimation should be applied to the selected ligands. SHEF is therefore a method which can be used as a potential fast and efficient filter prior to more efficient techniques in the vHTS context. As so, it confirms that techniques using purely geometrical representations of the active site and the candidate ligands can provide positive results [30].

In this paper, basic test experiments have been performed. In fact we have implemented SHEF into an integrated package for high-throughput virtual screening, the VSM-G platform. The combined use of SHEF with a classical docking program using this software was validated as a relevant enrichment technique in large-scale virtual screening experiments [31]. Additionally, although SHEF focuses on geometrical complementarity, it could be extended to include chemical features as to provide a more extensive measure of protein-ligand binding. Such an extension of SH molecular surfaces already gave good results for similarity-based ligand-based drug design approaches [32]. Work for expanding SHEF in a similar fashion is in progress.

ACKNOWLEDGEMENTS

This study is supported by National Natural Scientific Foundation of China (Nos. 20325517 and 20573102), and the Franco-Chinese PRA (Projet de Recherche Avancée) project (B02-04). Alexandre Beautrait was supported by grants from INRIA (Institut National de Recherche en Informatique et en Automatique), Région Lorraine, and ARC (Association pour la Recherche sur le Cancer); Vincent Leroux by a postdoc fellowship of INCa (Institut National du Cancer). We thank Openeye for providing free access to OMEGA, FRED and VIDA software according to an academic license

FIGURES

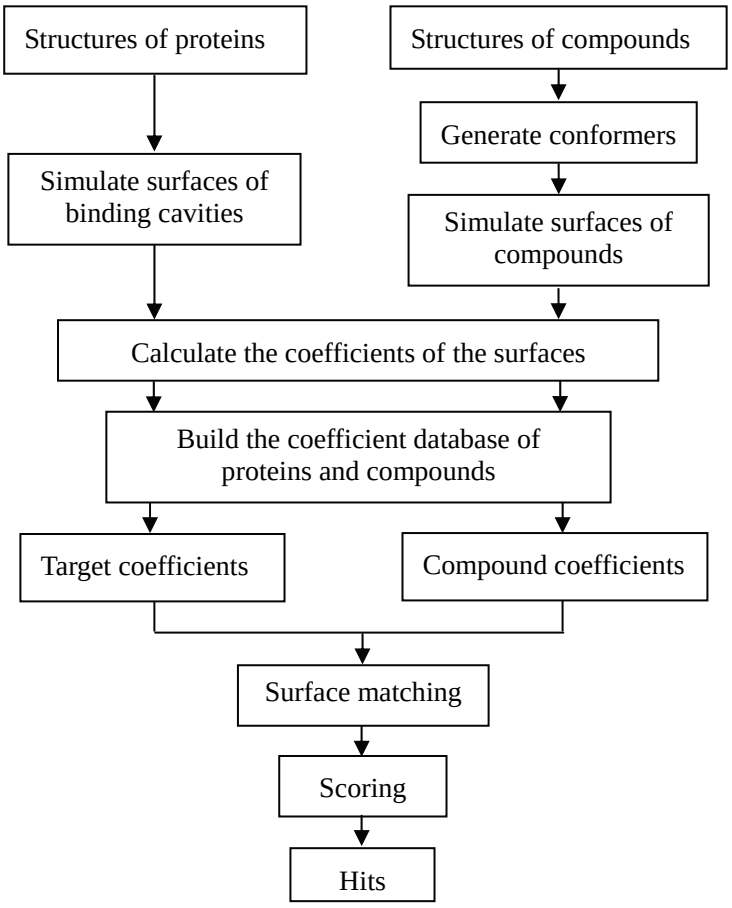


Figure 1 - The flowchart to build a SHEF filter in vHTS.

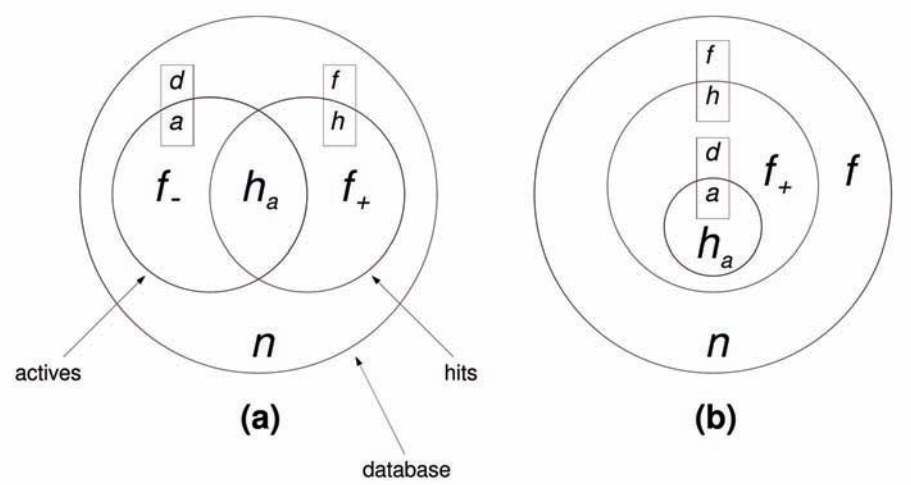


Figure 2 - Definition of molecular database sub-groups. The main circle represents the whole database, while the two inner circles represent the actives molecules and the hits as defined by the screening program. (a) General case. (b) Full coverage of the actives by the screening.

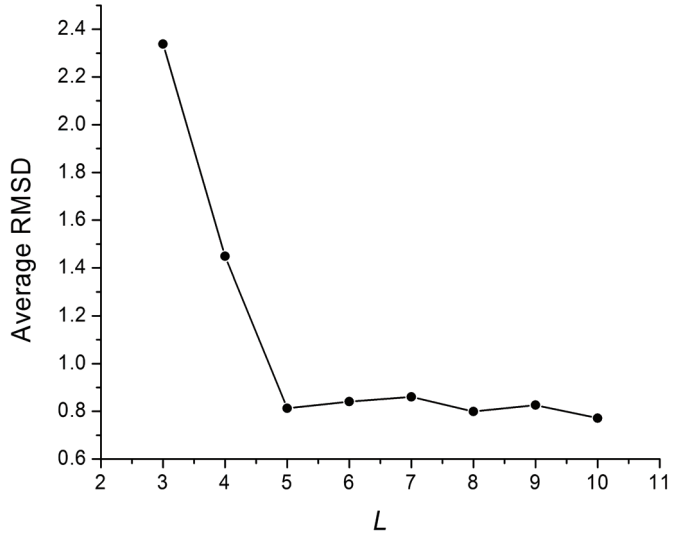


Figure 3 - The average RMSD of the docking results over 188 complexes with different values of the spherical harmonic expansion order L.

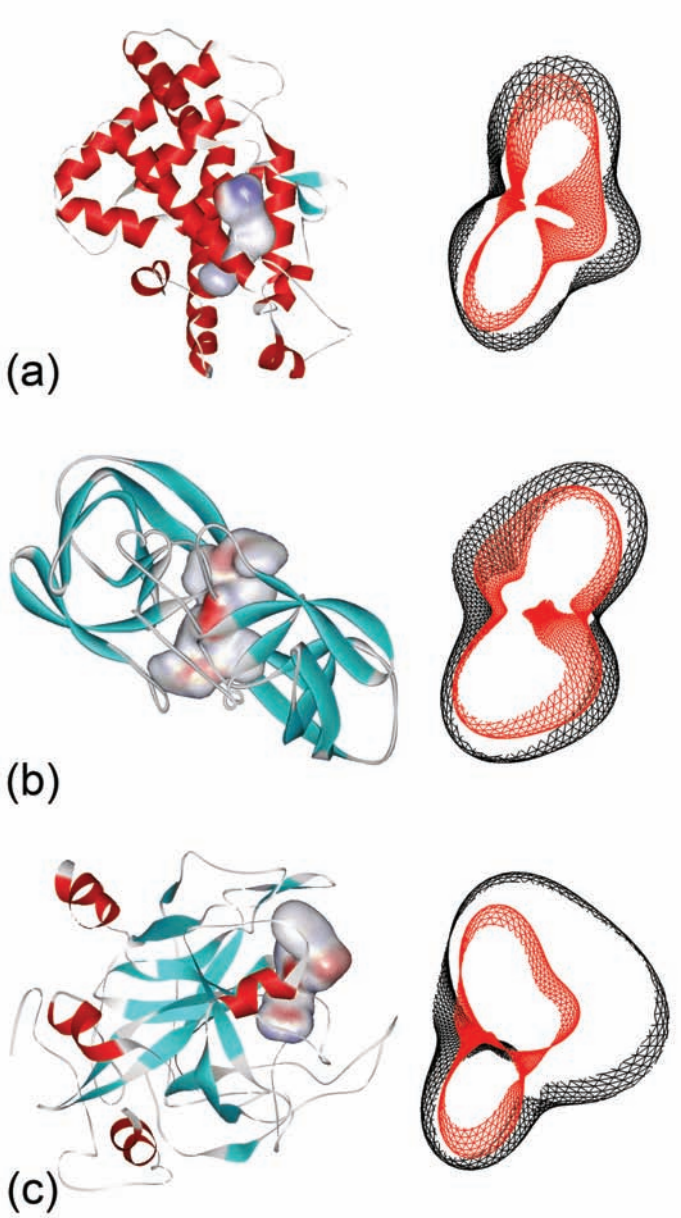


Figure 4 - The crystal structures of the three complexes (left), and their interface section figures generated by their optimized coefficients with L = 5 (right; red: ligand, black: cavity). (a) 1IE9. (b) 1HVK. (c) 1ETS.

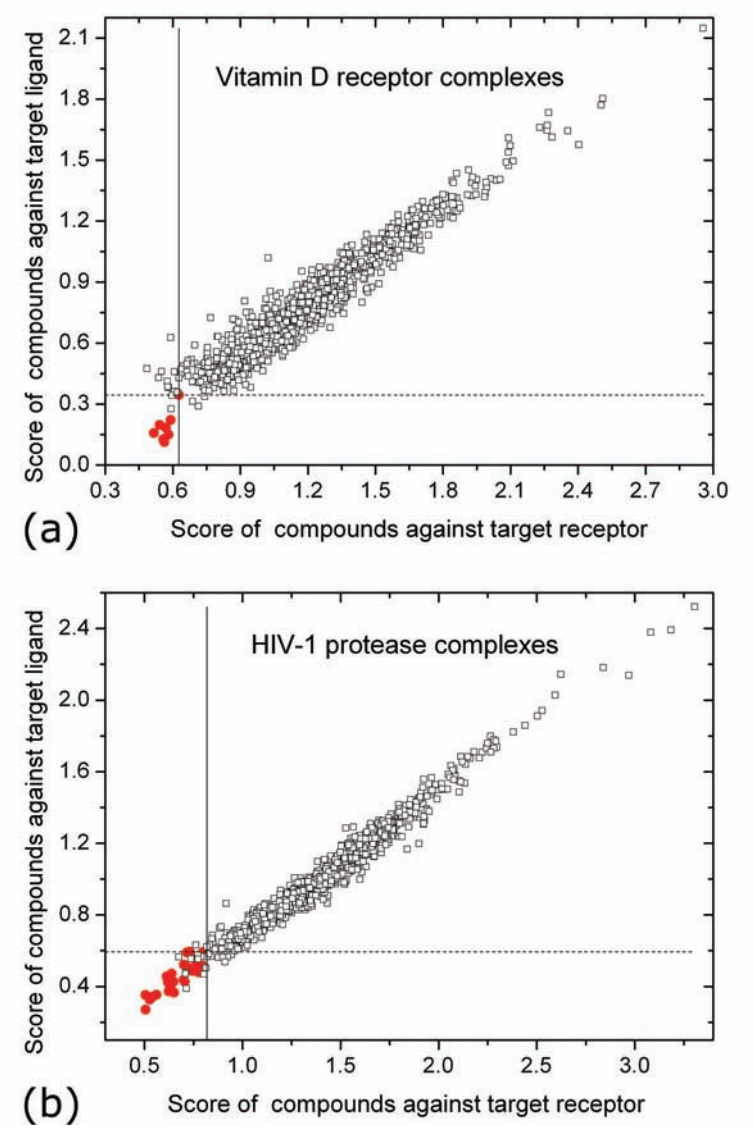


Figure 5 - SHEF filtering results against target receptor and target ligand, respectively. Known ligands (actives) are represented as solid circles, while hollow squares denote the decoys. Solid and dash lines: largest (worst) score amongst the actives against the reference binding site and its ligand respectively. SHEF hits are left to the solid line. (a) Vitamin D receptor complexes (target and reference ligand from PDB structure 1IE9). (b) HIV-1 protease complexes (target and reference ligand from PDB structure 1HVK).

TABLES

No	PDB	RMSD	No	PDB	RMSD	No	PDB	RMSD	No	PDB	RMSD
1	1A8G	0.510	48	1DRF	0.140	95	1LNA	0.469	142	1TNH	1.308
2	1A9M	0.285	49	1DWB	2.034	96	1LST	0.355	143	1TNI	0.338
3	1ABE	1.907	50	1DWD	0.611	97	1LYB	0.885	144	1TNJ	0.754
4	1ABF	0.671	51	1ELA	0.705	98	1MBI	0.320	145	1TNK	1.957
5	1ACJ	0.677	52	1EPO	0.652	99	1MCF	0.688	146	1TNL	1.213
6	1ACM	1.921	53	1EPP	0.807	100	1MCH	0.521	147	1TPH	0.463
7	1ADD	0.333	54	1ETR	0.779	101	1MFC	1.405	148	1TPP	1.469
8	1AHA	0.276	55	1ETS	0.842	102	1MMB	0.411	149	1UKZ	1.041
9	1AJV	0.412	56	1ETT	0.487	103	1MRK	0.912	150	1ULB	0.354
10	1AJX	0.153	57	1FKG	0.161	104	1MUP	0.413	151	1WAP	0.267
11	1APT	0.704	58	1FLR	0.835	105	1NCO	2.072	152	2ACK	0.400
12	1APU	0.788	59	1GHB	0.624	106	1NRR	0.394	153	2ADA	0.236
13	1APV	0.493	60	1GPY	2.725	107	1ODW	0.450	154	2AK3	0.689
14	1APW	0.559	61	1H8D	0.810	108	1ODX	0.539	155	2CGR	1.246
15	1ATL	0.543	62	1HBV	0.330	109	1OKM	0.604	156	2CHT	2.629
16	1B5G	0.506	63	1HDC	0.646	110	1OS0	0.383	157	2CPP	0.344
17	1BAP	1.913	64	1HDT	1.052	111	1PE8	0.194	158	2CTC	0.410
18	1BBP	0.381	65	1HEW	0.312	112	1PHF	0.198	159	**2DBE**	**8.768**
19	1BNM	1.061	66	1HIH	0.105	113	1PHG	0.404	160	2GBP	0.462
20	1BNN	0.658	67	1HIV	0.293	114	1POC	0.452	161	2IFB	0.991
21	1BNQ	1.228	68	1HOS	0.199	115	1PPB	0.316	162	2R04	0.417
22	1BNT	0.745	69	1HPS	0.443	116	1PPC	0.543	163	2R07	0.527
23	1BNU	0.350	70	1HPV	0.145	117	1PPK	0.562	164	2TMN	0.797
24	1BNV	0.569	71	1HRI	0.519	118	1PPL	0.920	165	2TSC	0.727
25	1BRA	2.387	72	1HSL	2.369	119	1QBR	0.143	166	3ER3	1.341
26	1BYB	0.799	73	1HTF	0.907	120	1QBT	0.144	167	3HVT	0.445
27	1BYG	1.602	74	1HTG	0.548	121	1QBU	0.198	168	3PTB	0.595
28	1C2T	0.370	75	1HVI	0.212	122	1QF0	0.252	169	3TMN	0.480
29	1C83	0.281	76	1HVJ	0.134	123	1QF1	0.056	170	3TPI	0.467
30	1CBS	0.832	77	1HVK	0.138	124	1QF2	0.585	171	4AAH	0.508
31	1CBX	0.568	78	1HVL	0.074	125	1RGK	0.629	172	4DFR	0.771
32	1CIM	1.330	79	**1HVR**	**10.031**	126	1RGL	0.709	173	4HVP	0.275
33	1COM	2.875	80	1HXB	0.412	127	1RJK	0.705	174	4PHV	0.443
34	1COY	0.664	81	1HYT	0.891	128	1RK3	0.532	175	4TMN	0.269
35	1CPS	0.918	82	1ICN	0.559	129	1RKG	0.237	176	5ER2	1.657
36	1CTT	1.728	83	1IDA	0.115	130	1RKH	0.632	177	5HVP	0.128
37	1D3H	0.255	84	1IE8	0.588	131	1RNE	0.652	178	5P21	0.602
38	1D4P	0.915	85	1IE9	0.500	132	1ROB	0.381	179	5TLN	1.030
39	1DB1	0.514	86	1IGJ	1.046	133	1S0Z	0.743	180	6ABP	0.836
40	1DBJ	0.611	87	1INC	0.748	134	1S19	0.579	181	7CPA	0.224
41	1DBK	0.572	88	1JAP	0.875	135	1SNC	0.462	182	7HVP	0.414
42	1DBM	0.61	89	1KEL	0.695	136	1STP	2.210	183	7LPR	1.251
43	1DD7	0.354	90	1KR6	0.420	137	1THL	0.245	184	7TIM	0.527
44	1DID	0.584	91	1KS7	0.860	138	1TKA	1.095	185	8ATC	2.502
45	1DIE	2.699	92	1LAH	0.308	139	1TLP	0.507	186	8CPA	0.492
46	1DIF	0.130	93	1LDM	1.646	140	1TMN	0.564	187	8GCH	0.976
47	1DMP	0.077	94	1LIC	1.242	141	1TNG	0.995	188	9HVP	0.159

Table 1 - Docking results of 188 complexes: RMSD values between the ligands from the experimental structures and the SHEF docking predictions. The average RMSD over the 188 complexes is 0.813 Å. Note: the 1HVR and 2DBE ligands have symmetrical structures and were "flipped" upon docking, hence the large RMSD values, which are indeed not correlated to a bad prediction.

Protein Family	Known ligands/Active compounds	PDB ID of complexes
I: Vitamin D receptor complexes	9	1DB1 1IE8 1IE9 1RJK 1RK3 1RKG 1RKH 1S0Z 1S19
II: HIV-1 protease complexes	30	1A8G 1A9M 1AJV 1AJX 1DIF 1DMP 1HBV 1HIH 1HIV 1HOS 1HPS 1HPV 1HTF 1HTG 1HVI 1HVJ 1HVK 1HVL 1HVR 1HXB 1ODW 1ODX 1QBR 1QBT 1QBU 4HVP 4PHV 5HVP 7HVP 9HVP

Table 2 - PDB codes of the structures composing the two protein families used in the filtering test.

(a)

a = 9; d = 1000	F^* (%)	h	Y (%)	E (%)	GH^* (%)
SHEF	97.9	21	42.9	48.0	56.5
FRED	92.0	81	11.1	12.5	30.9

(b)

a = 30; d = 1000	F^* (%)	h	Y (%)	E (%)	GH^* (%)
SHEF	96.0	41	73.2	25.1	79.0
FRED	95.2	49	61.2	21.0	69.6

Table 3 - Effectiveness of SHEF and FRED measured after maximum filtering upon total coverage (all actives recovered). Metrics are expressed as percentages. (a) Vitamin-D receptor complexes. (b) HIV-1 protease complexes.

REFERENCES

(1) Walters, W. P.; Stahl, M. T.; Murcko, M. A. Virtual screening - an overview. *Drug Discov. Today* **1998**, *3* (4), 160-178.

(2) Cavasotto, C. N.; Orry, A. J. Ligand docking and structure-based virtual screening in drug discovery. *Curr. Top. Med. Chem.* **2007**, *7* (10), 1006-1014.

(3) Reddy, A. S.; Pati, S. P.; Kumar, P. P.; Pradheep, H. N.; Sastry, G. N. Virtual screening in drug discovery - a computational perspective. *Curr. Protein Pept. Sci.* **2007**, *8* (4), 329-351.

(4) Seifert, M. H. J.; Kraus, J.; Kramer, B. Virtual high-throughput screening of molecular databases. *Curr. Opin. Drug Discovery Dev.* **2007**, *10* (3), 298-307.

(5) Brooijmans, N.; Kuntz, I. D. Molecular recognition and docking algorithms. *Annu. Rev. Biophys. Biomol. Struct.* **2003**, *32*, 335-373.

(6) Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins: Struct., Funct., Bioinf.* **2004**, *57* (2), 225-242.

(7) Leach, A. R.; Shoichet, B. K.; Peishoff, C. E. Prediction of protein-ligand interactions. Docking and scoring: successes and gaps. *J. Med. Chem.* **2006**, *49* (20), 5851-5855.

(8) Sousa, S. F.; Fernandes, P. A.; Ramos, M. J. Protein-ligand docking: current status and future challenges. *Proteins: Struct., Funct., Bioinf.* **2006**, *65* (1), 15-26.

(9) Joseph-McCarthy, D.; Baber, J. C.; Feyfant, E.; Thompson, D. C.; Humblet, C. Lead optimization via high-throughput molecular docking. *Curr. Opin. Drug Discovery Dev.* **2007**, *10* (3), 264-274.

(10) Jain, A. N. Scoring functions for protein-ligand docking. *Curr. Protein Pept. Sci.* **2006**, *7* (5), 407-420.

(11) Abagyan, R.; Totrov, M. High-throughput docking and lead generation. *Curr. Opin. Chem. Biol.* **2001**, *5* (4), 375-382.

(12) Proschak, E.; Rupp, M.; Derksen, S.; Schneider, G. Shapelets: possibilities and limitations of shape-based virtual screening. *J. Comput. Chem.* **2007**, *in the press*.

(13) Deng, Z.; Chuaqui, C.; Singh, J. Structural information fingerprint (SIFt): a novel method for analyzing three-dimensional protein-ligand binding interactions. *J. Med. Chem.* **2004**, *47* (2), 337-344.

(14) Bleicher, K. H.; Böhm, H.-J.; Müller, K.; Alanine, A. I. Hit and lead generation: beyond high-throughput screening. *Nat. Rev. Drug Discovery* **2003**, *2* (5), 369-378.

(15) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Research* **2000**, *28* (1), 235-242.

(16) Leicester, S. E.; Finney, J. L.; Bywater, R. P. Description of molecular surface shape using Fourier descriptors. *J. Mol. Graphics* **1998**, *6* (2), 104-108.

(17) Max, N. L.; Getzoff, E. D. Spherical harmonic molecular surfaces. *IEEE Comput. Graph. Appl.* **1988**, *8* (4), 42-50.

(18) Duncan, B. S.; Olson, A. J. Approximation and characterization of molecular surfaces. *Biopolymers* **1993**, *33* (2), 219-229.

(19) Barnett, M. P. Transformation of harmonics for molecular calculations. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (4), 1158-1165.

(20) Cai, W.; Zhang, M.; Maigret, B. New approach for representation of molecular surface. *J. Comput. Chem.* **1998**, *19* (16), 1805-1815.

(21) Cai, W.; Shao, X.; Maigret, B. Protein-ligand recognition using spherical harmonic molecular surfaces: towards a fast and efficient filter for large virtual throughput screening. *J. Mol. Graphics Modell.* **2002**, *20* (4), 313-328.

(22) Ritchie, D. W.; Kemp, G. J. L. Fast computation, rotation, and comparison of low resolution spherical harmonic molecular surfaces. *J. Comput. Chem.* **1999**, *20* (4), 383-395.

(23) Ritchie, D. W.; Kemp, G. J. L. Protein docking using spherical polar Fourier correlations. *Proteins: Struct., Funct., Genet.* **2000**, *39* (2), 178-194.

(24) Liu, D. C.; Nocedal, J. On the limited memory BGFS method for large scale optimization. *Math. Program.* **1989**, *45*, 503-528.

(25) Milne, G.; Nicklaus, M.; Driscoll, J.; Wang, S.; D, Z. National Cancer Institute drug information system 3D database. *J. Chem. Inf. Comput. Sci.* **1994**, *34* (5), 1219-1224.

(26) Voigt, J.; Bienfait, B.; Wang, S.; Nicklaus, M. Comparison of the NCI open database with seven large chemical structural databases. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (3), 702-712.

(27) OpenEye Science Software: Santa Fe, NM. http://www.eyesopen.com

(28) *Pharmacophore perception, development, and use in drug design*. **1999**. Biotechnology Series, ed. Güner, O. International University Line, La Jolla, CA.

(29) Putta, S.; Lemmen, C.; Beroza, P.; Greene, J. A novel shape-feature based approach for virtual library screening. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (5), 1230-1240.

(30) Jiang, F.; Kim, S. "Soft docking": matching of molecular surface cubes. *J. Mol. Biol.* **1991**, *219* (1), 79-102.

(31) Beautrait, A.; Leroux, V.; Chavent, M.; Ghemtio, L.; Devignes, M.-D.; Smaïl-Tabbone, M.; Cai, W.; Shao, X.; Moreau, G.; Bladon, P.; Yao, J.; Maigret, B. Multiple-step virtual screening using VSM-G: Overview and validation of fast geometrical matching enrichment. *J. Mol. Model.* **2007**, *accepted*.

(32) Mavridis, L.; Hudson, B. D.; Ritchie, D. W. Towards high throughput 3D virtual screening using spherical harmonic molecular surface representations. *J. Chem. Inf. Model.* **2007**, *47* (5), 1787-1796.

ARTICLE #2

"Multiple-step virtual screening using VSM-G: Overview and validation of fast geometrical matching enrichment"

MULTIPLE-STEP VIRTUAL SCREENING USING VSM-G. OVERVIEW AND VALIDATION OF FAST GEOMETRICAL MATCHING ENRICHMENT

Alexandre Beautrait [1], Vincent Leroux [1], Matthieu Chavent [1], Léo Ghemtio [1], Marie-Dominique Devignes [1], Malika Smaïl-Tabbone [1], Wensheng Cai [2], Xuegang Shao [2], Gilles Moreau [3], Peter Bladon [4], Jianhua Yao [5], Bernard Maigret [1,*]

[1] *Nancy Université, LORIA, Groupe ORPAILLEUR, Campus scientifique, BP 239, 54506 Vandœuvre-lès-Nancy Cedex, France.*
[2] *Nankai University, Department of Chemistry, Tianjin 300071, P.R. China.*
[3] *30 Avenue Jean Jaurès, 94220 Charanton, France.*
[4] *Interprobe Chemical Services, Gallowhill House, Larch Avenue, Lenzie Kirkintilloch, Glasgow G66 4HX, Scotland, UK.*
[5] *Shanghai Institute of Organic Chemistry, Laboratory of Computer Chemistry and Chemoinformatics, 354 Fenglin rd, Shanghai 200032, P.R. China.*

* Corresponding author: bernard.maigret@loria.fr, +33 (0)3 54 95 86 08 (phone), +33 (0)3 83 27 56 52 (fax).

ABSTRACT

Numerous methods are available for use as part of a virtual screening strategy, but yet none of those is solely able to guarantee both a level of confidence comparable to experimental screening and a computing efficiency that could drastically cut down the costs of early phase drug discovery campaigns. We present here VSM-G (**V**irtual **S**creening **M**anager for computational **G**rids), a virtual screening platform that combines several structure-based drug design tools. VSM-G aims to be as user-friendly as possible while retaining enough flexibility to accommodate other *in silico* techniques as they are developed.

In order to illustrate VSM-G concepts, we present a proof-of-concept study of a fast geometrical matching method based on spherical harmonics expansions surfaces. This technique is implemented in VSM-G as the first module of a multiple-step sequence tailored for high-throughput experiments. We show that using this protocol, notable enrichment of the input molecular database can be achieved against a specific target, here the LXR nuclear receptor. The benefits, limitations and applicability of such an approach are discussed. Possible improvements of both the geometrical matching technique and its implementation within VSM-G are suggested.

KEYWORDS

multiple-step virtual screening; VSM-G; structure-based drug design; geometrical matching; spherical harmonics surfaces; SHEF; GOLD; molecular database enrichment.

INTRODUCTION

The search for new drugs is time-consuming and expensive [1]; any method that speeds up the process is beneficial. Recently virtual screening (VS) techniques [2] have gained much interest in many drug development strategies [3]. VS has two obvious advantages: the speed with which one can screen a large library of compounds and the small initial capital investment compared to the cost of an in vitro high-throughput screening (HTS) program. The first aim of HTS and VS is to reduce a molecular database to few hit compounds for a protein target. VS is considered successful when, combined or not with HTS, it leads to confirmed hits for a cost lower to that of HTS alone. Research in this area is particularly active and several success stories have been reported [4-7]. Thus it is now widely accepted that VS calculations can complement HTS experiments [8, 9].

VS methods can have two distinct purposes. The first one is the exclusion of a large number of compounds which have little or no activity, leading to a limited set of molecules which are more probable hits [10]; such a method is referred to as a *filter*. In the literature, database filtering against a given target is often referred to as *enrichment* [11, 12]. The second purpose is the identification of a small number of candidates likely to be potent, by ranking input compounds. In all VS filters there is a trade-off between speed and accuracy; filters are optimized for speed. The fastest filters can handle up to a few million molecules, but are notoriously imprecise in reducing this number to less than a thousand while retaining all potential hits. More costly techniques, which can be used in lead optimization strategies, can tackle this problem [13, 14] but not with several million molecules as input and sensible computation times [15]. Therefore VS protocols are often based on a single or a few fast filters, and used prior to experimental screening. However, in that case, VS usage is limited to that of a pre-filter for HTS, reducing the number of compounds to be tested experimentally and hence the cost of experiments by at least one order of magnitude [16, 17].

We have devised a platform for virtual screening, called VSM-G (Virtual Screening Manager for computational Grids). Our objective with VSM-G is to provide a user-friendly tool that would give scientists a large range of *in silico* strategies for finding hits. Two kinds of approaches can be employed here – ligand-based and structure-based [18, 19]. At present VSM-G uses structure-based methods to rank input compounds according to their affinity for a target. Thus it can prioritize them for experimental testing. Ligand-based modules, such as substructure search, can be involved as pre-processing steps to screen molecular databases and reduce the number of compounds to be considered subsequently. This initial operation can precede the central element of the platform, the *screening funnel*, a multi-step structure-based filtering process that hierarchically combines several docking methods.

After describing the VSM-G platform, we will present a proof-of-concept study in the filtering/enrichment context using the liver-X-receptor β (LXRβ) as a target for a screening calculation against a diverse ligand database. The VSM-G screening funnel was used, consisting of a fast geometrical matching filter preceding flexible docking. This approach is compared to using flexible docking alone for VS. The benefits and limitations of geometrical matching as part of the screening funnel approach, in terms of computing efficiency, applicability and relevance, are discussed.

OVERVIEW OF THE VSM-G PLATFORM

Aims and scope of VSM-G

The first step of the pre-clinical drug discovery process can be simplified as a work of exploration at the intersection of distinct spaces [20]. The first of these is the proteome, whose exploration in the drug design context involves its restriction to the sub-space of proteins whose interactions could be significant therapeutically as novel targets – the *target space*. The second space starts from the even larger ensemble of synthesizable small chemical structures. The exploration here involves sorting out molecules with no or unwanted biological effects, restraining the chemical space [21] to the so-called *drug space* [22]. Eventually, merging the target space with the drug space leads to a third ensemble of receptor-ligand associations that have to be explored successfully in order to solve the drug discovery problem. Provided that the ensembles of targets and candidate molecules have been previously reduced efficiently to avoid a combinatorial explosion, this is still a long and arduous process.

VSM-G rationalizes these searches. It focuses on the exploitation and management of current knowledge of the proteome-to-target and chemical-to-drug steps. It also relies on a specific protocol relying on structure-based virtual screening methods regarding the final ligand-to-hit process. Its workflow has been designed so to match the processes described above and is summarized in **figure 1**. The basic organization of the platform is therefore divided in three distinct parts: two for the preparation of input data (ligands and protein targets respectively) and a third one which is a multi-layer funnel for the *in silico* screening.

Current status

The key features of VSM-G are as follows:

1. Wide coverage of the VS process, from ligand and target preparation to the screening setup, the monitoring of the calculation processes and finally the results' analysis.

2. Unified and user-friendly graphical interface (see **figure 2**). Seamless integration of the modules, *e.g.* intercommunication procedures, such as file format conversions, are automatic and transparent to the user.

3. Easy maintenance of the code, with modular design and choice of widely used programming languages (Java, C, C++ and Fortran).

4. Access to grid technology to take advantage of distributed computing involving computer- and cluster-grids.

5. VSM-G relies on third-party software for performing specific tasks, or in order to provide several choices of techniques for a given purpose. Due to its modular design, VSM-G is readily useable even if those external programs are not installed on the host computer. One of the main development goals of VSM-G is to provide at least one free, open-source solution for each task, which is not currently the case (*e.g.* at the moment GOLD is the only choice for performing flexible docking).

The VSM-G features regarding the ligand database preparation and its target-related capabilities are listed on **chart 1** and **chart 2** respectively. Current development is mostly concentrated on the screening funnel part.

The screening funnel: a multiple-step strategy

Presently a wide variety of virtual screening programs are available, and it is generally assumed that a well-chosen combination of methods will give better results than a single one. The interest for such multiple-step VS protocols has been stressed in various papers, often as a combination of a single structure-based docking calculation with ligand-based approaches as pre-filters [5, 23]. Post-processing refinements starting from docking results have also been reviewed [15, 24]. Alternatively, several methods can be employed at different stages within a given docking program [25]. The use of several docking programs in the same protocol [26] is less frequent. Moreover, most programs require significant expertise in setting up and analyzing the results. More generally each technique features a specific

balance between the speed of calculations and the reliability of results [27]. Open software tools overcoming such limitations are lacking. The virtual screening implementation in the VSM-G platform is constituted by a series of different structure-based methods, organized sequentially in a funnel strategy. The techniques range from simple methods to more sophisticated ones, profiting from the speed of the former and the accuracy of the latter. At each step of the process, the filter discards inappropriate compounds. The most simple and quick filters are being used at an early stage in the filtering process, allowing the more time consuming processes to be used in later stages. The multiple-step screening funnel strategy is shown in **figure 3**.

METHODOLOGY

Outline of the proof-of-concept study

Most docking methods are not efficient enough for use in high-throughput VS (*i.e.* the time required to process $>10^6$ molecules is out of reach with modern hardware). Fast filtering prior to docking might be a workaround. Ligand-based methods can also prove useful here, but unless large training sets are available for the target, they are of limited value. Geometrical matching procedures, which are orders of magnitude faster than common docking methods, can be employed in this particular context [28], and can lead to discovery of hits [29], but few studies exist to estimate their impact in a general VS experiment.

The geometrical matching procedure evaluated here is a two-step process. First, the MSSH program [30, 31] approximates the geometry of molecular structures using a series of spherical harmonics functions. This representation is very compact as all information is contained in the expansion coefficients, while the corresponding surfaces still provide a good level of detail. Additionally, this process can be done once and for all for each protein and ligand conformer. Afterwards, evaluating the surface complementarity between a target active site and a ligand is performed through simple and efficient operations [32] specific to spherical harmonics algebra. This very fast procedure is performed with the SHEF program [33], which identifies and scores the geometrically-optimal orientation of each ligand conformation for the target. These techniques are described in depth by Cai *et al.* [30, 31, 33]

In this paper we study a VSM-G-operated screening funnel using MSSH/SHEF followed by flexible docking using GOLD [34, 35]. Such an approach involves using SHEF results to filter out part of the input ligand database before proceeding to the second funnel step relying on GOLD. In this proof-of-concept study we did no such filtering; all molecules of the test set are evaluated with both techniques in order to simulate the screening funnel for all levels of filtering between the two steps.

Target preparation

The liver X receptors (LXRs) [36] represent attractive targets for the development of new therapeutic agents for treating multiple (especially cardiovascular) diseases [37]. Several structures of the ligand binding domain of LXR, co-crystallized with various ligands, have been determined by X-Ray crystallography. Reports on structural analysis reveal great plasticity of the ligand binding pocket; it is able to accommodate ligands with noticeably different shapes and sizes [38]. In this work we study more specifically the LXRβ isoform, for which we took as a starting point different X-ray structures available from the Protein DataBank (PDB) [39]: 1P8D [40], 1PQ6 [38] and 1PQ9 [38]. For each of these structures the most complete chain was retained: chain A for 1P8D and chain B for 1PQ6 and 1PQ9. In all cases the binding area was complete and the Cα trace superimposed well, allowing missing fragments to be added using homology modeling. Protonation was performed at pH 7 with VSM-G. The imidazole tautomer of the active site histidine residue is the N^{d1}-H one [41].

Figure 4 shows that the three binding site conformations, represented by their MSSH-generated surfaces imported into VMD [42], are clearly distinct geometrically. The 1PQ9 cavity is significantly smaller (810 $Å^3$) than 1PQ6 (996 $Å^3$) and 1P8D (1014 $Å^3$). 1PQ6 has a less-spherical, more specific shape. Therefore, it could be expected that (i) 1P8D is the least selective upon ligand binding, (ii) 1PQ6 shape specificity could be overcome by ligand flexibility, and (iii) the 1PQ9 conformation should filter out more structures based on their size.

The protein-ligand binding modes depicted in the three experimental structures have also been analyzed. The shared characteristics are dominated by hydrophobic interactions with F_{271}, F_{329} and F_{340}. 1PQ6 allows for a possible specific charge-charge interaction with R_{319}. R_{319} already makes an internal interaction with E_{281} in the 1P8D conformation, dampening the strength of possible ligand interaction. In the case of 1PQ9 neither of those residues is accessible as the pocket size is restricted by a particular F_{329} orientation.

Ligand database preparation

The starting database is composed by compounds commercially available in March 2006 from three suppliers, ChemDiv [43], Enamine [44] and Comgenex [45]. Filtering using Lipinski's rule-of-five [46] was performed, allowing a single violation for each structure, giving a total of 598,375 unique molecules. In order to reduce the database size while retaining as many chemical diversity as possible, we used the ScreeningAssistant software [47]. This tool characterizes each molecule of the database using SSKey-3D 54-bit fingerprints [48], allowing for similarity estimation between pairs by computing

Tanimoto coefficients [49]. Database clustering can then govern the generation of diversity-maximized subsets. In our case, we targeted a 10,000 molecule subset and obtained a database of 8,383 compounds.

A reference diverse database was defined by merging the initial 598,375 molecules database with the Chimiothèque Nationale (CN) database [50, 51]. Diversity of each of the three subsets (the 598,375 database, the 8,383 diversity set and the 31,220 CN) was measured as fractions of the total diversity [47]. Results are depicted in **table 1**. It appears that the 8,383 subset and the larger CN database are of comparable diversity. The former is therefore suitable as input data for a VS validation experiment. Interestingly, from the large scale database to the diversity subset we only traded ~40% of the diversity for a 98.6% size reduction.

The 8,383-compounds database was pre-processed into VSM-G ligand preparation modules, which made it suitable for the docking programs used afterwards. Molecules were first converted into 3D and then their protonation state was set arbitrarily at pH = 7. As MSSH/SHEF is a rigid shape-matching procedure, a conformational search was performed (retaining at most 400 conformers per compound), giving 1,102,299 conformers.

Parameterization of the virtual screening programs

1,102,299 conformers were docked using SHEF in the three target conformations, giving a total of 3,306,897 rigid docking calculations. Using GOLD, 8,383 molecules were docked, giving 25,149 flexible docking calculations. The programs parameters that were used, favoring reliability over speed, are listed in **chart 3**.

Definition and relevance of reference data

The reference data for evaluating SHEF performance is constituted by GOLD results and not by experimental data. Like all docking programs, GOLD does not provide 100% success in reproducing conformations and binding free energies of protein-ligand complexes [52]. Hence the reference set is approximate and cannot be used to measure SHEF performance precisely. However, our aim here is simply to demonstrate SHEF usefulness as part of the VSM-G screening funnel, in a large-scale VS context. Consequently, a reference set large enough statistically and chemically diverse seems appropriate despite GOLD-related limitations.

In order to evaluate filtering, the reference molecular database has to be divided in two subsets, the first corresponding to the (presumably) most potent molecules (referred to as the *hit compounds* subset) that shall be conserved upon filtering, and the second subset being considered as the inactive structures for the target. The GOLD score values are being used to rank ligands against the three target conformations, and the 10% best-ranked ligands are selected from each of the three sets. This cutoff value is set arbitrarily. Ranks are being used to select ligands instead of the score values because molecular dynamics simulations performed in our laboratory on LXRβ indicate that important induced fit effects [53] could occur upon ligand binding. This suggests the GOLD scoring function, which does not account for the receptor internal energy, may only correlate with the global free energy of binding across a single receptor conformer [54].

As shown in **figure 5**, the three ensembles of 838 selected structures overlap, giving a classification of *hits* into different families regarding their selectivity for the three target conformations. Out of a total of 1,414 molecules, 670 (47%) bind specifically on one of the three conformations, 356 (25%) bind on all the three conformations, the rest binding on two out of three. The amount of selective molecules on each conformation is 20%, 24%, and 36% for 1P8D, 1PQ6 and 1PQ9 respectively, which is in agreement with the structural specificities highlighted previously.

Analysis of results

An in-house program was created for representing relationships between the screening results of two different techniques for the same set of input data. **Figure 6** explains the principles of the generated graphical representation. Both ranks ranges are divided in twenty 5% blocks, a sensible trade-off between graphics clarity and the amount of represented information. Three particular cases are provided as examples. **Figure 6.a** depicts random selection, and on the opposite, **figure 6.c** corresponds to a perfect correlation. A given filtering process will obviously have results between these two. **Figure 6.b** is another ideal case for filtering, but only for a precise filtering amount (which may or may be not satisfactory).

The Spearman [55] ρ and Kendall [56] τ coefficients are employed as measures of correlation:

$$\rho = 1 - \frac{6}{n(n^2 - 1)} \sum_{i=1}^{n} \Delta r_i^{\,2}$$

$$\tau = -1 + \frac{4}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \delta(r_j > r_i)$$

r_i is the SHEF ranking of the i^{th}-ranked GOLD structure; Δr_i is the difference between these two ranks ($\Delta r_i = r_i - i$). δ is the boolean function: δ (true) = 1 while δ (false) = 0.

The rankings, in both cases, are in ascending order from the best predicted binding molecule to the worst. We also have $0 \leq \rho \leq 1$ and $-1 \leq \tau \leq 1$, 0 indicating an absence of correlation (random selection) and 1 perfect correlation (same rankings).

Other metrics are used in order to evaluate filtering performance. Given a definition of what is a hit structure and what is not for a specific target, we can describe the *quality* q of a molecular database of n structures as the ratio between the number of hit compounds and the total number of structures:

$$q = \frac{n_{hits}}{n}$$

The *enrichment* e of a database by a filtering process and for a given filtering ratio f ($0 \leq f \leq 1$, f being the amount of *filtered out* candidates) can be defined as the ratio between the quality of the reduced database and the quality of the initial database:

$$e(f) = \frac{q(f)}{q(0)}$$

Enrichment is commonly used to evaluate the efficiency of molecular database method. By definition random selection does not affect quality, so its efficiency is 1 for any filtering amount. The maximum enrichment that can be obtained for a given filtering level is when all hits are retained, which corresponds to:

$$e_{\max}(f) = \frac{1}{1-f}$$

The *filtering efficiency* E is eventually defined as the relative distance of the filtering method from random filtering (E = 0) to maximum enrichment (E = 1):

$$E(f) = \frac{e(f)-1}{e_{\max}(f)-1}$$

RESULTS

Influence of target conformation on GOLD and SHEF results

The density plots of **figure 7** give a picture of how target conformation specificities influence GOLD and SHEF results. The SHEF correlation between 1P8D and 1PQ9 (**figure 7.b**) is greater than those between 1PQ6 and both 1P8D (**figure 7.a**) and 1PQ9 (**figure 7.c**). This is in agreement with the observation that the 1PQ6 shape is the most specific. In the case of GOLD, it first appears that 1P8D and 1PQ6 results are highly correlated (**figure 7.d**). The correlations with 1PQ9 (**figures 7.e and 7.f**) are lower. A significant amount of structures performing well with both 1P8D and 1PQ6 are ranked low with 1PQ9, indicating a group of ligands whose size fits well into the former active site conformations but not in the smaller 1PQ9. Surprisingly, such an expected group does not appear in SHEF results.

Therefore, SHEF, which is a surface-based method, appears more sensitive to the active site shape specificities than GOLD, which relies on a classical atom coordinates-based representation of molecular structures. But in contrast to GOLD, SHEF appears unable to assess size constraints correctly. This could be related not to SHEF itself but rather to its current implementation within the VSM-G screening funnel. Indeed, only the best conformer score is retained for ranking each compound: the diversity of geometrically-acceptable conformations (referred to as *adaptability*) is not taken into account. This could lead to SHEF producing false positives with ligands occupying almost all the active site volume. These ligands might require a minimal adaptability in order to provide a good chance to satisfy chemical constraints upon binding, in addition to geometrical complementarity.

Relationship between SHEF and GOLD classifications

Figures 7.g, 7.h and 7.i depict the relationships between SHEF and GOLD ranks for 1P8D, 1PQ6 and 1PQ9 respectively. Given the fundamental differences between these two programs, it is not surprising to see lower correlation between SHEF and GOLD than between two different target conformations for either SHEF or GOLD. We are, however, far from the random case depicted in **figure 6.a**, thus it is clear that noticeable enrichment using SHEF is already observed at this point.

If the general profile of the three density plots is similar, they differ regarding the distributions of false positives, *i.e.* populations located at the bottom right corners, corresponding to molecules whose binding ranks are overestimated by SHEF according to GOLD results. In agreement with previous observations, it appears that SHEF generates most false positives when docking on the 1PQ9 conformation, while correlation between GOLD and SHEF is best in the 1PQ6 case, which presents a more specific shape that should favor SHEF efficiency.

Interestingly, **figure 7.j** shows that the correlation between the SHEF and GOLD consensus rankings is higher than the average of the GOLD-SHEF correlation for the three receptor conformations. Additionally, such an approach could be more interesting than the 1PQ6-only filtering of **figure 7.h**, which naturally favors ligands more specific to 1PQ6. Even if the corresponding correlation is higher, it is probably more important to favor diversity regarding target conformations when no precise information is known concerning their relative stability.

SHEF as a first-step enrichment filter in the screening funnel protocol

It should first be noted that the ligands present in the 1P8D, 1PQ6 and 1PQ9 experimental structures, redocked using GOLD, fall into the range of the *hits* subset as defined previously. These reference ligands are also amongst the top 2% structures according to SHEF calculations. Therefore, unless the filtering ratio is set too high, they would be retrieved in a SHEF/GOLD screening funnel experiment.

Taking as reference the SHEF consensus ranking, we plotted the variation of the population of GOLD *hits* as a function of the filtering ratio. The resulting curve is shown in **figure 8** together with the enrichment curves that would result from random selection and from the ideal case where the 1,414 *hits* are all ranked before the other 6,969 molecules. A clear enrichment is observed on all ranges of filtering. There is still much room for improvement, but present SHEF performance is interesting considering that SHEF and GOLD are not in the same league in terms of speed and precision. In the virtual screening context, if the number of molecules to screen is too high for GOLD using available computing power, SHEF could provide a rational solution for decreasing the number of candidate molecules without limiting too much the chances of finding novel hit compounds for a given target.

Correlation between SHEF efficiency and the nature of the protein-ligand binding mode

We will now focus on results for two particular filtering ratios, chosen arbitrarily: 0.1 (low filtering, 90% of molecules retained) and 0.5 (half the molecules filtered out). In order to determine whether particular families of molecules could influence SHEF filtering efficiency, the variation of all *hits* populations as defined in **figure 5** for the four possible SHEF rankings (on the 1P8D, 1PQ6, 1PQ9 targets, and multiple-target consensus) was collected. The results are shown in **table 2**. This data was translated in terms of filtering efficiency E(f) in **table 3**. The main result can be interpreted as follows: if we apply respectively 10% and 50% filtering using the multiple-target SHEF filter, amongst all *hits* we will retain respectively 90.8% and 52.8% of what would have been lost using random selection.

The comparison between the four available filters based on SHEF rankings suggests that the use of the multiple-target consensus ranking should be the best choice. This is in agreement with the observations made analyzing **figures 7.g, 7.h and 7.i**. More interestingly, analysis of SHEF efficiency of the different *hits* sub-groups reveals that molecules specific to the 1PQ6 target conformation according to GOLD are performing poorly with SHEF (see **table 3**, ”1PQ6-specific” line). It has been shown that the specific 1PQ6 shape is taken into account by SHEF, but 1PQ6 also presents a second particularity: the accessibility of a charged residue. The corresponding 1PQ6-specific ligands most probably share a binding mode dominated by electrostatic effects that SHEF, as it only compares geometries, is unable to assess. Contrarily, the molecules that are defined as *hits* for all of the three LXRβ pocket conformations are those for which SHEF filtering is the most efficient for both values of filtering (see **table 3**, “1P8D+1PQ6+1PQ9” line). These molecules might have a high degree of adaptability, allowing SHEF to perform well in identifying the conformations that have the best steric complementarity.

DISCUSSION AND CONCLUDING REMARKS

In this study, we wanted to present an overview of VSM-G, and then more precisely to evaluate the usefulness of the SHEF geometrical matching procedure as part of the VSM-G multiple-step high-throughput VS procedure. We have chosen, as the reference data, score values from the flexible docking program GOLD. This allows for a qualitative assessment of MSSH/SHEF efficiency as a first fast filter for the VSM-G multiple-step procedure. Thus, even considering the limitations of our validation test, results are clear enough to demonstrate that SHEF, and by extension its association as the first module in the VSM-G screening protocol, can actually be useful for *in silico* drug discovery.

This paper has highlighted precisely the conditions for obtaining good performance from MSSH/SHEF. It appears that for flexible receptors prone to induced fit effects upon complexation, a filtering based on a consensus ranking of SHEF results for multiple target conformers should be favored. More importantly, basic information regarding the types of interactions involved in ligand binding is crucial for deciding if MSSH/SHEF should be used and if so to what extent. Enrichment can only be expected when binding is not largely dominated by chemical interactions such as electrostatic effects or hydrogen bonding. Active sites that are known to favor hydrophobic interactions might be targets of choice for a structure-based drug design strategy involving MSSH/SHEF as part of a multiple-step VS procedure set up using the VSM-G program.

Limitations of the spherical harmonics-based geometrical matching procedure have been pointed out. As with all structure-based *in silico* techniques, there are two fundamental aspects of how the protein-ligand binding is modeled. Firstly, the way search space is defined, and secondly, how this space is explored. An improvement of SHEF in the first area would involve taking into account basic chemical properties to extend the complementarity score that is currently computed. Such an approach has already been tried out in the ligand-based drug design area [57]. Regarding the exploration strategy, in its current implementation in VSM-G, SHEF acts as a rigid docking program that only selects a single conformer out of a list for a given structure; this approach has been shown here to produce significant numbers of false positives in some cases. An alternative could be to use a diverse set of docked conformers for each ligand, the selection between them being made by a second module in the screening funnel protocol. Various techniques are being considered in this regard [58-60].

In any case, it is uncertain that improvements of the SHEF algorithm would necessary be worthwhile. At the present time the main advantage of the MSSH/SHEF approach is its speed. With the safe parameters used in this report, SHEF is typically 2-3 orders of magnitude faster for processing $>10^6$ conformers than GOLD for docking the corresponding $\sim10^4$ structures. MSSH is still 1 order of magnitude faster than GOLD, and its calculations can be done once and for all for a given molecular database. Enhancements of the MSSH and SHEF programs should obviously not be made at the cost of the loss of such a computing speed advantage that allows for performing large scale structure-based VS.

In further work, we will focus on selection rather than on filtering capability. This will include a proof-of-concept study of the usefulness of post-docking optimizations and molecular dynamics calculations as funnel modules following geometrical matching and flexible docking. Next, we will illustrate the whole screening funnel strategy through an actual large scale hit discovery campaign using computer grid architectures. The relevance of using advanced techniques like target sampling and grid computations in such a context will also be highlighted.

ACKNOWLEDGMENTS

We thank Yesmine Asses, Safia Kellou and Amel Maouche for their feedback. Alexandre Beautrait was supported by grants from INRIA (Institut National de Recherche en Informatique et en Automatique), Région Lorraine, and ARC (Association pour la Recherche sur le Cancer); Vincent Leroux by a post-doctoral fellowship from the INCa (Institut National du Cancer); Matthieu Chavent by a joined fellowship between CNRS (Centre National pour la Recherche Scientifique) and Région Lorraine. We thank Openeye for providing free access to OMEGA and VIDA software according to an academic license, Chemaxon for supplying MarvinBeans Java library, CCDC for the trial version of the GOLD program, and the laboratory of chemoinformatics at the Orléans University for the ScreeningAssistant program.

FIGURES

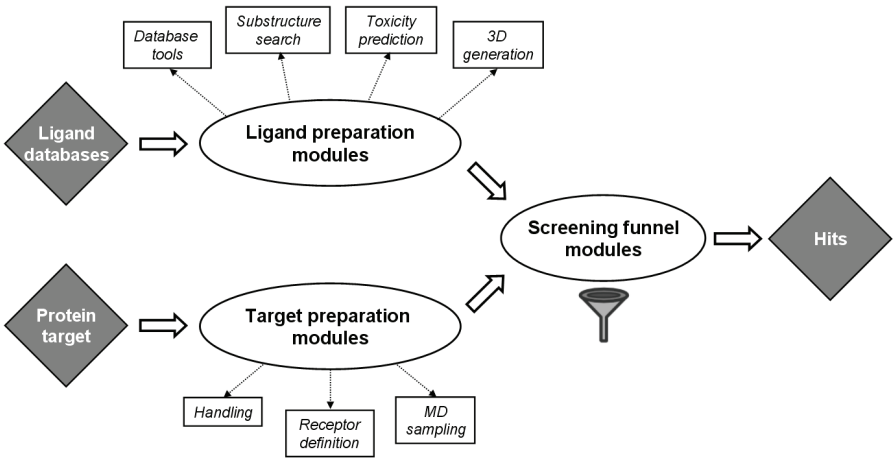


Figure 1 - Basic VSM-G workflow for hit discovery.

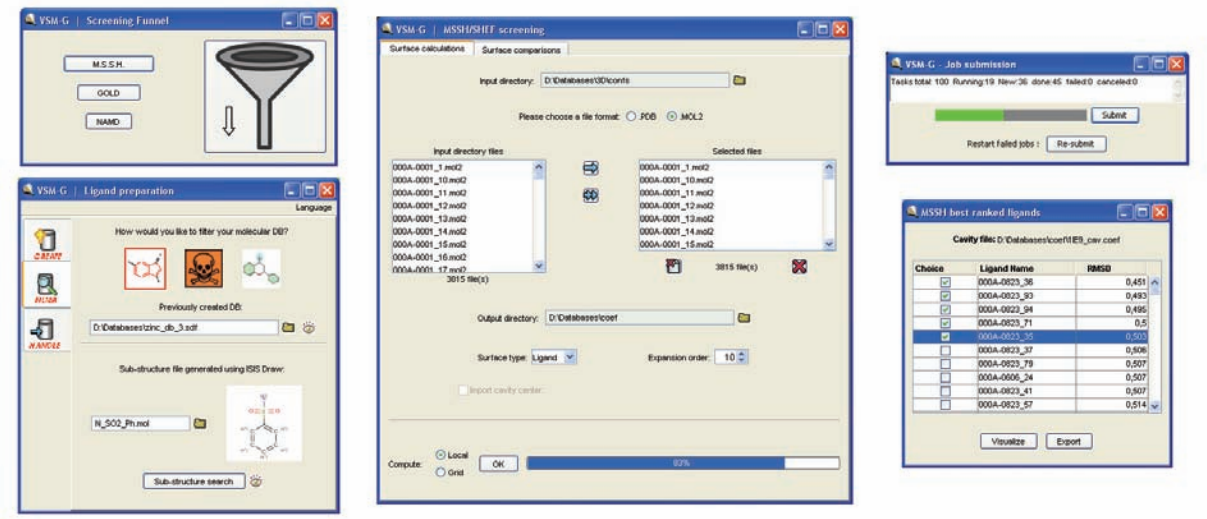


Figure 2 - Some screenshots of the VSM-G graphical interface.

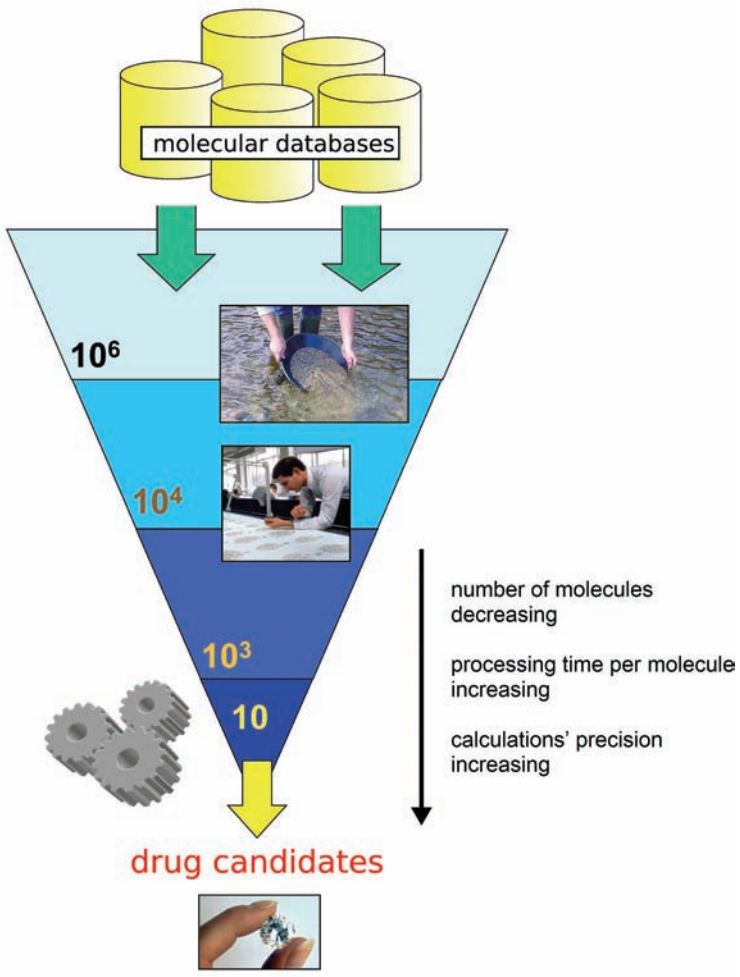


Figure 3 - Basic principle of the virtual screening funnel process.

.

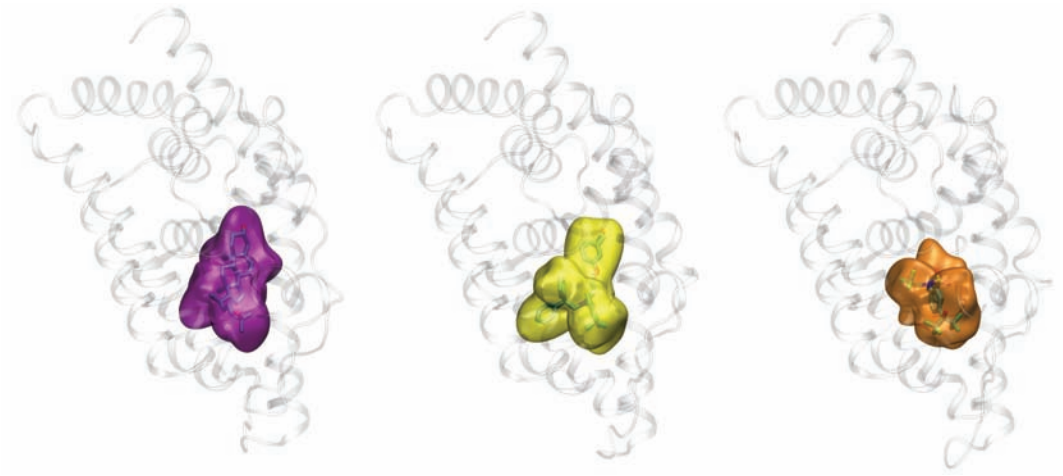


Figure 4 - Shapes of the 1P8D, 1PQ6 and 1PQ9 active sites (from left to right) as approximated by spherical harmonics expansion surfaces using MSSH. The X-ray ligands filling the active sites are shown

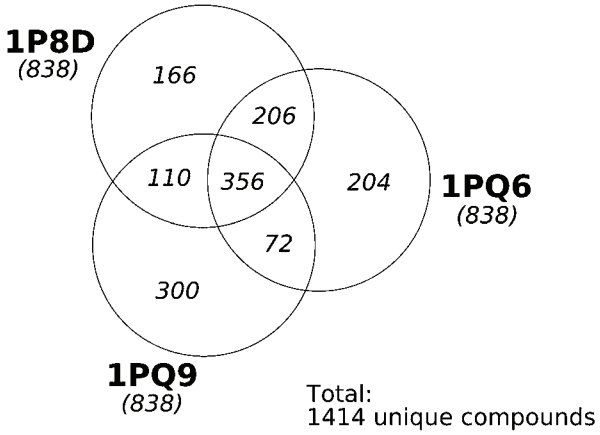


Figure 5 - Populations of hits defined from GOLD results of the 8,383-compounds diverse database. For each target conformation (1P8D, 1PQ6 and 1PQ9) the top-scoring 10% structures are defined as hits. The overlapping of these three sets is represented. There are a total of 1,414 hit compounds that is defined as the target subset that has to be conserved through the filtering process.

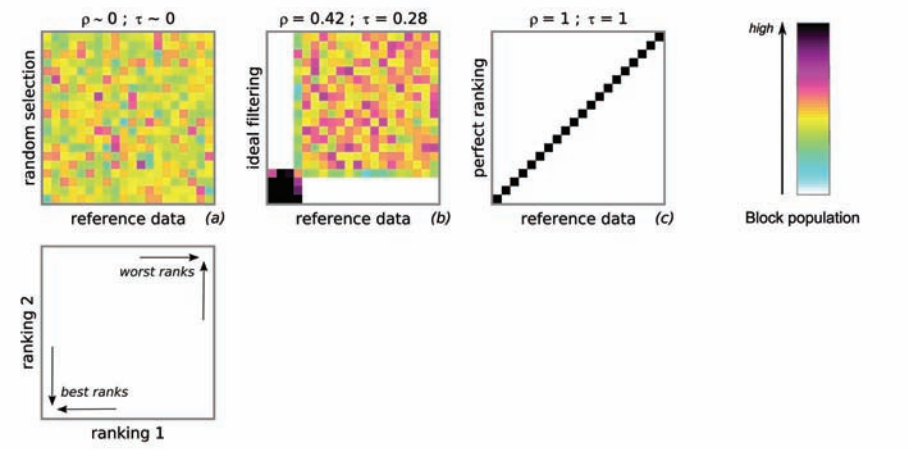


Figure 6 - Explanation of the density plots representation for ranks correlation. Three particular cases are exemplified: (a) random selection, (b) ideal filtering, (c) perfect correlation.

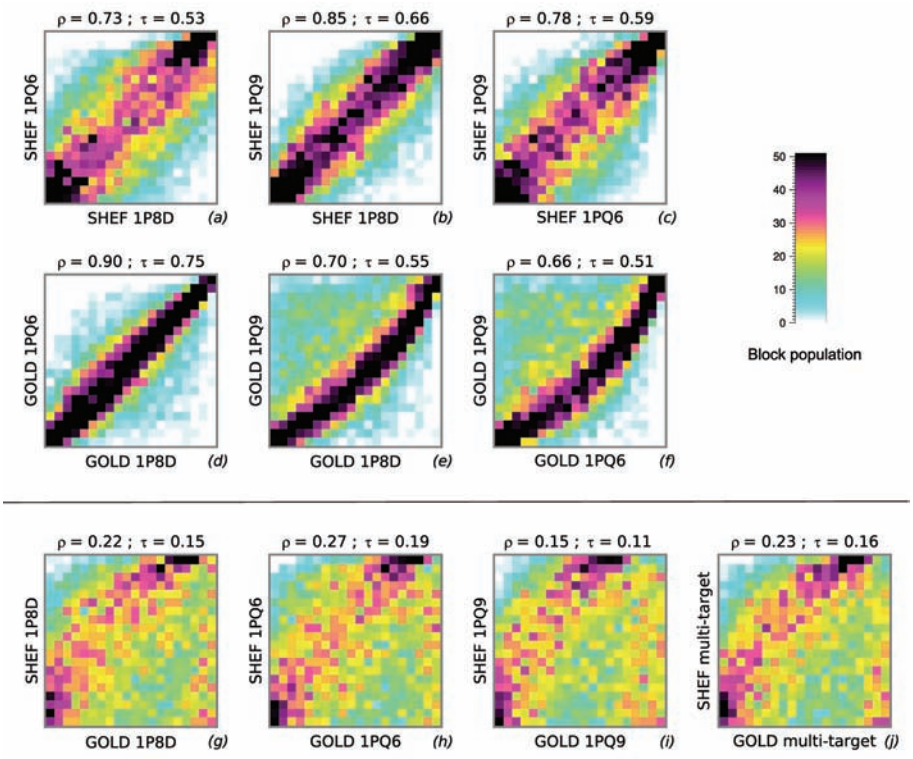


Figure 7 - Density plots between rankings. The 6 first plots (a, b, c, d, e, f) depict the relationships between the different target conformations, for SHEF (a, b, c) and GOLD (d, e, f). Target conformation influence on these two programs can therefore be observed. The 4 last plots (g, h, i, j) show the relationship between SHEF and GOLD results, for the three target conformations (g, h, i), then using multiple-target rankings (j).

The scale is set so that the average $(5\%)^2$ block density is 8383 / 400 ~ 21. Further explanations on these representations can be found on figure 6.

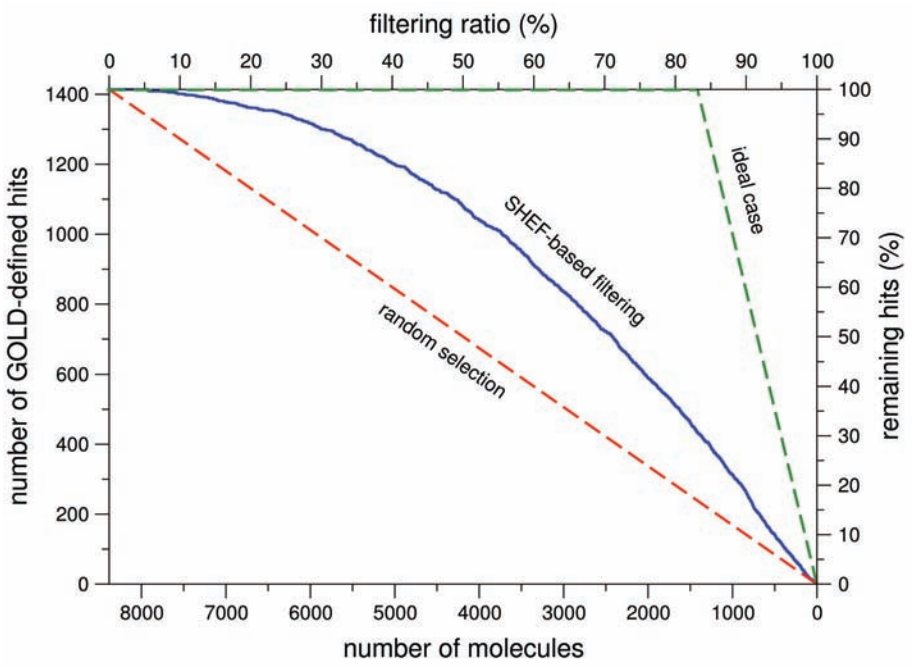


Figure 8 - Enrichment curve of SHEF as measured in the validation experiment, depending on the amount of applied filtering.

Reference results are the GOLD rankings, which were used to define a target subset of 1414 structures (referred to as hits) out of the starting 8383. The multiple-target rankings were used in both cases (i.e. the rank of each molecule is the best rank amongst the 1P8D, 1PQ6 and 1PQ9 classifications).

The two dotted curves represent random selection and perfect correlation (in which SHEF would reproduce GOLD results perfectly); thus the filtering efficiency E for a filtering ratio f can be measured as the relative y position of the SHEF enrichment curve between these two.

CHARTS

Chart 1

Current VSM-G features: ligand database preparation.

Database creation and handling

- generation of virtual combinatorial libraries from chemical scaffolds and fragments
- merging of molecular files, with detection of duplicate structures
- support for different file formats, the most popular SDF [61] and MOL2 [62] as output
- conversion between formats using in-house code or OpenBabel [63]
- implementation of the MarvinBeans library [64] and VIDA [65] for database browsing (if available)

Substructure search

- flexible criteria through combinations of simple operators (and, or, not, have, at least, at most…)
- support for SMILES [66], SDF and RDF [61] as input
- internal use of a canonical topology coding that greatly reduces the complexity of the requests
- quickly searches through millions of compounds on desktop computers once the coding is performed

Toxicity prediction

- implementation of PCT [67], a carcinogenicity prediction program based on SAR
- exclusion of presumably toxic compounds
- possible enrichment of the database of substructures associated with poor chemical stability or toxicity

3D structure generation

- fragment-based 3D structure generation program
- the fragment database (> 10,000 structures) can be enhanced / extended by the user
- CORINA [68], which shares the same concept, can be used alternatively (if available)
- post-processing options: protonation (at pH = 7); conformational sampling using OMEGA [65] (if available)

Chart 2

Current VSM-G features: target preparation.

Handling of protein structures

- automatic checking and cleaning of input PDB files with respect to PDB standards [69]
- protein structures can be checked using the MOLPROBITY server [70]
- correction of protonation states: link to the H++ web server [71]
- relaxation of the hydrogen positions upon energy minimization
- link to the STING [72] web-based suite of programs for data mining

Receptor definition

- holoproteins: receptor assumed to be located at the center of mass of the ligand
- apoproteins: generation of an interactive protein 2D map with MSSH [30, 31] and VMD [42] for picking up surface receptors
- manual definition of receptors can be imported from VMD selections, and exported to funnel modules
- handling of resident water molecules, potentially useful with some docking programs [73]

Multiple target conformations management

- handling of multiple X-ray structures
- enrichment through MD sampling [15, 74], using VMD and NAMD [75]
- clustering, averaging and minimization of conformations from NMR data or MD sampling

Chart 3

Parameters for MSSH, SHEF and GOLD used for the validation study simulating the use of MSSH/SHEF for filtering prior to GOLD calculations.

MSSH [30, 31] / SHEF [33]

- spherical harmonics expansion of order 10
- cavity coordinates defined using the ligand center of mass

GOLD [35]

- default genetic algorithm parameters
- 50 dockings / molecule
- early termination option: docking stopped if the top 5 conformations fall within 1.5 Å RMSD range
- cavity definition: flood fill (works well when the receptor is not open and extended)
- same cavity coordinates as with MSSH/SHEF
- scoring function: GoldScore

TABLES

Database	number of compounds	drug-like compounds	lead-like compounds	drug-like diversity	lead-like diversity	**global diversity**
large-scale	598,327	563,777 (94.2%)	195,332 (32.6%)	84.3%	82.3%	81.8%
diversity subset	8,383	7,875 (93.9%)	3,178 (37.9%)	50.0%	43.5%	48.3%
CN	31,220	27,403 (87.8%)	20,295 (65%)	41.4%	44.8%	43.7%

Table 1 - Diversity analysis of the reference database used in this paper, here referred to as the diversity subset of 8,383 compounds. In the table, 100% diversity is that of the union of the large-scale and CN databases. All values are computed by the ScreeningAssistant software. Please refer to Monge et al. for details on how drug-like and lead-like compounds are defined, and how molecular database diversity is measured.

Table contents : population of the different GOLD *hit* groups after SHEF filtering		initial population	SHEF-based filters							
			1P8D		1PQ6		1PQ9		**multiple-target**	
			10%	50%	10%	50%	10%	50%	10%	50%
GOLD-based *hit* groups	1P8D	838	834	672	835	702	827	652	832	688
	1PQ6	838	828	620	833	664	817	591	826	644
	1PQ9	838	832	632	837	707	835	660	838	687
	1P8D-specific	166	165	130	165	125	164	117	165	130
	1PQ6-specific	204	198	116	201	125	192	96	197	121
	1PQ9-specific	300	295	194	299	222	297	212	300	213
	1P8D+1PQ6	206	203	156	204	151	197	137	201	142
	1P8D+1PQ9	110	110	90	110	97	110	90	110	93
	1PQ6+1PQ9	72	71	52	72	59	72	50	72	58
	1P8D+1PQ6+1PQ9	356	356	296	356	329	356	308	356	323
	all hits	1414	1398	1034	1407	1108	1388	1010	**1401**	**1080**

Table 2 - Evolution of the GOLD hits subsets population when applying 10% and 50% SHEF-based filtering. The groups defined on figure 6 are studied separately, while regarding SHEF filtering, the results for each of the 3 target conformations are presented as well as those using the multiple-target consensus ranking. Note: the multiple-target / all hits results (bottom right) can be measured directly on the figure 8 curve.

Table contents : SHEF filtering efficiency (%)		SHEF-based filters							
		1P8D		1PQ6		1PQ9		**multiple-target**	
		10%	50%	10%	50%	10%	50%	10%	50%
GOLD-based *hit* groups	1P8D	95.2	60.4	96.4	67.5	86.9	55.6	92.8	64.2
	1PQ6	88.1	48.0	94.0	58.5	74.9	41.1	85.7	53.7
	1PQ9	92.8	50.8	98.8	68.7	96.4	57.5	100	64.0
	1P8D-specific	94.0	56.6	94.0	50.6	88.0	41.0	94.0	56.6
	1PQ6-specific	70.6	13.7	85.3	22.5	41.2	-5.9	65.7	18.6
	1PQ9-specific	83.3	29.3	96.7	48.0	90.0	41.3	100	42.0
	1P8D+1PQ6	85.4	51.5	90.3	46.6	56.3	33.0	75.7	37.9
	1P8D+1PQ9	100	63.6	100	76.4	100	63.4	100	69.1
	1PQ6+1PQ9	86.1	44.4	100	63.9	100	38.9	100	61.1
	1P8D+1PQ6+1PQ9	100	66.3	100	84.8	100	73.0	100	81.5
	all hits	88.7	46.3	95.0	56.7	81.6	42.9	**90.8**	**52.8**

Table 3 - Values of the SHEF filtering efficiency E(f) for f = 10% and f = 50%. These values are directly correlated to those of table 2.

REFERENCES

(1) DiMasi, J. A.; Hansen, R. W.; Grabowski, H. G. The price of innovation: new estimates of drug development costs. *J. Health. Econ.* **2003**, *22*, 151-185.

(2) Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **2004**, *432*, 862-865.

(3) Stahura, F. L.; Bajorath, J. Virtual screening methods that complement HTS. *Comb. Chem. High Throughput Screening* **2004**, *7* (4), 259-269.

(4) Perola, E.; Xu, K.; Kollmeyer, T. M.; Kaufmann, S. H.; Prendergast, F. G.; Pang, Y. P. Successful virtual screening of a chemical database for farnesyltransferase inhibitor leads. *J. Med. Chem.* **2000**, *43* (3), 401-408.

(5) Grüneberg, S.; Stubbs, M. T.; Klebe, G. Successful virtual screening for novel inhibitors of human carbonic anhydrase: Strategy and experimental confirmation. *J. Med. Chem.* **2002**, *45*, 3588-3602.

(6) Vangrevelinghe, E.; Zimmermann, K.; Schoepfer, J.; Portmann, R.; Fabbro, D.; Furet, P. Discovery of a potent and selective protein kinase CK2 inhibitor by high-throughput docking. *J. Med. Chem.* **2003**, *46* (13), 2656-2662.

(7) Kraemer, O.; Hazemann, I.; Podjarny, A. D.; Klebe, G. Virtual screening for inhibitors of human aldose reductase. *Proteins: Struct., Funct., Bioinf.* **2004**, *55*, 814-823.

(8) Doman, T. N.; McGovern, S. L.; Witherbee, B. J.; Kasten, T. P.; Kurumbail, R.; Stallings, W. C.; Conolly, D. T.; Shoichet, B. K. Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. *J. Med. Chem.* **2002**, *45*, 2213-2221.

(9) Bajorath, J. Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discovery* **2002**, *1*, 882-894.

(10) Abagyan, R.; Totrov, M. High-throughput docking and lead generation. *Curr. Opin. Chem. Biol.* **2001**, *5*, 375-382.

(11) Xu, H.; Agrafiotis, D. K. Retrospect and prospect of virtual screening in drug discovery. *Curr. Top. Med. Chem.* **2002**, *2*, 1305-1320.

(12) Krovat, E. M.; Langer, T. Impact of scoring functions on enrichment in docking-based virtual screening: An application study on renin inhibitors. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (3), 1123-1129.

(13) Huo, S.; Wang, J.; Cieplak, P.; Kollman, P. A.; Kuntz, I. D. Molecular dynamics and free energy analyses of Cathepsin D-inhibitor interactions: Insight into structure-based ligand design. *J. Med. Chem.* **2002**, *45* (7), 1412-1419.

(14) Jenwitheesuk, E.; Samudrala, R. Improved prediction of HIV-1 protease inhibitor binding energies by molecular dynamics simulations. *BMC Struct. Biol.* **2003**, *3*.

(15) Alonso, H.; Bliznyuk, A. A.; Gready, J. E. Combining docking and molecular dynamic simulations in drug design. *Med. Res. Rev.* **2006**, *26* (5), 531-568.

(16) Waszkowycz, B.; Perkins, T. D. J.; Sykes, R. A.; Li, J. Large-scale virtual screening for discovering leads in the postgenomic era. *IBM Syst. J.* **2001**, *40* (2), 360-376.

(17) Bleicher, K. H.; Böhm, H.-J.; Müller, K.; Alanine, A. I. Hit and lead generation: beyond high-throughput screening. *Nat. Rev. Drug Discovery* **2003**, *2* (5), 369-378.

(18) Veselovsky, A. V.; Ivanov, A. S. Strategy of computer-aided drug design. *Curr. Drug Targets: Infect. Disord.* **2003**, *3* (1), 33-40.

(19) Jain, A. N. Virtual screening in lead discovery and optimization. *Curr. Opin. Drug Discovery Dev.* **2004**, *7* (4), 396-403.

(20) Ofran, Y.; Punta, M.; Schneider, R.; Rost, B. Beyond annotation transfer by homology: novel protein-function prediction methods to assist drug discovery. *Drug Discov. Today* **2005**, *10* (21), 1475-1482.

(21) Dobson, C. M. Chemical space and biology. *Nature* **2004**, *432*, 824-828.

(22) Oprea, T. I.; Gottfries, J. Chemography: The art of navigating in chemical space. *J. Comb. Chem.* **2001**, *3* (2), 157-166.

(23) So, S.-S.; Karplus, M. Evaluation of designed ligands by a multiple screening method: Application to glycogen phosphorylase inhibitors constructed with a variety of approaches. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 613-647.

(24) Lyne, P. D. Structure-based virtual screening: an overview. *Drug Discov. Today* **2002**, *7* (20), 1047-1055.

(25) Wang, J.; Kollman, P. A.; Kuntz, I. D. Flexible ligand docking: A multistep strategy approach. *Proteins: Struct., Funct., Genet.* **1999**, *36* (1), 1-19.

(26) Miteva, M. A.; Lee, W. H.; Montes, M. O.; Villoutreix, B. O. Fast structure-based virtual ligand screening combining FRED, DOCK, and Surflex. *J. Med. Chem.* **2005**, *48*, 6012-6022.

(27) Leroux, V.; Maigret, B. Should structure-based virtual screening techniques be used more extensively in modern drug discovery? *Computers and Applied Chemistry* **2007**, *24* (1), 1-10.

(28) Yamagishi, M. E. B.; Martins, N. F.; Neshich, G.; Cai, W.; Shao, X.; Beautrait, A.; Maigret, B. A fast surface-matching procedure for protein-ligand docking. *J. Mol. Model.* **2006**, *12*, 965-972.

(29) Singh, J.; Chuaqui, C. E.; Boriack-Sjodin, P. A.; Lee, W. C.; Pontz, T.; Corbley, M. J.; Cheung, H.-K.; Arduini, R. M.; Mead, J. N.; Newman, M. N.; Papadatos, J. L.; Bowes, S.; Josiah, S.; Ling, L. E. Successful shape-based virtual screening: the discovery of a potent inhibitor of the type I TGFß receptor kinase (TßRI). *Bioorg. Med. Chem. Lett.* **2003**, *13* (24), 4355-4359.

(30) Cai, W.; Zhang, M.; Maigret, B. New approach for representation of molecular surface. *J. Comput. Chem.* **1998**, *19* (16), 1805-1815.

(31) Cai, W.; Shao, X.; Maigret, B. Protein-ligand recognition using spherical harmonic molecular surfaces: towards a fast and efficient filter for large virtual throughput screening. *J. Mol. Graphics Modell.* **2002**, *20* (4), 313-328.

(32) Ritchie, D. W.; Kemp, G. J. L. Fast computation, rotation, and comparison of low resolution spherical harmonic molecular surfaces. *J. Comput. Chem.* **1999**, *20* (4), 383-395.

(33) Cai, W.; Xu, J.; Shao, X.; Leroux, V.; Beautrait, A.; Maigret, B. SHEF: a vHTS geometrical filter using coefficients of spherical harmonics molecular surfaces. *J. Mol. Model.* **2007**, *submitted*.

(34) Jones, G.; Willett, P.; Glen, R. C. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol.* **1995**, *245* (1), 43-43.

(35) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267* (3), 727-748.

(36) Lala, D. S. The liver X receptors. *Curr. Opin. Investig. Drugs* **2005**, *6* (9), 934-943.

(37) Collins, J. L. Therapeutic opportunities for liver X receptor modulators. *Curr. Opin. Drug Discovery Dev.* **2004**, *7* (5), 692-702.

(38) Färnegårdh, M.; Bonn, T.; Sun, S.; Ljunggren, J.; Ahola, H.; Wilhelmsson, A.; Gustafsson, J.-Å.; Carlquist, M. The three-dimensional structure of the liver X receptor ß reveals a flexible ligand-binding pocket that can accommodate fundamentally different ligands. *J. Biol. Chem.* **2003**, *278* (40), 38821-38828.

(39) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235-242.

(40) Williams, S.; Bledsoe, R. K.; Collins, J. L.; Boggs, S.; Lambert, M. H.; Miller, A. B.; Moore, J.; McKee, D. D.; Moore, L.; Nichols, J.; Parks, D.; Watson, M.; Wisely, B.; Willson, T. M. X-ray crystal structure of the liver X receptor beta ligand binding domain: regulation by a histidine-tryptophan switch. *J. Biol. Chem.* **2003**, *278* (29), 27138-27143.

(41) Steiner, T.; Koellner, G. Coexistence of both histidine tautomers in the solid state and stabilisation of the unfavourable Nδ-H form by intramolecular hydrogen bonding:crystalling L-His-Gly hemihydrate. *Chem. Commun. (Cambridge, U.K.)* **1997**, *13*, 1207-1208.

(42) Humphrey, W.; Dalke, A.; Schulten, K. VMD - Visual Molecular Dynamics. *J. Mol. Graphics* **1996**, *14*, 33-38.

(43) ChemDiv - The chemistry of cures. http://www.chemdiv.com

(44) Enamine - Smart chemistry solutions. http://www.enamine.net

(45) Albany Molecular Research - AMRIDirect chemical compound database. http://www.amridirect.com

(46) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23* (1), 3-25.

(47) Monge, A.; Arrault, A.; Marot, C.; Morin-Allory, L. Managing, profiling and analyzing a library of 2.6 million compounds gathered from 32 chemical providers. *Mol. Divers.* **2006**, *10* (3), 389-403.

(48) Xue, L.; Godden, J.; Bajorath, J. Database searching for compounds with similar biological activity using short binary bit string representations of molecules. *J. Chem. Inf. Comput. Sci.* **1999**, *39* (5), 881-886.

(49) Tanimoto, T. T. Non-linear model for a computer assisted medical diagnostic procedure. *Trans. N-Y Acad. Sci.* **1961**, *2* (23), 576-580.

(50) Hibert, M.; Haiech, J. Des gènes aux médicaments : nouveaux défis, nouvelles stratégies. *M. S. Méd. Sci.* **2000**, *16* (12), 1332-1339.

(51) Chimiothèque Nationale. http://chimiotheque-nationale.enscm.fr/

(52) GOLD CCDC/Astex validation test set results. http://www.ccdc.cam.ac.uk/products/life_sciences/validate/gold_validation/

(53) Koshland Jr., D. The key–lock theory and the induced fit theory. *Angew. Chem., Int. Ed. Engl.* **1994**, *33* (23-24), 2375-2378.

(54) Redocking experiments of LXRß reference ligands present in the X-ray structures back up this hypothesis. Using GOLD, the 1PQ6 ligand redocked in the 1PQ6 binding pocket conformation yields a significantly higher score than the 1PQ9 ligand redocked in the 1PQ9 conformation. However, according to experimental data, the 1PQ9 ligand is indeed clearly more potent on LXRß than the 1PQ6 one, further indicating that the protein-ligand interaction could not be the dominant term in the free energy of binding.

(55) Spearman, C. The proof and measurement of associtaion between two things. *Am. J. Psychol.* **1904**, *15* (1), 72-101.

(56) Kendall, M. A new measure of rank correlation. *Biometrika* **1938**, *30* (1-2), 81-89.

(57) Mavridis, L.; Hudson, B. D.; Ritchie, D. W. Toward high throughput 3D virtual screening using spherical harmonic molecular surface representations. *J. Chem. Inf. Model.* **2007**, *47*, 1787-1796.

(58) Massova, I.; Kollman, P. A. Combined molecular mechanical and continuum solvent approach (MM-PBSA/GBSA) to predict ligand binding. *Perspect. Drug Discov. Des.* **2000**, *18* (1), 113-135.

(59) Gilson, M. K.; Zhou, H.-X. Calculation of protein-ligand binding affinities. *Annu. Rev. Biophys. Biomol. Struct.* **2007**, *36*, 21-42.

(60) Marcou, G.; Rognan, D. Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. *J. Chem. Inf. Model.* **2007**, *47* (1), 195-207.

(61) MDL, SD file format. http://www.mdl.com/solutions/white_papers/ctfile_formats.jsp

(62) Tripos, Mol2 file format. http://www.tripos.com/data/support/mol2.pdf

(63) Open Babel project. http://www.openbabel.sourceforge.net

(64) ChemAxon Ltd., Budapest, Hungary. http://www.chemaxon.com/products.html

(65) OpenEye Science Software: Santa Fe, NM. http://www.eyesopen.com

(66) Weininger, D. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28* (1), 31-36.

(67) Liao, Q.; Yao, J. H.; Li, F.; Yuan, S. G.; Doucet, J.-P.; Panaye, A.; Fan, B. T. CISOC-PCST: a predictive system for carcinogenic toxicity. *SAR QSAR Environ. Res.* **2004**, *15* (3), 217-235.

(68) Sadowski, J. From atoms and bonds to three-dimensional atomic coordinates: Automatic model builders. *Chem. Rev.* **1993**, *93*, 2567-2581.

(69) PDB file format. http://www.rcsb.org/pdb/static.do?p=file_formats/pdb/index.html

(70) Davis, I. W.; Leaver-Fay, A.; Chen, V. B.; Block, J. N.; Kapral, G. J.; Wang, X.; Murray, L. W.; Arendall, W. B., III; Snoeyink, J.; Richardson, J. S.; Richardson, D. C. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res.* **2007**, *in the press*.

(71) Gordon, J. C.; Myers, J. B.; Folta, T.; Shoja, V.; Heath, L. S.; Onufriev, A. H++: a server for estimating pKas and adding missing hydrogens to macromolecules. *Nucleic Acids Res.* **2005**, *33* (Web server issue), W368-W371.

(72) Neshich, G.; Mancini, A. L.; Yamagishi, M. E.; Kuser, P. R.; Fileto, R.; Pinto, I. P.; Palandrani, J. F.; Krauchenco, J. N.; Baudet, C.; Montagner, A. J.; Higa, R. H. STING Report: convenient web-based application for graphic and tabular presentations of protein sequence, structure and function descriptors from the STING database. *Nucleic Acids Res.* **2005**, *33* (Database issue), D269-D274.

(73) Verdonk, M. L.; Chessari, G.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Nissink, J. W. M.; Taylor, R. D.; Taylor, R. Modeling water molecules in protein-ligand docking using GOLD. *J. Med. Chem.* **2005**, *48*, 6504-6515.

(74) Wong, C. F.; Kua, J.; Zhang, Y.; Straatsma, T. P.; McCammon, J. A. Molecular docking of balanol to dynamics snapshots of protein kinase A. *Proteins: Struct., Funct., Bioinf.* **2005**, *61* (4), 850-858.

(75) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **2005**, *26* (16), 1781-1802.

ARTICLE #3

"Induced fit in Liver X Receptor beta: A molecular dynamics-based investigation"

INDUCED FIT IN LIVER X RECEPTOR BETA: A MOLECULAR DYNAMICS-BASED INVESTIGATION

Alexandre Beautrait [1,#], Arnaud S. Karaboga [2,#], Michel Souchet [2,*], Bernard Maigret [1,*]

[1] *Nancy Université, groupe ORPAILLEUR, UMR CNRS/UHP 7503, LORIA, Campus scientifique, BP 239, 54506 Vandœuvre-lès-Nancy Cedex, France*
[2] *Laboratoires Fournier, a Solvay Pharmaceuticals company, 50 rue de Dijon, 21121 Daix, France*

[#] These authors participated equally to this work.

[*] Corresponding authors:
Bernard Maigret, email: bernard.maigret@loria.fr - Phone/Fax: +33 (0)3 54 95 86 08 / +33 (0)3 83 27 56 52
Michel Souchet, email: michel.souchet@solvay.com - Phone/Fax: +33 (0)3 80 44 75 31 / +33 (0)3 80 44 76 53

ABSTRACT

Ligand induced fit phenomenon occurring at the ligand binding domain of the Liver X Receptor beta (LXRβ) was investigated by means of molecular dynamics. Reliability of a 4 ns trajectory was tested from two distinct LXRβ crystal complexes 1PQ6B/GW and 1PQ9B/T09 characterized by an open and a closed state of the pocket, respectively. Crossed complexes 1PQ6B/T09 and 1PQ9B/GW were then submitted to the same molecular dynamic conditions which were able to recover LXRβ conformations similar to the original crystallography data. Analysis of "open to closed" and "closed to open" conformational transitions pointed out the dynamic role of critical residues lining the ligand binding pocket involved in the local remodeling upon ligand binding (*e.g.* Phe271, Phe329, Phe340, Arg319, Glu281). Altogether, the present study indicates that the molecular dynamic protocol is a consistent approach for managing LXRβ-related induced fit process. This protocol could therefore be used for refining ligand docking solutions of a structure-based design strategy.

KEYWORDS

induced fit; liver X receptor; molecular dynamics; protein flexibility; structure-based drug design

ABBREVIATIONS

LXR: liver X receptor; MD: molecular dynamics; DBD: DNA binding domain; LBD: ligand binding domain; LBP: ligand binding pocket; DR-4: direct repeat 4; RXR: retinoic X receptor; ECH: 24(S),25-epoxycholesterol ligand; T09: T0901317 ligand; GW: GW3965 ligand; 1PQ6B: chain B of LXRβ from PDB entry 1PQ6; 1PQ9B: chain B of LXRβ from PDB entry 1PQ9; AF-2: activation function 2; SBDD: structure-based drug design.

INTRODUCTION

Liver X receptors LXRα (NR1H3) and LXRβ (NR1H2) are members of the nuclear hormone receptor superfamily of ligand-activated transcription factors [1]. Two LXR isoforms have been identified: LXRα and LXRβ share 77% sequence identity in their DNA binding domain (DBD) and ligand binding domain (LBD) [2] but differ by their tissue distribution [3]. Indeed, LXRα is mainly expressed in liver, intestine, macrophage and foam cells, whereas LXRβ is ubiquitous. Heterodimers of both LXRs with retinoid X receptor (RXR) bind to hormone response elements of the DR-4 type which are direct repeats of two similar hexa-nucleotide half-sites separated by four nucleotides [4].

Monooxidized cholesterol derivatives (oxysterols), such as 22(R)-hydroxycholesterol, 24(S),25-epoxycholesterol (ECH, **Fig. 1**) and 24(S)-hydroxycholesterol are generally accepted as endogenous LXR activators [5, 6]. It is noteworthy that ECH was co-crystallized with LXRβ LBD later on [7]. Investigations of more potent LXR agonists led to the identification of the first generation of selective dual LXRα/β synthetic ligands T0901317 (T09 [8-10], **Fig. 1**) and GW3965 (GW [8], **Fig. 1**). Elucidation of crystallographic structures of LXR LBD in complex with the natural ligand ECH, and the two synthetic ligands T09 and GW have confirmed that LXR LBD adopted the canonical nuclear receptor fold of twelve α-helices [11] (**Fig. 2**) and have shown that agonists can affect the conformation of the ligand-dependent activation function 2 (AF-2) residing in the C-terminal H12 helix [12, 13]. Indeed, upon agonist binding, H12 is packed against the protein and closes the ligand binding pocket (LBP), thereby leading to an active conformation of the receptor allowing co-factors recruitment, a prerequisite step which controls expression of target genes.

These non-steroidal exogenous ligands were used as pharmacological tools to further explore the biology of LXR: they demonstrated its critical importance in the regulation of transcription of multiple genes involved in lipid metabolism[14], but also, more recent studies underlined implication of LXR in inflammation [15] and glucose metabolism [16]. In addition, administration of T09 [17] or GW [18] to mice was shown to have an anti-atherogenic effect. Nevertheless, the main drawback of these LXR activators is the increase in the amount of serum and liver triglycerides [18, 19]. Therefore, dissociating the favorable effects on cholesterol metabolism from the unfavorable effects on fatty acid metabolism represents a challenge in the development of new LXR agonists used for the treatment of atherosclerosis and prevention of cardiovascular diseases.

In order to design LXR agonists with desired biological profiles, a structure-based drug design (SBDD) approach was initiated, and analysis of reported crystallography data was carried out with the aim to deciphering molecular mechanisms underlying ligand-receptor recognition. It appeared that LXR pocket accommodates ligands and that related amino acid side chains move according to the nature of chemical scaffold. Thus, considering this observed induced fit [20] process in our SBDD strategy turned out to be necessary.

The present study reports the use of molecular dynamics for investigating the flexibility of the LXRβ pocket upon binding of two markedly different reference agonists T09 and GW. It highlights precisely molecular mechanisms driving the related induced fit phenomenon.

MATERIALS AND METHODS

Structural analysis

Superimposition of the LBDs complexed with T09 (PDB code: 1PQ9 [8], chain B) and GW (code PDB: 1PQ6 [8], chain B) on those of ECH (PDB code: 1P8D [7], chain A) was carried out with Relibase+ [21, 22]. Related volume of each LBP was calculated using VOIDOO [23]. Since GW induces the largest pocket, 28 residues located within 4.5 Å of this ligand (**Fig. 3**) are considered for subsequent structural analysis of both crystal data and MD simulations of all protein-ligand complexes.

Molecular Dynamics (MD) simulations

A series of MD simulations have been carried out using as starting points the X-Ray coordinates of LXRβ complexed with T09 and GW (1PQ9 and 1PQ6 PDB codes, respectively). In the case of GW, 1PQ6 chain B was selected since it does not contain missing fragments (most complete chain). 1PQ9 chain B of LXRβ/T09 was chosen because of the higher level of resolution (2.1 Å). But due to ligand splitting observed in this complex, we decided to use T09 coordinates from 1PQC [8] chain B complex. Missing fragments of 1PQ9 chain B were reconstructed by homology with 1PQ6 chain B LBD using the Modeller method [24], and consequently all LBD sequences ranged from Leu220 to His460.

In total, four molecular systems were investigated by MD: (i) two are considered as reference systems, as dealing with the original X-ray structures with bound agonists denoted in this report as 1PQ6B/GW and 1PQ9B/T09 respectively (ii) the other two complexes are crossed systems (1PQ6B/T09 and 1PQ9B/GW) in which the two agonists in their original conformation have been inverted. While the placement of the T09 compound within the large 1PQ6B binding pocket is straightforward, it is noteworthy that in the case of GW inserted in 1PQ9B, severe steric clashes occured between the Phe329 sidechain and the phenyl group of GW propoxy phenylacetic acid moiety. In such a situation, as molecular mechanics is not able to separate the imbricate moieties, a slight manual rotation of the Phe329 ring was necessary to get a consistent starting point. This was followed by few steps (100 iterations) of steepest descent energy minimization in which the whole system was considered as fixed except the two phenyl rings in order to keep the system as close as possible to the original 1PQ9B X-ray structure.

Because it was not possible from the X-ray data to determine the protonation state of His435 located within the LBP, we have considered its two tautomeric states, namely $N^{\delta 1}$–H or N^{e2}–H in our calculations.

Reference and crossed systems were then solvated with a 80x80x80 $Å^3$ box of TIP3P explicit water molecules and Na+ ions were added for ensuring the electrostatic neutrality. The different systems encompass a similar number of atoms (~51,000). The MD program NAMD [25] was employed in conjunction with the CHARMM22 forcefield[26] to describe the receptors, the ligands, the ions and the water molecules. The missing parameters for the ligands were added in the original forcefield [27]. Each system was firstly energy minimized (6,400 steps of conjugate gradients), next equilibrated (500 ps MD) and a trajectory of 4 ns was then produced. The MD simulations were carried out in the NPT ensemble: Langevin dynamics and Langevin piston methods were applied to keep the temperature (300 K) and the pressure (1 atm) of the system fixed. The equations of motion were integrated with a 2 fs timestep, using the r-RESPA algorithm [28] to update short- and long-range contributions at different frequencies. Chemical bonds between hydrogen and heavy atoms were constrained to their equilibrium value by means of the SHAKE algorithm [29]. Long-range electrostatic forces were treated using the particle-mesh Ewald approach [30]. Conservation of the secondary structure elements along the MD trajectories were checked by using the Timeline plug-in in VMD [31].

RESULTS AND DISCUSSION

Structural analysis of X-ray data

Examination of the available X-ray crystal structures of LXRβ LBDs complexed with the natural ligand ECH [7] and the two non-steroidal synthetic ligands [8] T09 and GW provided insight into key elements linked to ligand recognition as well as the mechanisms of activation associated with each ligand.

The overall fold of these structures is similar to those previously reported for other nuclear receptors: the LXR LBD, composed of twelve α-helices (H1 to H12) and of a two-strands anti-parallel β-sheet located between H5 and H6, adopts the canonical three-layered α-helical fold. Superimposing the LXRβ LBDs complexed with T09 and GW on the one complexed with ECH leads to C^{α}-RMSD values of 1.1 Å and 1.0 Å, respectively, indicating a very close overall architecture (**Fig. 2**). Particularly, the C-terminal AF-2 helix presents a similar positioning, packed against the core of the protein, characteristic of the active conformation of the receptor. This is in agreement with the agonist activity of the three ligands. Despite these global structural similarities, each ligand shows a specific binding mode within the LBP.

The first crystal structure solved at a resolution of 2.8 Å by Williams *et al.* (PDB code 1P8D) revealed that the natural ECH ligand makes primarily hydrophobic contacts with its cholesterol skeleton, but also two distant polar interactions (**Fig. 3A**). Indeed, its hydroxyl group is oriented towards a polar sub-cavity near H1 to form a hydrogen bond with Glu281 (3.2 Å) and its epoxy moiety on the aliphatic chain is in the vicinity of the activation helix H12 and interacts through a polar contact with His435 (3.5 Å) on H11. This latter interaction packs His435 imidazole ring against the Trp457 indole ring on H12, maintaining an electrostatic contact that stabilizes the AF-2 helix in the active conformation. In this complex, ECH induces an elongated LBP of 675 $Å^3$ (**Fig. 4A**).

In all the structures in complex with T09 (PDB codes 1PQ9 and 1PQC), the ligand interacts in a similar way with both His435 and Trp457 residues. Indeed, the hexafluoro-carbinol moiety of T09 forms a strong hydrogen bond interaction (2.6 Å) with His435 (**Fig. 3B**). This histidine residue consequently promotes the electrostatic interaction with Trp457 to maintain H12 in the agonist conformation. In contrast to ECH, T09 is a compact and short ligand that could not reach the upper part of the cavity near the polar residues of H1 and consequently, does not interact with Glu281. However, in the lower part of the cavity, the residues encapsulate T09 and delimit a more compact LBP of 466 $Å^3$ (**Fig. 4B**). Compared to the 1P8D structure, the greatest conformational change is observed for Phe329 side chain which closes the pocket in direction of H1 and thus, creates a new

hydrophobic sub-cavity with Phe271 and Phe340 residues: these three phenylalanines rearrange themselves around the phenylsulfonyl group of T09 by making π-stacking interactions.

Farnegardh *et al.* also reported the crystallographic structure of GW (PDB code 1PQ6) in complex with human LXRβ with a resolution of 2.4 Å (**Fig. 3C**). Instead of the polar interaction observed with ECH and T09, GW contacts His435 (3.6 Å) through its hydrophobic 2-chloro-3-trifluoromethylbenzyl group. Nevertheless, this interaction is sufficient to maintain His435 and Trp457 in the same conformation as observed in the two other structures, suggesting that the resulting electrostatic interaction plays a crucial role in the stabilization of H12 and is required for the activation of the receptor. At the opposite side of the pocket near H1, the carboxylate moiety carried by GW is located in the polar sub-cavity where it forms a hydrogen bonds network with Arg319 (2.8 Å), Ser242 (2.9 Å) and a water molecule (2.5 Å). GW is bulkier than the two other ligands previously described and therefore induces structural rearrangements of the closest residues leading to a larger LBP of 831 $Å^3$ (**Fig. 4C**): particularly, the bisphenyl group of GW fits into the hydrophobic sub-cavity occupied by phenylsulfonyl moiety of T09 but due to its bulkiness, the bisphenyl group requires the displacement of the phenylalanines side chains namely Phe329, Phe271 and Phe340.

According to this structural analysis performed with the X-ray structures of the different LXRβ complexes, we assumed that the LBP is able to exhibit at least two extreme shapes upon ligand binding. The first LBP deriving from the small T09 scaffold could be considered as a “closed” state, where the Phe329 side chain adopts a conformation (χ_1 value of -163°) that restrains the cavity to its hydrophobic core near H12 and prevents such a compact ligand to reach the polar sub-cavity in the H1 area. The second LBP induced by the large GW pattern could be considered as an “open” state, where the Phe329 rotamer (χ_1 value of -67°) allows to form a larger pocket spreading from H12 residues to H1 and consequently, making an extension of the hydrophobic core pocket towards the polar sub-cavity.

Taken together, this structural analysis on LXRβ LBDs highlights:

(i) The role of key residues which enable the accommodation of ligands of different sizes and shapes, characterizing an induced fit phenomenon;

(ii) The importance of the His435/Trp457 switch as a key interaction required for maintaining H12 in an “agonist position” and consequently, for keeping activator profile of ligands.

Molecular Dynamics simulations

Reference systems

In order to check the validity of our MD simulation protocol to be used for studying the identified induced fit phenomenon, we have first investigated the behavior of the 1PQ6B/GW and 1PQ9B/T09 reference systems.

Our results revealed that the evolution of the two reference systems is very stable along the corresponding MD trajectories: the average values of C^{α}-RMSD obtained with 1PQ6B/GW and 1PQ9B/T09 are of 1.24±0.12 Å and 1.25±0.12 Å, respectively. All the secondary structure elements of the LXRβ LBD are maintained during the 4 ns MD (**Fig. 5**) including the critical helix 12.

If we focus on the 28 residues forming the LBP itself (**Figs. 6A** and **6B**), it appeared that the C^{α} trace is also very well maintained (average RMSD value of 0.84±0.12 Å and of 0.99±0.08 Å for 1PQ6B/GW and 1PQ9B/T09, respectively) and that the side chains of these residues contributed to weak distortions with an average RMSD value of 1.24±0.18 Å (1PQ6B/GW) and of 1.33±0.10 Å (1PQ9B/T09).

Nevertheless, among these 28 residues, a deviation above 2 Å is observed for Arg319 (average RMSD of 2.19±0.26 Å in 1PQ9B/T09) and Glu281 (average RMSD of 2.33±0.25 Å in PQ6B/GW). Interestingly, these two residues are located in the polar sub-cavity near H1 and are surrounded by water molecules. In the case of 1PQ9B/T09 MD simulation, Arg319 does not interact with the T09 ligand and can consequently move away from its crystallographic conformation to form a stabilizing salt bridge with Glu281. By contrast, Arg319 residue remained in interaction with the carboxylate moiety of GW during the MD of the second reference system, 1PQ6B/GW, preventing the formation of the same stabilizing salt bridge with Glu281. As a consequence this latter residue is pushed away from its starting position and its acidic group is shielded with water molecules.

Except for Glu281 and Arg319, the overall low distortion of LBP residues is in line with the conformational stability also measured by the average RMSD on ligands T09 (0.60±0.20 Å) and GW (0.84±0.20 Å) suggesting that all the main receptor-ligand interactions mentioned in the structural analysis of the two crystallographic complexes are maintained during MD simulations of the reference systems. We should therefore conclude that the reference systems are stable in our simulation conditions and that our results are consistent with the experimental structural data.

In order to refine our MD protocol with the reference systems, we also checked the protonation state of the sole histidine (His435) present in the LBP according to the role of this residue with regard to the interaction with the two ligands. Our calculations performed with both $N^{\delta1}$–H and $N^{\varepsilon2}$–H tautomers of His435 indicated that the protonation state seemed to have no noticeable impact on the interaction energy with GW ($E_{N\delta1-H}$ - $E_{N\varepsilon2-H}$ = -0.72 kcal/mol; **Fig. 7**). This is in agreement with the hydrophobic contact of GW with His435 observed in the structural analysis. However, the protonation state of His435 plays a crucial role in the interaction with T09 ($E_{N\delta1-H}$ - $E_{N\varepsilon2-H}$ = -4.65 kcal/mol; **Fig. 7**). This emphasizes that protonation on $N^{\delta1}$ position allows $N^{\varepsilon2}$ nitrogen to accept a hydrogen from the T09 carbinol group. Consequently, the $N^{\delta1}$–H protonation state for His435 residue has been considered in our MD protocol conditions and will be used for studying the two crossed systems defined in the next section.

Crossed systems

We decided to investigate the identified induced fit mechanism by means of the two following crossed systems: 1PQ6B/T09 complex represents the small T09 ligand placed inside the “open” state of the binding cavity and 1PQ9B/GW complex where the large GW ligand is forced inside the “closed” conformation. Both 1PQ6B/T09 and 1PQ9B/GW crossed systems were submitted to the same simulation protocol as the reference systems mentioned before. Analyses of the recorded MD trajectories were done considering particular subsets taken from the 28 residues involved in T09 and GW recognition. Our results highlighted (i) the remodeling of the hydrophobic sub-cavity involving rearrangements of the three phenylalanines Phe329, Phe271 and Phe340; (ii) the role of the polar residues Arg319 and Glu281 and water molecules within the polar sub-cavity near H1; (iii) the behavior of the critical His435/Trp457 switch.

Hydrophobic sub-cavity: aromatic triad formed by the three phenylalanines: Phe271,Phe329 and Phe340

MD trajectory of 1PQ6B/T09 crossed system showed a rapid conformational transition for Phe329 in less than 10 ps during the early stage of the MD trajectory (**Fig. 8A**) leading to the “closed” state observed in the reference 1PQ9B/T09 system. The movement of Phe329 is mostly due to a rotation around its χ_1 angle of about 80° to mimic the crystallographic rotamer found in the 1PQ9B X-ray structure (χ_1 of -163°). The resulting position of Phe329 is favored by the ability to form π-π interactions with the phenylsulfonyl of T09. In addition, we observed that, after a stable regime during

half of the simulation (from 1.5 to 3.5 ns), Phe329 χ_1 value varies punctually. Visual inspection of the trajectory revealed that this χ_1 variation corresponds to a displacement between both phenylsulfonyl and Phe329. Indeed, the highest values of $\delta\chi_1$ observed in **Fig. 8A** after 3.5 ns are related to different geometries of π–π interactions [32, 33].

Similarly, phenylsulfonyl moiety of T09 interacts with the two additional neighbor residues of the hydrophobic sub-cavity Phe271 and Phe340 (**Fig. 4B**) and stabilizes their position as shown with their respective low $\delta\chi_1$ values (**Fig. 8A**). Nevertheless, their final orientation required an adaptation of χ_2 angle as observed in the initial 300 ps period (**Fig. 8B**) in order to make them close to the original crystallographic rotamers (**Fig. 9A**). Subsequent weak fluctuations of χ_2 values are related to concerted π-π interactions between Phe271 and Phe340 and phenylsulfonyl group of T09 (**Fig. 8B**).

In the case of the second crossed system 1PQ9B/GW, the "closed to open" transition of Phe329 occured also in the equilibration phase in less than 10 ps (**Fig. 8C**). This major conformational change is characterized by marked $\delta\chi_1$ variation followed by a stable position of Phe329 side chain resulting from a π-π interaction with phenoxy group of GW as observed in the corresponding crystallographic complex. We should keep in mind that the initial high $\delta\chi_1$ value of 150° reflects the manual rotamer adjustment and the subsequent slight minimization of the complex (see Materials and Methods section) needed to place GW in the « closed » pocket. The minimization step (6,400 conjugate gradient iterations) prior to MD also led to the removal of the other pre-existing protein-ligand clashes: particularly, Phe340 shifted to accommodate bisphenyl moiety of GW, inducing a concerted rotation of Phe271. As a consequence, the change in $\delta\chi_2$ previously detected in the other crossed system for Phe271 and Phe340 (**Fig. 8B**) did not occur during the equilibration phase of 1PQ9B/GW. In addition, the low variability of $\delta\chi1$ and $\delta\chi_2$ indicates a local stabilization of the receptor mainly driven by π-π interactions between Phe271 and Phe340 and the bisphenyl moiety of GW (**Figs. 8C** and **8D**).

Other MD simulations related to induced fit phenomena were already reported, highlighting the use of this theoretical approach to handle such a complex mechanism[34-37]. In the case of LXR, the main feature derives from a major and fast conformational transition appearing early in the simulation time course (during the equilibration phase) and leading to a target conformation (close to the X-ray one) stable over the 4 ns production trajectories. This behavior can be explained by a rapid change in χ_1 value of Phe329 which occurs when interchanging the ligands. The aromatic cluster within the LBP seems therefore to play a crucial role for driving its adaptation to ligands, with different shapes and sizes, through π–π interactions.

Polar sub-cavity delineated by Glu281 and Arg319

Similarly to both reference systems, the polar sub-cavity of each crossed system undergone a remodeling during MD simulations:

(i) At the end of the equilibration phase of 1PQ6B/T09 crossed complex, some of the water molecules found in the sub-cavity are driven out and a salt bridge is formed between Glu281 and Arg319. This salt bridge is similar as the one described previously for the reference 1PQ9B/T09 system and is observed until the end of the simulation.

(ii) In the 1PQ9B/GW crossed system, the polar sub-cavity is partially occupied with the acidic moiety of GW. Surprisingly, instead of moving away from both the Arg319 and the ligand, as observed in the reference 1PQ6B/GW complex trajectory, Glu281 residue is maintained in the vicinity of the Arg319 guanidinium and GW carboxylate. Interestingly, to avoid the unfavorable Glu281-carboxylate/GW-carboxylate interaction, a water molecule is strongly anchored between these groups along the MD simulation. The computed average distance between Glu281 and GW C-carboxylate atoms is of 5.3 Å, which is in agreement with the average value found for such kind of bridged interactions within proteins [38]. While frequently exchanged during the simulation, another water molecule bound between the GW carboxylate and Arg319 guanidinium group is found at a position corresponding to the bridging water molecule observed in the crystallographic complex (**Fig. 3C**).

The His435/Trp457 switch

As observed in the crystallographic complexes, the interaction between the ligand and His435 sidechain is maintained during MD simulation of both 1PQ6B/T09 and 1PQ9B/GW crossed systems (**Fig. 9**):

(i) In the case of 1PQ6B/T09 crossed system, a stable hydrogen bond is observed during the MD between His435 and carbinol moiety of T09 with an average distance of 2.7 Å compared to 2.6 Å measured in the X-ray 1PQ9B structure (**Fig. 3B**).

(ii) In the 1PQ9B/GW crossed system, the hydrophobic interaction observed between His435 and GW in the MD (average distance of 3.3 Å) is similar to the one found in the 1PQ6B crystal structure (3.6 Å, **Fig. 3C**).

The stabilization of His435 by the ligands, characterized in the MD of the two crossed systems, enables to maintain the Trp457 position in a similar way as observed in the corresponding X-ray structures. As a result, the His435/Trp457 switch was conserved and provided a topological requirement involved in the positioning of AF-2 activation helix.

CONCLUSION

X-ray structure analysis of LXRβ complexed with chemically diverse ligands revealed the plasticity of the LBP residues which adapt themselves to the ligands shape and size. A MD protocol was tested with two reference crystallographic systems 1PQ9B/T09 and 1PQ6B/GW showing markedly different shapes of the binding pocket. The resulting 4 ns trajectories confirmed the reliability of the MD protocol and its potential use for "crossed studies" where the two ligands were exchanged to each other.

As a matter of fact, our MD approach showed the ability to recover the original LXR conformation associated to each ligand as obtained by X-ray crystallography. This confirms that the present MD conditions were able to manage the induced fit phenomenon occurring in the LXR context, and to gain more insight into the molecular mechanism. Thus, MD data pointed out that encapsulation of the small T09 ligand in the GW pocket or the adaptation of the LBP to the large GW ligand is mainly driven by a rapid conformational change of Phe329 gate occurring at the very beginning of the simulation period. The subsequent closing or opening of the LXRβ LBP is further refined by small displacements of neighbor phenylalanines in order to optimize interactions with the respective ligand giving rise to concerted movements along the rest of the MD trajectory as exemplified with Phe271 and Phe340. Interestingly, MD calculations in explicit solvent stressed the contribution of the polar sub-cavity in the overall induced fit phenomenon: (i) a salt bridge could be formed between Arg319 and Glu281 and helped stabilize LXRβ LBP locally (ii) water molecules participated to polar interactions either with the salt bridge or with the acidic group present on GW ligand. The latter polar network is consistent with the crystallographic data showing a bound water molecule in this polar area.

Thus, the present work emphasizes the important role of solvation/desolvation process in the ligand-receptor recognition especially for compounds bearing a charged group like GW. This also may explain why the presence of an acidic moiety does not account for an increase in LXR potency of GW (EC_{50} values of 190 nM) compared to T09 (EC_{50} values of 85 nM) but rather for a decrease by one order of magnitude [39]. The current results also highlight the restricted conformational flexibility of the His435/Trp457 switch, which is in agreement with its critical role in the positioning of the AF-2 helix required for activation of LXR receptor.

Altogether, the present study opens new ways of investigation of LXR plasticity and related induced fit process with several applications for drug design. As indicated previously, LXR conformational adaptation mainly occurred early and rapidly which allows the use of a shorter MD protocol as a post-processing step in a ligand docking approach of a SBDD strategy without performing computer intensive calculations. Such a refinement step is usually needed for gaining insight into the topological environment surrounding any new lead compound[40] and thus, for improving accuracy of docking method carried out with the aim of designing analogs. Finally, as exemplified with this work and similarly to other studies [41, 42], a MD-based approach can be seen as a relevant method for generating novel LXR conformations used as molecular filters in a virtual screening strategy.

ACKNOWLEDGMENTS

The authors thank the Fournier/Solvay and LORIA collaborators who have reviewed this manuscript. AB is grateful to Association pour la Recherche sur le Cancer (ARC) for financial support (Grant N° JR/MLD/MDV - A06/4).

FIGURES

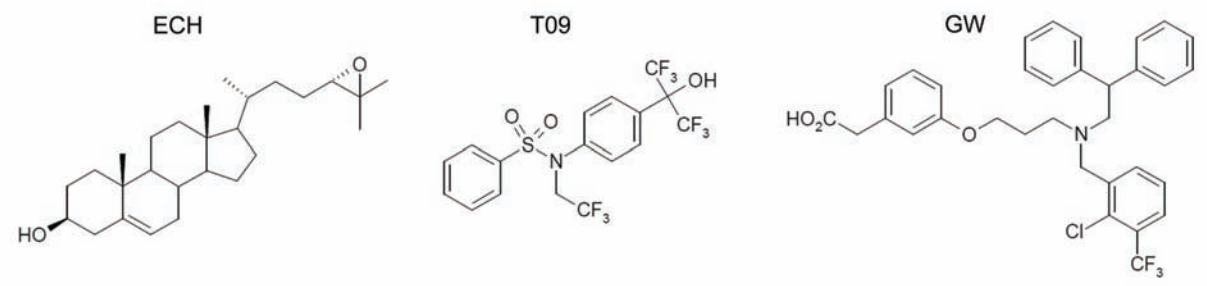


*Fig. 1 - Chemical structures of three LXR agonists: 24(S),25-epoxycholesterol (**ECH**), T0901317 (**T09**), and GW3965 (**GW**).*

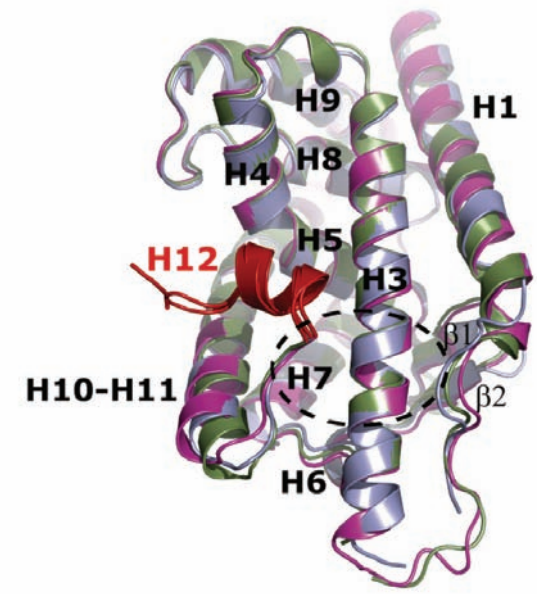


Fig. 2 - Superimposition of three LXRβ LBD crystal structures. Traces of proteins complexed with ECH, T09, and GW are colored in blue, green and magenta, respectively. The LBD is composed of twelve α-helices (labeled H1 to H12) displayed as a ribbon and two β-strands (namely β1 and β2) represented by arrows. The ligand binding pocket location is encircled by a dashed line. C-terminal helices 12 (Activation Function 2) are highlighted in red.

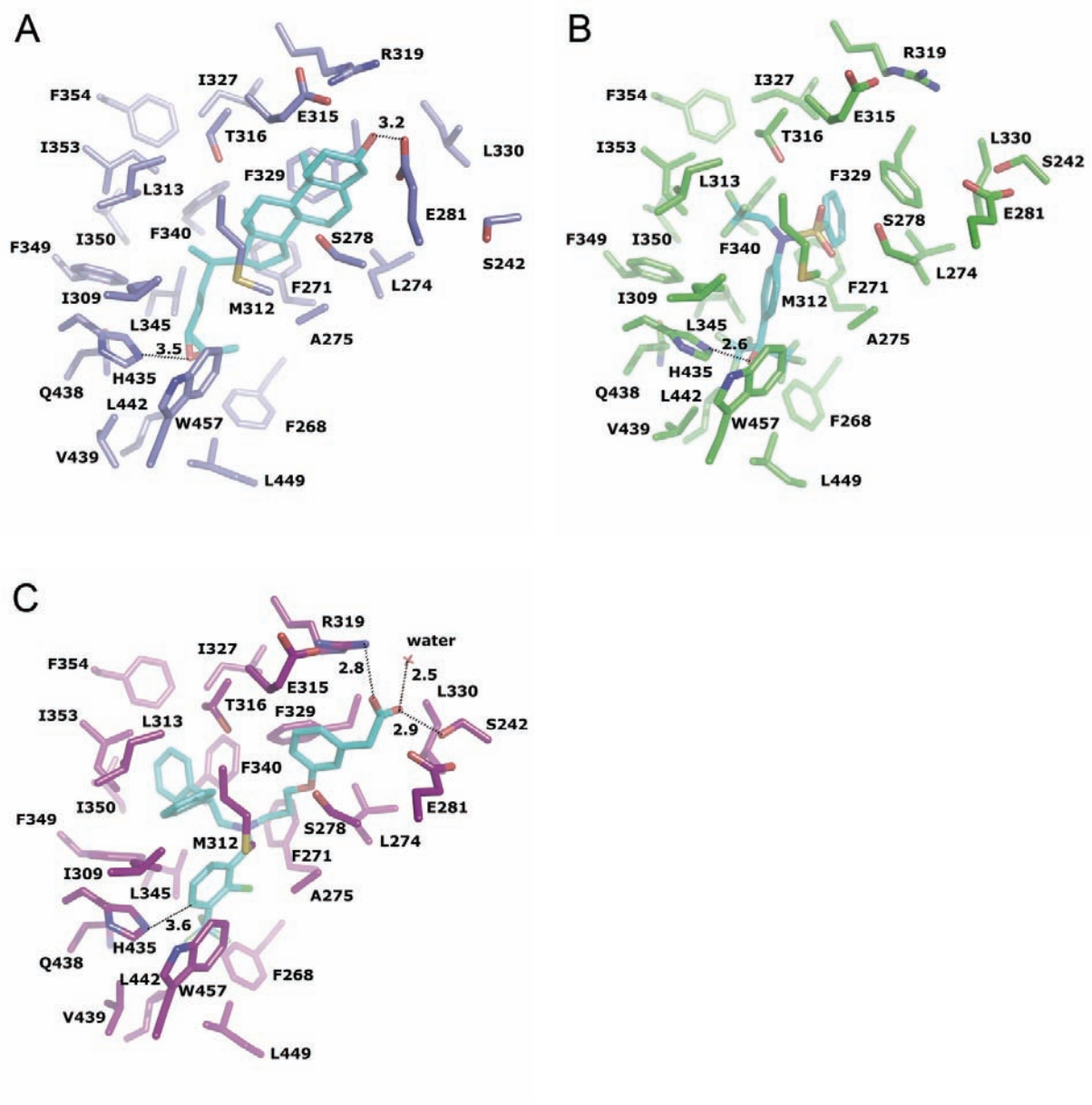


*Fig. 3 - Different protein-ligand patterns from crystal structures of LXRβ complexes. Interacting residues (the 28-residues subset considered in this work), surrounding ECH (**A**), T09 (**B**) and GW (**C**) ligands (colored in cyan) are labeled and colored in blue, green and magenta, respectively. Hydrogen bonds are depicted as dashed lines and their length is indicated in Angström. A water molecule is represented by a red cross (**C**).*

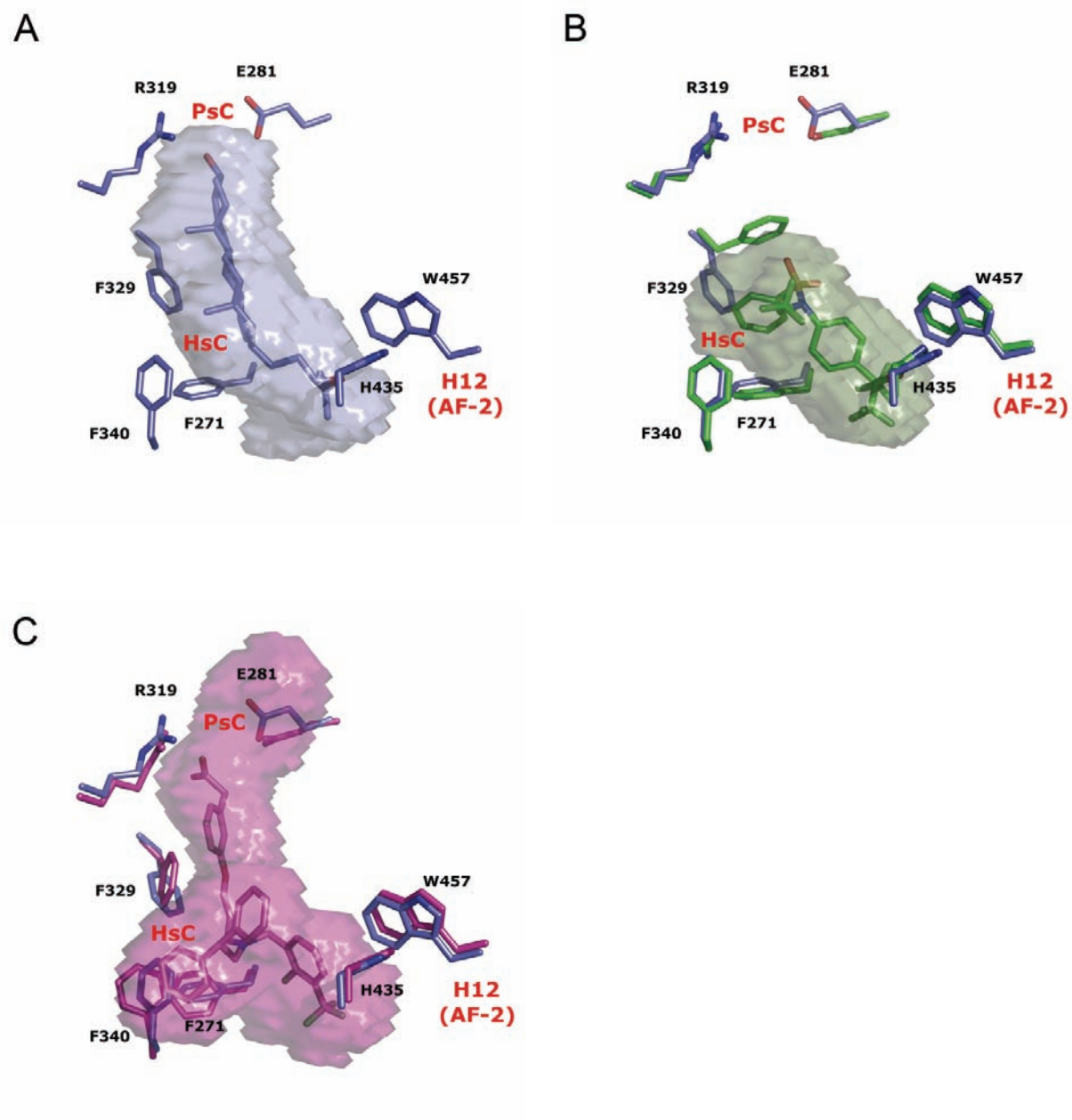


*Fig. 4 - Comparison of ligand binding pocket from X-ray structures of LXRβ complexed with ECH (blue, **A**), T09 (green, **B**) and GW (magenta, **C**). Pocket volumes were computed with VOIDOO and are represented by transparent surfaces colored according to their respective ligand. Panel A shows key residues in interaction with ECH as blue sticks. Panel B and C show the superimposition of these same residues within T09 (green) and GW (magenta) pockets, respectively. PsC: polar sub-cavity; HsC: hydrophobic sub-cavity; H12: helix 12; AF-2: activation function 2.*

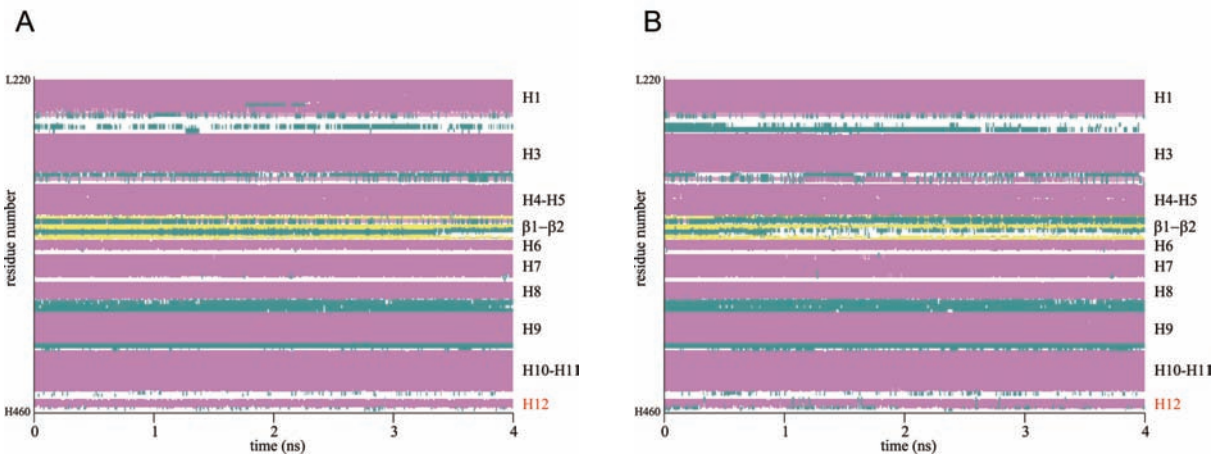


*Fig. 5 - Secondary structure elements evolution of the LBD along MD trajectory, for the two reference systems 1PQ6B/GW (**A**) and 1PQ9B/T09 (**B**). α-helices, 3_{10}-helices, β-sheets, β-turns, and coils are shown in purple, pink, yellow, cyan and white, respectively.*

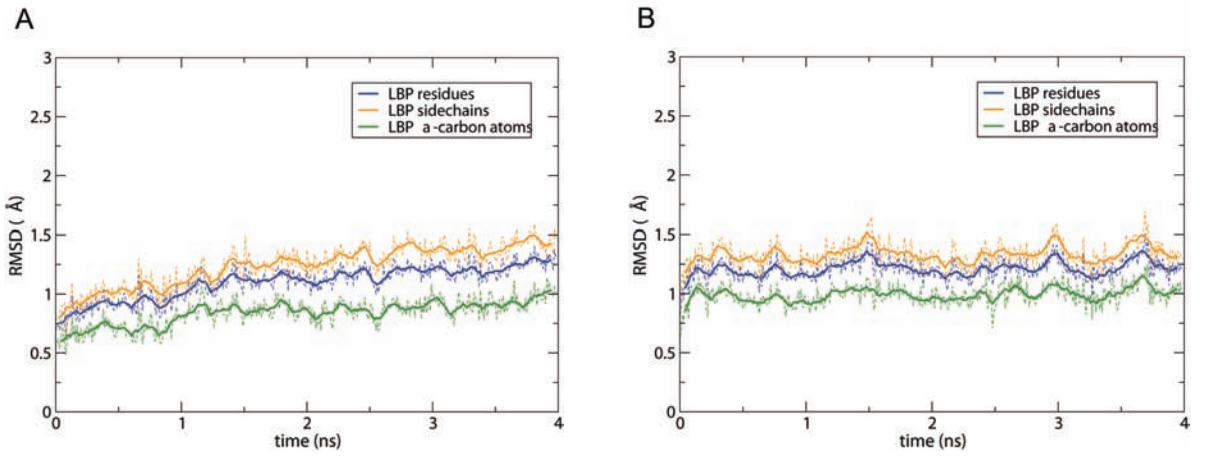


*Fig. 6 - RMS deviations for the 28 residues lining the LBP along MD trajectory for the two reference systems 1PQ6B/GW (**A**) and 1PQ9B/T09 (**B**).*

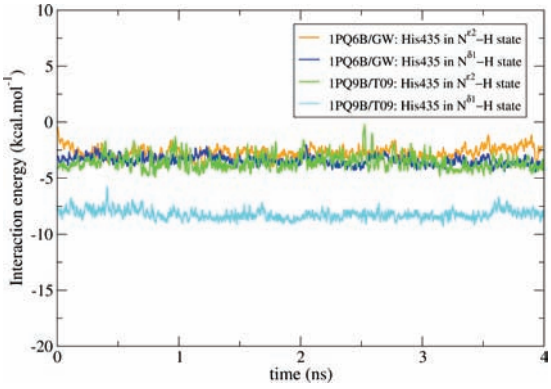


Fig. 7 - Influence of the protonation state of His435 on interaction energy with each ligand of the reference systems 1PQ6B/GW and 1PQ9B/T09 along the MD simulations. Both $N^{\delta 1}$–H and $N^{\varepsilon 2}$–H His435 tautomeric forms are reported.

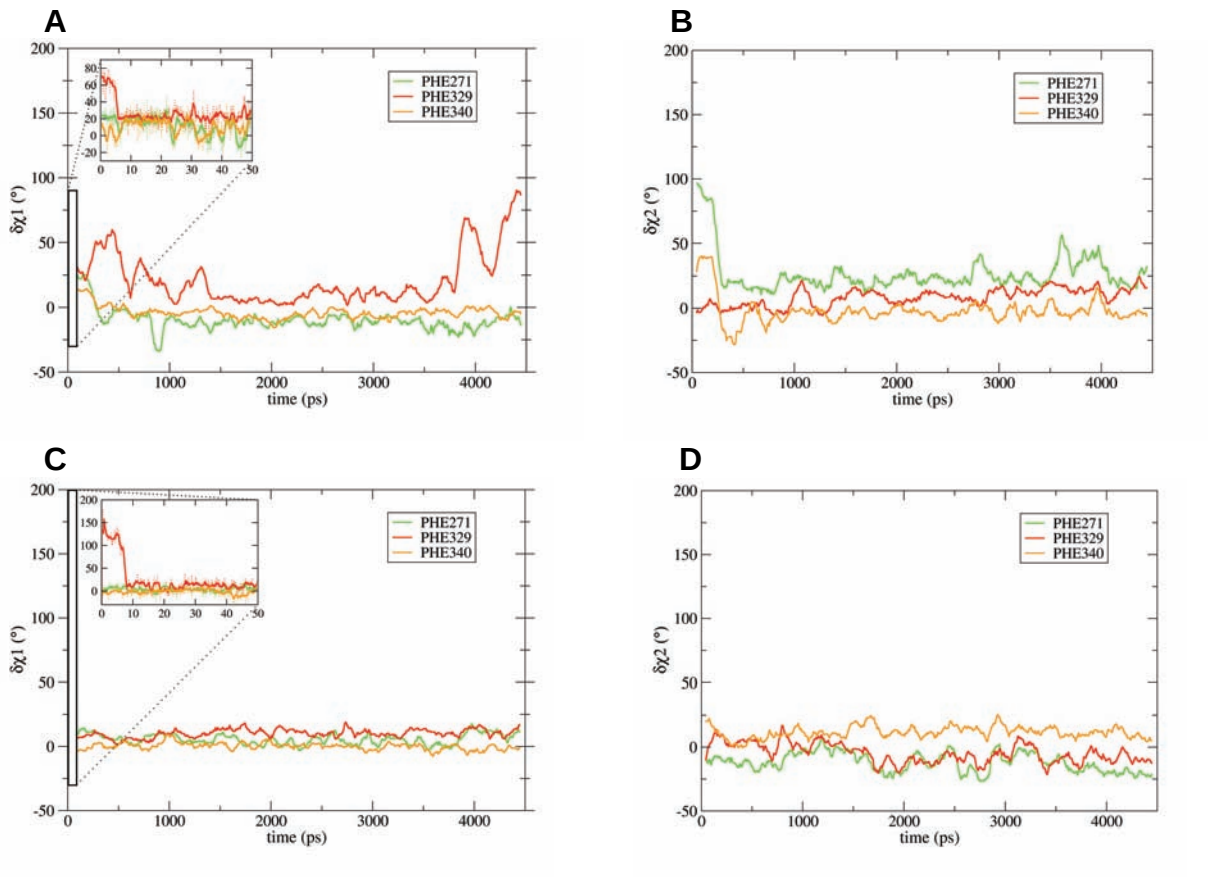


*Fig. 8 - Torsion angle deviation from crystal structures for the three phenylalanines (Phe271, Phe329 and Phe340) lining the LBP hydrophobic sub-cavity along the MD trajectory (including equilibration phase). Panels **A** and **B** show their $\delta\chi_1$ and $\delta\chi_2$ deviation in the 1PQ6B/T09 crossed system from the 1PQ9B X-ray structure. Similarly, panels **C** and **D** depict the same features for 1PQ9B/GW crossed system with respect to 1PQ6B crystal structure. On panels **A** and **C**, a frame (ranging from 0 to 50 ps) highlights the very beginning of the simulation.*

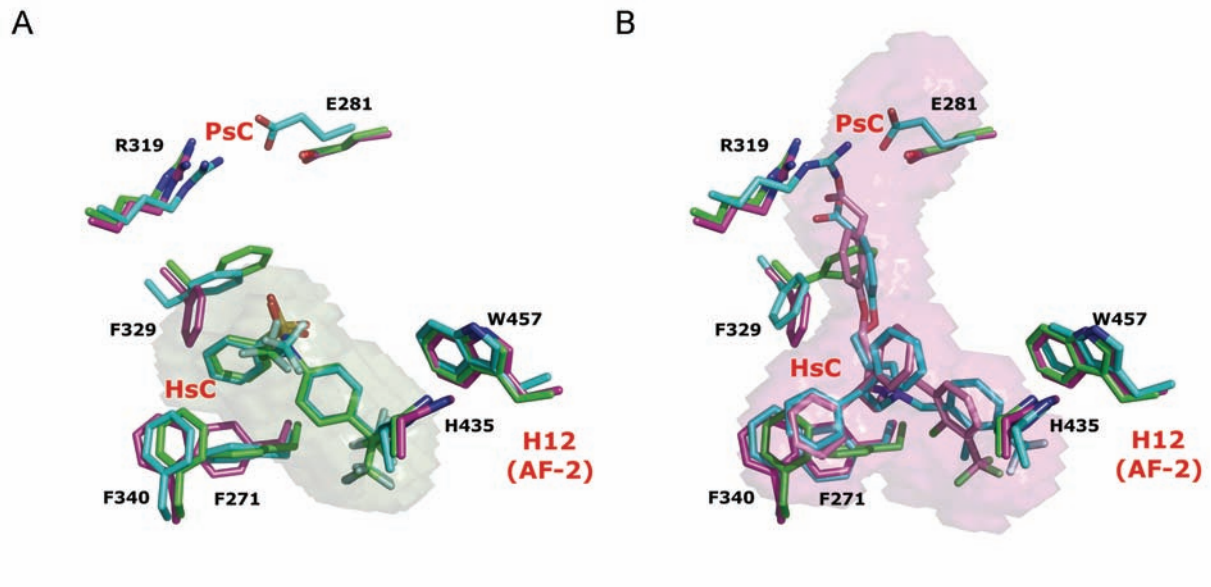


*Fig. 9 - MD trajectory snapshots of the two crossed systems 1PQ6B/T09 (**A**) and 1PQ9B/GW (**B**), compared to the crystal structures. LBP key residues are represented as sticks and colored in cyan, magenta and green for the MD snapshots, 1PQ6B and 1PQ9B structures, respectively. Following the same color code, GW and T09 ligands are displayed as sticks. PsC: polar sub-cavity; HsC: hydrophobic sub-cavity; H12: helix 12; AF-2: activation function 2.*

REFERENCES

(1) Mangelsdorf, D. J.; Thummel, C.; Beato, M.; Herrlich, P.; Schutz, G.; Umesono, K.; Blumberg, B.; Kastner, P.; Mark, M.; Chambon, P.; Evans, R. M. The nuclear receptor superfamily: the second decade. *Cell* **1995**, *83* (6), 835-839.

(2) Repa, J. J.; Mangelsdorf, D. J. The role of orphan nuclear receptors in the regulation of cholesterol homeostasis. *Annu. Rev. Cell Dev. Biol.* **2000**, *16*, 459-481.

(3) Zhang, Y.; Mangelsdorf, D. J. LuXuRies of lipid homeostasis: the unity of nuclear hormone receptors, transcription regulation, and cholesterol sensing. *Mol. Interv.* **2002**, *2* (2), 78-87.

(4) Repa, J. J.; Mangelsdorf, D. J. Nuclear receptor regulation of cholesterol and bile acid metabolism. *Curr. Opin. Biotechnol.* **1999**, *10* (6), 557-563.

(5) Janowski, B. A.; Willy, P. J.; Devi, T. R.; Falck, J. R.; Mangelsdorf, D. J. An oxysterol signalling pathway mediated by the nuclear receptor LXR alpha. *Nature* **1996**, *383* (6602), 728-731.

(6) Lehmann, J. M.; Kliewer, S. A.; Moore, L. B.; Smith-Oliver, T. A.; Oliver, B. B.; Su, J. L.; Sundseth, S. S.; Winegar, D. A.; Blanchard, D. E.; Spencer, T. A.; Willson, T. M. Activation of the nuclear receptor LXR by oxysterols defines a new hormone response pathway. *J. Biol. Chem.* **1997**, *272* (6), 3137-3140.

(7) Williams, S.; Bledsoe, R. K.; Collins, J. L.; Boggs, S.; Lambert, M. H.; Miller, A. B.; Moore, J.; McKee, D. D.; Moore, L.; Nichols, J.; Parks, D.; Watson, M.; Wisely, B.; Willson, T. M. X-ray crystal structure of the liver X receptor beta ligand binding domain: regulation by a histidine-tryptophan switch. *J. Biol. Chem.* **2003**, *278* (29), 27138-27143.

(8) Farnegardh, M.; Bonn, T.; Sun, S.; Ljunggren, J.; Ahola, H.; Wilhelmsson, A.; Gustafsson, J. A.; Carlquist, M. The three-dimensional structure of the liver X receptor beta reveals a flexible ligand-binding pocket that can accommodate fundamentally different ligands. *J. Biol. Chem.* **2003**, *278* (40), 38821-38828.

(9) Hoerer, S.; Schmid, A.; Heckel, A.; Budzinski, R. M.; Nar, H. Crystal structure of the human liver X receptor beta ligand-binding domain in complex with a synthetic agonist. *J. Mol. Biol.* **2003**, *334* (5), 853-861.

(10) Svensson, S.; Ostberg, T.; Jacobsson, M.; Norstrom, C.; Stefansson, K.; Hallen, D.; Johansson, I. C.; Zachrisson, K.; Ogg, D.; Jendeberg, L. Crystal structure of the heterodimeric complex of LXRalpha and RXRbeta ligand-binding domains in a fully agonistic conformation. *EMBO J.* **2003**, *22* (18), 4625-4633.

(11) Wurtz, J. M.; Bourguet, W.; Renaud, J. P.; Vivat, V.; Chambon, P.; Moras, D.; Gronemeyer, H. A canonical structure for the ligand-binding domain of nuclear receptors. *Nat. Struct. Mol. Biol.* **1996**, *3* (1), 87-94.

(12) Bourguet, W.; Ruff, M.; Chambon, P.; Gronemeyer, H.; Moras, D. Crystal structure of the ligand-binding domain of the human nuclear receptor RXR-alpha. *Nature* **1995**, *375* (6530), 377-382.

(13) Egea, P.; Klaholz, B.; Moras, D. Ligand-protein interactions in nuclear receptors of hormones. *FEBS Lett.* **2000**, *476* (1), 62-67.

(14) Castrillo, A.; Tontonoz, P. Nuclear receptors in macrophage biology: at the crossroads of lipid metabolism and inflammation. *Annu. Rev. Cell Dev. Biol.* **2004**, *20* (1), 455-480.

(15) Zelcer, N.; Tontonoz, P. Liver X receptors as integrators of metabolic and inflammatory signaling. *J. Clin. Invest.* **2006**, *116* (3), 607-614.

(16) Mitro, N.; Mak, P. A.; Vargas, L.; Godio, C.; Hampton, E.; Molteni, V.; Kreusch, A.; Saez, E. The nuclear receptor LXR is a glucose sensor. *Nature* **2007**, *445* (7124), 219-223.

(17) Terasaka, N.; Hiroshima, A.; Koieyama, T.; Ubukata, N.; Morikawa, Y.; Nakai, D.; Inaba, T. T-0901317, a synthetic liver X receptor ligand, inhibits development of atherosclerosis in LDL receptor-deficient mice. *FEBS Lett.* **2003**, *536* (1-3), 6-11.

(18) Joseph, S. B.; McKilligin, E.; Pei, L.; Watson, M. A.; Collins, A. R.; Laffitte, B. A.; Chen, M.; Noh, G.; Goodman, J.; Hagger, G. N.; Tran, J.; Tippin, T. K.; Wang, X.; Lusis, A. J.; Hsueh, W. A.; Law, R. E.; Collins, J. L.; Willson, T. M.; Tontonoz, P. Synthetic LXR ligand inhibits the development of atherosclerosis in mice. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99* (11), 7604-7609.

(19) Schultz, J. R.; Tu, H.; Luk, A.; Repa, J. J.; Medina, J. C.; Li, L.; Schwendner, S.; Wang, S.; Thoolen, M.; Mangelsdorf, D. J.; Lustig, K. D.; Shan, B. Role of LXRs in control of lipogenesis. *Genes Dev.* **2000**, *14* (22), 2831-2838.

(20) Koshland Jr., D. The key–lock theory and the induced fit theory. *Angew. Chem., Int. Ed. Engl.* **1994**, *33* (23-24), 2375-2378.

(21) Gunther, J.; Bergner, A.; Hendlich, M.; Klebe, G. Utilising structural knowledge in drug design strategies: applications using Relibase. *J. Mol. Biol.* **2003**, *326* (2), 621-636.

(22) Hendlich, M.; Bergner, A.; Gunther, J.; Klebe, G. Relibase: design and development of a database for comprehensive analysis of protein-ligand interactions. *J. Mol. Biol.* **2003**, *326* (2), 607-620.

(23) Kleywegt, G. J.; Jones, T. A. Detection, delineation, measurement and display of cavities in macromolecular structures. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **1994**, *50* (Pt 2), 178-185.

(24) Fiser, A.; Sali, A. Modeller: generation and refinement of homology-based protein structure models. *Methods enzymol.* **2003**, *374*, 461-491.

(25) Phillips, J.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R.; Kale, L.; Schulten, K. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **2005**, *26* (16), 1781-1802.

(26) MacKerell Jr., A. D.; Bashford, D.; Bellott, M.; Dunbrack Jr., R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher III, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiórkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem.* **1998**, *102* (18), 3586-3616.

(27) Momany, F.; Rone, R. Validation of the general purpose QUANTA 3.2/CHARMm force field. *J. Comput. Chem.* **1992**, *13* (7), 888-900.

(28) Humphreys, D.; Freisner, R.; Berne, B. A multiple-time-step molecular dynamics algorithm for macromolecules. *J. Phys. Chem.* **1994**, *98*, 6885-6892.

(29) Kräutler, V.; van Gunsteren, W. F.; Hünerberger, P. H. A fast SHAKE algorithm to solve distance constraint equations for small molecules in molecular dynamics simulations. *J. Comput. Chem.* **2001**, *22* (5), 501-508.

(30) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An N.log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98* (12), 10089-10092.

(31) Humphrey, W.; Dalke, A.; Schulten, K. VMD - Visual Molecular Dynamics. *J. Mol. Graph.* **1996**, *14*, 33-38.

(32) Hunter, C.; Lawson, K.; Perkins, J.; Urch, C. Aromatic interactions. *J. Chem. Soc., Perkin Trans. 2* **2001**, 651-669.

(33) McGaughey, G.; Gagne, M.; Rappe, A. π-Stacking interactions. Alive and well in proteins. *J. Biol. Chem.* **1998**, *273* (25), 15458-15463.

(34) Senapati, S.; Bui, J.; McCammon, J. Induced fit in mouse acetylcholinesterase upon binding a femtomolar inhibitor: A molecular dynamics study. *J. Med. Chem.* **2005**, *48* (26), 8155-8162.

(35) Sotriffer, C.; Krämer, O.; Klebe, G. Probing flexibility and "induced-fit" phenomena in aldose reductase by comparative crystal structure analysis and molecular dynamics simulations. *Proteins: Struct., Funct., Bioinf.* **2004**, *56* (1), 52-66.

(36) Verli, H.; Guimaraes, J. Insights into the induced fit mechanism in antithrombin-heparin interaction using molecular dynamics simulations. *J. Mol. Graph. Model.* **2005**, *24*, 203-212.

(37) Macek, P.; Novak, P.; Krizova, H.; Zidek, L.; Sklenar, V. Molecular dynamics study of major urinary protein-pheromone interactions: A structural model for ligand-induced flexibility increase. *FEBS Lett.* **2006**, *580* (2), 682-684.

(38) Singh, J.; Thornton, J. *Atlas of protein side-chain interactions*. Vol. I and II **1992**. IRL press, Oxford, UK.

(39) Collins, J.; Fivush, A.; Watson, M.; Galardi, C.; Lewis, M.; Moore, L.; Parks, D.; Wilson, J.; Tippin, T.; Binz, J.; Plunket, K.; Morgan, D.; Beaudet, E.; Whitney, K.; Kliewer, S.; Willson, T. Identification of a nonsteroidal liver X receptor agonist through parallel array synthesis of tertiary amines. *J. Med. Chem.* **2002**, *45* (10), 1963-1966.

(40) Alonso, H.; Bliznyuk, A.; Gready, J. Combining docking and molecular dynamic simulations in drug design. *Med. Res. Rev.* **2006**, *26* (5), 531-568.

(41) Sivanesan, D.; Rajnarayanan, R.; Doherty, J.; Pattabiraman, N. In-silico screening using flexible ligand binding pockets: a molecular dynamics-based approach. *J. Comput.-Aided Mol. Des.* **2005**, *19* (4), 213-228.

(42) Wong, C.; Kua, J.; Zhang, Y.; Straatsma, T.; McCammon, J. Molecular docking of balanol to dynamics snapshots of protein kinase A. *Proteins: Struct., Funct., Bioinf.* **2005**, *61* (4), 850-858.

I-6. Emploi de VSM-G dans une campagne de criblage à grande échelle

Dans le cadre d'un contrat de collaboration avec les Laboratoires Fournier (membre du groupe Solvay), nous avons employé la plateforme VSM-G pour cribler une large chimiothèque sur le récepteur nucléaire LXRβ qui est une de leurs cibles d'étude. L'objectif de cette campagne de criblage était double : identifier de nouveaux composés bioactifs ciblant LXRβ, mais également profiter de cette occasion pour tester la plateforme VSM-G en grandeur nature. Les résultats préliminaires de cette étude ne font et ne feront pas l'objet de publication pour des raisons de confidentialité. Néanmoins, nous présenterons la stratégie adoptée pour cette campagne, ainsi qu'un aperçu des résultats obtenus.

I-6.1. Échantillonnage de la cible

L'examen de la flexibilité du site actif de LXRβ a montré combien celui-lui est sujet au phénomène d'*induced fit*. Par conséquent, il nous a semblé nécessaire de disposer d'un ensemble de conformations du récepteur afin de prendre en compte la flexibilité de la cible durant le criblage. Pour ce faire, nous avons utilisé les 3 conformations de LXRβ disponibles publiquement, co-cristallisées avec différents agonistes (codes PDB : 1P8D, 1PQ6 et 1PQ9 ; ces dernières ayant déjà été employées pour l'étude de l'article #2) auxquelles nous avons ajouté d'autres conformations issues d'une simulation par DM de la cible sous sa forme libre.

La simulation par DM a été conduite dans les conditions de l'ensemble thermodynamique NPT, c'est-à-dire à pression et température constantes et en prenant en compte l'effet du solvant de façon explicite. Afin d'obtenir un échantillonnage conformationnel du site actif qui soit suffisamment représentatif et divers, un peu plus de 30 ns de simulation ont été nécessaires[*]. Pour des raisons pratiques, nous avons décomposé la trajectoire de cette simulation en une collection de 600 conformations, extraites à intervalles de temps régulier (1 conformation / ~50 ps)[†].
Afin d'y soutirer un certain nombre de conformères divers, nous avons choisi de "clustériser" cette trajectoire de 600 conformations sur la base des RMSD croisés des résidus identifiés dans l'article #3 comme jouant un rôle clé dans la flexibilité du site actif. A partir des calculs, des cartes de RMSD croisés ont été construites, lesquelles ont ensuite servi pour la "clustérisation".

[*] Cela représente près de 3 semaines de calcul sur une grappe de PC de 22 processeurs.

[†] Pour s'assurer qu'à une telle fréquence d'extraction n'entraînait pas de perte d'information, nous avons comparé ces cartes de RMSD avec celles issues en extrayant avec une fréquence de 1 conformation / 20 ps (~1600 conformations récoltées) : qualitativement, les résultats sont équivalents.

Mais, dans un premier temps, plusieurs sous-ensembles de résidus du site actif ont été considérés afin de voir leur influence sur les cartes de RMSD (figure ci-dessous).

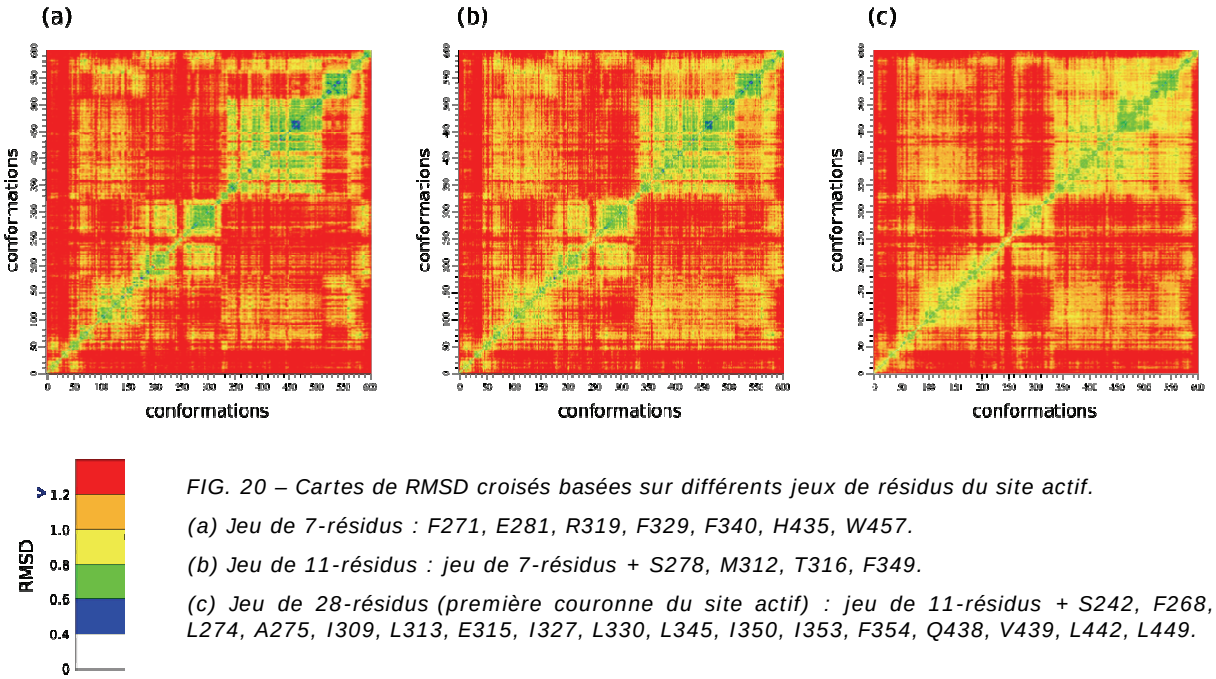


FIG. 20 – Cartes de RMSD croisés basées sur différents jeux de résidus du site actif.

(a) Jeu de 7-résidus : F271, E281, R319, F329, F340, H435, W457.

(b) Jeu de 11-résidus : jeu de 7-résidus + S278, M312, T316, F349.

(c) Jeu de 28-résidus (première couronne du site actif) : jeu de 11-résidus + S242, F268, L274, A275, I309, L313, E315, I327, L330, L345, I350, I353, F354, Q438, V439, L442, L449.

La figure ci-après présente une carte de RMSD, basée sur le jeu de 28-résidus du site actif, utilisée pour la clustérisation. Celle-ci nous a permis de sélectionner 7 conformations distinctes et représentatives de la trajectoire, fournissant un jeu minimal caractérisant l'ensemble des états conformationnels observés au cours de la simulation par DM.

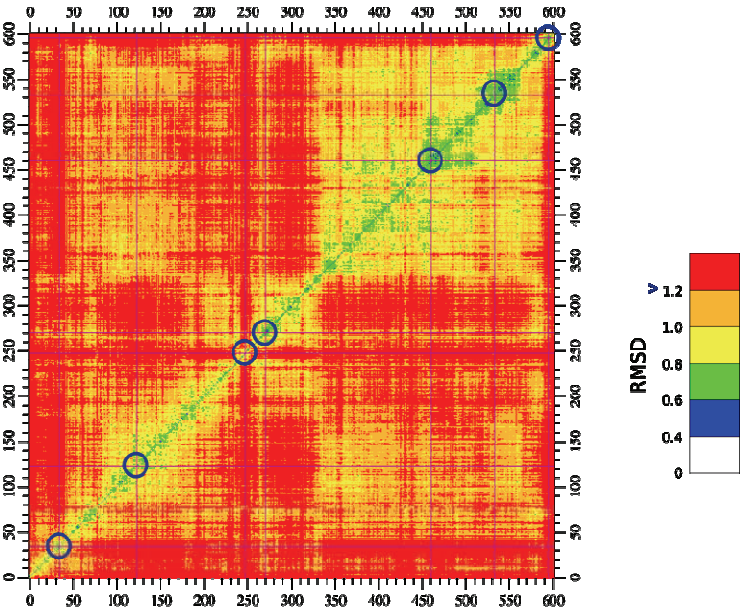


FIG. 21 - Carte de RMSD basées sur le jeu de 28-résidus du site actif. Les 7 conformations choisies pour ce criblage haut-débit sont encerclées.

Une analyse comparative (toujours sur la base des RMSD) des 7 conformations obtenues par DM a confirmé leur non redondance conformationnelle par rapport aux 3 conformations expérimentales. Pour ce criblage, nous avons donc utilisé au final un ensemble de 10 conformations représentatives de la variété structurale du domaine de liaison de LXRβ.

(a) (b)

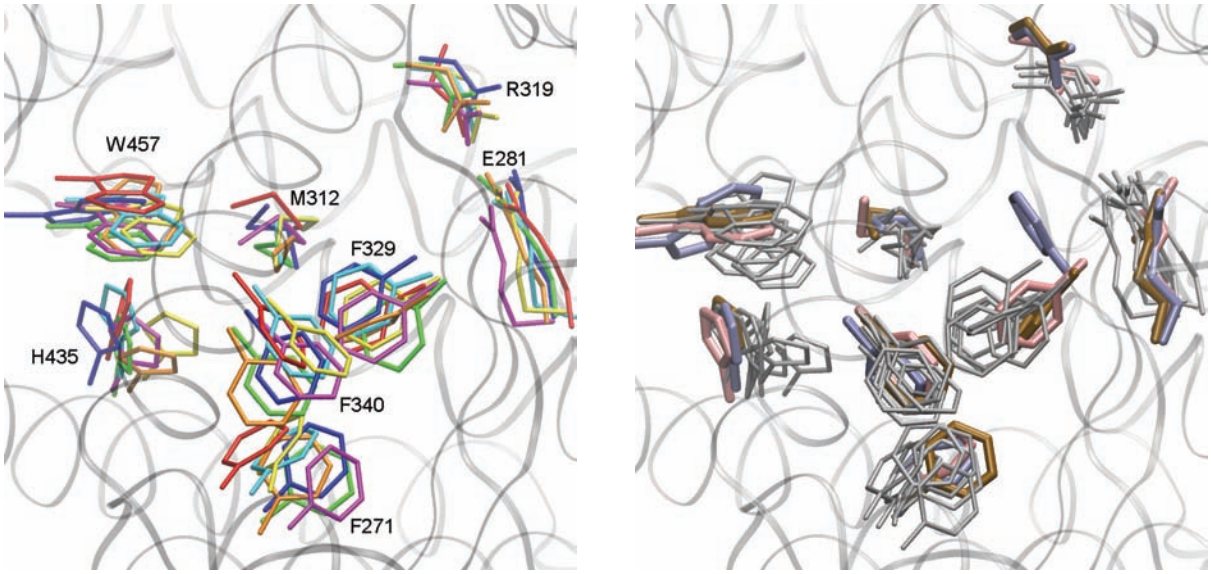


FIG. 22 - Superposition des conformations sélectionnées pour le criblage. La trace de la protéine est représentée en ruban et 8 résidus du site actif (jeu de 7-résidus + M312) sont en bâtons.
(a) Les 7 conformations issues de la DM uniquement.
*(b) Les 10 conformations : les 7 issues de la DM en gris et les 3 cristallographiques (**1P8D**, **1PQ6** et **1PQ9**).*

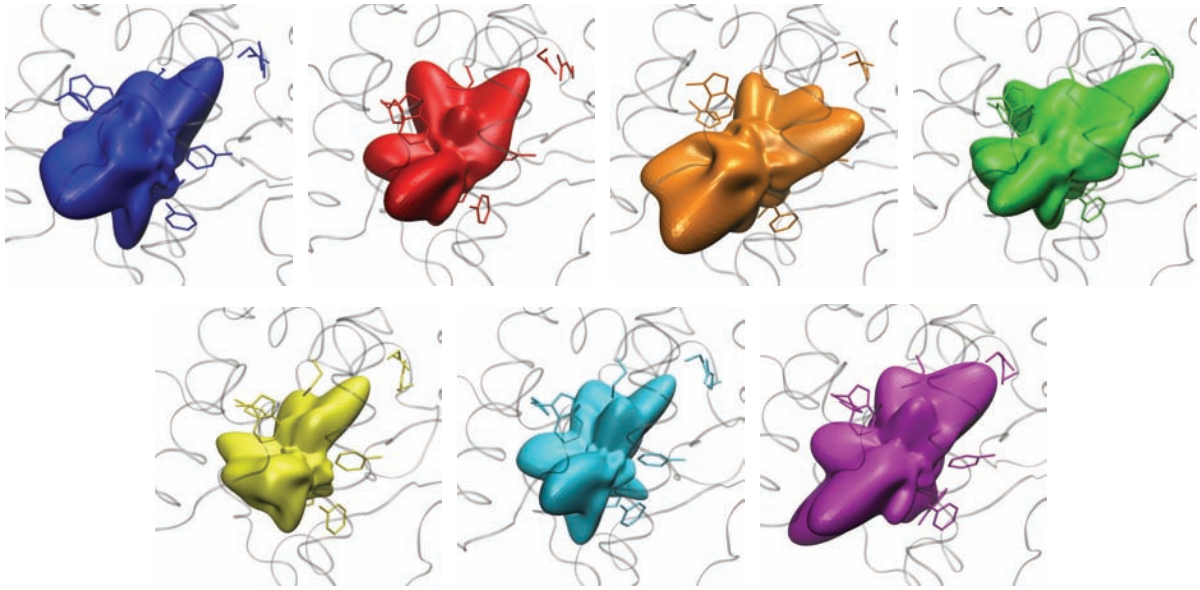


FIG. 23 - Surfaces des 7 sites actifs issus de la DM considérées par SHEF dans le criblage (même règles de couleurs que dans FIG. 22 (a)).

I-6.2. Échantillonnage des ligands de la chimiothèque

La chimiothèque à cribler est une compilation d'environ 600 000 composés, disponibles commercialement auprès de divers fournisseurs (Enamine, ChemDiv et AMRI, anciennement Comgenex). Dans un premier temps, cette chimiothèque a été pré-filtrée suivant les règles de Lipinski (avec une violation permise) et en excluant toute molécule potentiellement toxique, ou supposée telle, et celles comportant des groupements jugés trop réactifs.

Afin de prendre en compte la flexibilité des ligands de la chimiothèque au niveau du premier filtre de VSM-G (docking rigide avec SHEF), une génération conformationnelle a été réalisée avec le programme OMEGA [177]. En fixant à 400 le nombre maximum de conformères à générer par ligand, cet échantillonnage a produit ~90 000 000 conformères 3D. La masse de données correspondante (environ 600 giga-octets, comprenant pour chaque conformère les coordonnées atomiques, la topologie et la définition de surface) est stockée sur une machine dédiée. Nous pourrons ainsi réutiliser cette information pour des expériences ultérieures de criblage virtuel sur cette base.

I-6.3. Le criblage

La large chimiothèque a été criblée sur chacune des 10 conformations de LXRβ. Comme le criblage virtuel se prête bien au calcul distribué, nous avons profité de l'architecture du réseau national Grid'5000[*] [233] (1360 CPUs utilisés) et de nos *clusters* (16 CPUs) pour répartir les calculs SHEF. Globalement, cette campagne de criblage géométrique a nécessité environ 1 mois de calculs. De façon intéressante, la génération des coefficients d'harmoniques sphériques (programme MSSH présenté dans l'article #2) a occupé la quasi-totalité des temps de calcul (~ 90%), alors que le docking par SHEF s'est effectué sur 10 conformations de LXRβ. Un criblage virtuel ultérieur sur cette base de 90 millions de conformères ne nécessitera plus cette première étape. On peut ainsi estimer que, pour une conformation de récepteur donnée (plus nécessairement LXRβ), le docking de la base de conformères correspondant à ~ 600 000 molécules ne prendra désormais plus qu'environ une demi-heure (sur Grid'5000) ou bien une journée (sur un de nos *clusters*).

[*] Grid'5000 (www.grid5000.fr) nous a donné accès à son réseau, à titre expérimental, afin de nous permettre de définir les conditions optimales pour distribuer les calculs SHEF sur une architecture de grille de calculs.

Les résultats préliminaires de cette campagne de criblage sont, pour l'heure, limités à ceux du premier filtre SHEF de l'entonnoir de VSM-G, car le choix du second filtre dans le cadre de cette étude n'a pas encore été déterminé (cf. les perspectives dans la prochaine section). Néanmoins, les résultats à la suite du filtrage par SHEF sont prometteurs. Le tableau ci-après, dans lequel les grandes lignes de ce criblage haut-débit sont résumées, indique les résultats chiffrés en rapport avec quelques données expérimentales.

Données de départ

Chimiothèque : ~600 000 molécules

- Compilation de composés disponibles commercialement, pré-filtrage suivant les règles de Lipinski, et exclusion des molécules toxiques et trop réactives.
- Analyse conformationnelle avec 400 confs. maxi / molécule => **~90 000 000 conformères**.

Cible LXRβ : 10 conformations

- 3 conformations holo : codes PDB : 1P8D, 1PQ6 et 1PQ9.
- 7 conformations apo : échantillonnage à partir de la trajectoire de 30 ns d'une simulation de DM et clustérisation des conformations.

Résultats préliminaires

Évaluation du criblage virtuel

- Résultats SHEF : sélection du top-10 000 molécules sur chaque conformation de la cible.
- Comparaison avec les molécules déjà testées expérimentalement chez Fournier/Solvay.

Principaux résultats

- 4% des molécules sélectionnées par SHEF sont présentes dans la chimiothèque de la société Fournier/Solvay dont certains composés « hits LXR expérimentaux » : taux de réussite ou « hit rate » multiplié par 8.
- Parmi les autres molécules sélectionnées par SHEF, présentes dans la chimiothèque Fournier/Solvay mais non encore testées expérimentalement, 80% ont un fort intérêt potentiel car issues d'une sélection en amont selon des critères spécifiques (diversité, *etc.*).
- Les conformations sélectionnées à partir de la DM ont apporté une diversité significative de nouveaux ligands présentant un intérêt potentiel, au côté des trois structures cristallographiques utilisées pour le criblage. En modélisant la plasticité de la protéine LXRβ, elles participent à l'augmentation de l'enrichissement.

TAB. 2 - Principaux éléments de la campagne de criblage haut-débit.

I-7. Conclusion et perspectives

Le choix d'une stratégie de criblage virtuel dépend essentiellement de la nature du système à étudier et des moyens de calculs à disposition. Ainsi un protocole, efficace pour un système donné, ne le sera pas forcément pour un autre. Le protocole choisi pour étudier la cible LXRβ, s'il n'est donc pas nécessairement réutilisable sur n'importe quel autre système, m'aura conduit, à travers sa mise en œuvre, à acquérir une expérience précieuse. A l'avenir, cette expertise devrait me permettre de définir plus aisément un protocole spécifique de criblage virtuel.

Les travaux présentés dans cette partie ont concerné l'aspect filtrage (première étape d'un criblage haut-débit) de VSM-G, plutôt que sa capacité à classifier les molécules candidates en sortie de l'entonnoir et qui sont destinées, en tant que touches potentielles, à être testées expérimentalement.

Concernant le filtrage, les développements prévus visent principalement à ajouter la prise en charge des propriétés physico-chimiques, soit directement sur les surfaces utilisées par SHEF, soit à travers la mise en place d'un ou plusieurs filtres supplémentaires au sein de VSM-G, et ce afin d'améliorer la pertinence des prédictions. Il est aussi envisagé d'intégrer un module "*ligand-based*" de QSAR (relation quantitative structure-activité) [234, 235] à utiliser en amont de l'entonnoir de criblage de VSM-G.

La partie la moins représentée dans ce travail, qui concerne la classification des molécules à la suite du filtre SHEF, est toujours en réflexion car plusieurs options, non exclusives, sont possibles. Il est, par exemple, envisageable d'utiliser un programme de docking à la suite de SHEF (comme dans l'article #2 avec GOLD) qui soit si possible gratuit et à code ouvert. Une étude approfondie, visant à évaluer l'intérêt au sein de notre stratégie multi-étapes de différents programmes de docking disponibles, est d'ailleurs en cours dans notre laboratoire. En outre, en complément ou en remplacement de tels programmes de docking, une alternative consisterait à employer une méthode basée sur les empreintes structurales d'interaction [236, 237] afin d'évaluer d'une manière plus qualitative les modes de liaison des poses des ligands proposées par SHEF. À l'issue de cette procédure, l'optimisation des meilleurs complexes pourrait se faire par le biais de techniques telles qu'un minimiseur MM/PBSA par exemple [222]. Ces techniques plus élaborées en aval du premier filtre SHEF sont cruciales pour, entre autres, limiter le taux de faux positifs, permettant ainsi de raffiner les résultats des études de criblage à grande échelle, telles que celle décrite précédemment sur LXRβ.

L'étude de criblage à large échelle nous a permis de nous confronter aux limitations liées à la manipulation d'une quantité considérable de données. C'est ainsi que les développements techniques envisagés à court terme s'orientent également vers une meilleure automatisation des calculs massivement distribués (leur déploiement sur grilles de calculs, en particulier) et vers la conception d'une base de données relationnelle, afin de disposer d'un environnement unifié au niveau du stockage des connaissances. La mise en place d'une base de données s'avère déjà indispensable pour conserver les données issues des différents criblages réalisés (en particulier, les coefficients des surfaces générées pour le filtrage géométrique) et assurer leur accessibilité. Cela nous permettra aussi de réutiliser aisément ce patrimoine dans de futures campagnes. Une base de connaissances sur les données protéine-ligand est également en développement ; elle pourrait permettre, entre autres, en association avec les données de criblage accumulées, de proposer des contraintes pour guider le processus de criblage plus efficacement dans l'entonnoir de VSM-G.

Enfin, comme dans toute approche purement géométrique, SHEF, en tant que premier filtre de la stratégie multi-étapes de VSM-G, présente de sérieuses limitations dès lors que les systèmes étudiés ont des sites actifs plats ou très ouverts. En effet, à la différence des protéines comportant une cavité bien fermée, ce type de site actif ne possède pas de contraintes géométriques suffisamment discriminantes pour qu'un filtre purement géométrique soit efficace. C'est pourquoi, en parallèle avec les développements méthodologiques de la plateforme, il nous a semblé nécessaire d'étudier en détails un système protéine-ligand pour lequel VSM-G, dans sa version actuelle et avec le protocole employé pour étudier LXRβ, n'était pas utilisable. Ainsi, l'étude du domaine FAT (*Focal Adhesion Targeting domain*) de la protéine kinase FAK (*Focal Adhesion Kinase*) fait l'objet de la seconde partie résultats de ce travail. Une valorisation de l'expertise acquise à travers cette seconde application pourrait apparaître sous la forme d'indications qui permettraient de guider l'adaptation du protocole de VSM-G à de tels sites actifs particuliers.

II. ÉTUDE DU DOMAINE FAT DE LA KINASE D'ADHÉRENCE FOCALE FAK

Cette deuxième section de résultats concerne des travaux ciblant l'activité du domaine FAT de FAK. Le but recherché est clairement d'étendre le champ des connaissances sur FAK et, à ce titre, nous ne cherchons plus explicitement à faire du développement méthodologique. La plateforme VSM-G autour de laquelle orbitaient les résultats précédents n'est d'ailleurs pas adaptée à une cible telle que le domaine FAT, comme nous venons de l'évoquer.

Tout d'abord, nous passerons en revue l'état actuel des connaissances biologiques et structurales (notre contribution exceptée) concernant d'abord la protéine FAK, puis son domaine FAT. Nous résumerons ensuite nos objectifs initiaux et hypothèses de travail au sein du projet interdisciplinaire auquel nous avons participé. Une étude du système-cible sera alors présentée. Celle-ci contribue à éclaircir certaines interrogations le concernant et a permis de valider un modèle adapté aux simulations par dynamique moléculaire. Sur cette base, la mise au point de composés présumés actifs sera enfin décrite. La validation expérimentale de ce travail (tests biologiques des composés) est en cours.

II-1. Présentation de la protéine FAK

Les complexes d'adhérence (ou d'adhésion focale) sont situés sous la membrane cellulaire et forment les points de liaison entre la matrice extracellulaire et le cytosquelette (cf. FIG. 24) [238, 239]. Au sein de ces complexes, la kinase d'adhérence focale ou FAK (pour *Focal Adhesion Kinase*) [240, 241] occupe une place centrale en tant qu'intermédiaire pour la propagation de signaux émanant de divers récepteurs transmembranaires, en particulier ceux des intégrines [242-245]. Ainsi, cette protéine cytoplasmique ubiquitaire joue un rôle fondamental dans le contrôle d'importants processus cellulaires tels que l'adhérence, la prolifération, la survie et la motilité des cellules [245-248].

FAK possède de nombreux résidus tyrosine qui, une fois phosphorylés, agissent comme des points d'amarrage pour le recrutement et l'activation de plusieurs protéines de signalisation à domaines SH2[*] et SH3[†] [249]. Son activation se traduit par l'autophosphorylation d'un résidu tyrosine libérant un site

[*] Ce type de domaine constitue un motif protéique d'environ 100 acides aminés qui reconnaît et lie spécifiquement les tyrosines phosphorylées et qui joue un rôle clé dans le relais des cascades de signalisation.

[†] Les domaines SH3 sont des séquences d'environ 50 acides aminés qui reconnaissent et lient spécifiquement les régions riches en proline.

de liaison de haute affinité pour les kinases de la famille de Src notamment, dont les cascades de transduction aboutissent à l'activation de programmes géniques de régulation du cycle cellulaire et de migration. Ainsi, en tant que protéine adaptatrice, FAK interagit avec des protéines du cytosquelette telles que l'actine, la paxilline, la taline, aussi bien qu'avec les protéines participant au remaniement de ce même cytosquelette telles que PI3K ou p130Cas.

La dérégulation de plusieurs processus cellulaires pour lesquels FAK joue un rôle central est associée, à différents niveaux (prolifération, survie, développement, migration), au développement de cellules cancéreuses. Pour cette raison, FAK constitue une cible thérapeutique majeure de la lutte anti-cancer [250-254].

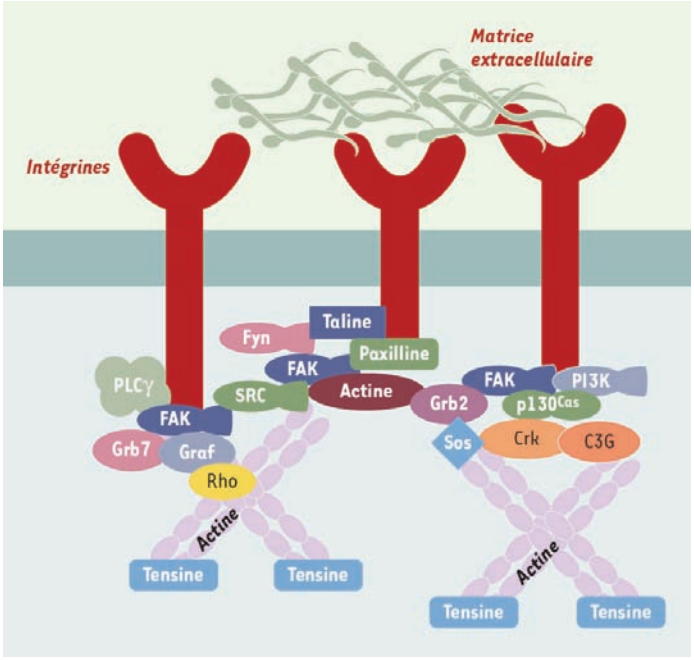


FIG. 24 – Représentation schématique du complexe d'adhérence focale. Tiré de (Cornillon et al., 2003). [247]

II-1.1. Organisation de la protéine FAK

FAK a un poids moléculaire de 125 kDa pour plus de 1 000 acides aminés. Elle forme, avec PYK2 (*proline-rich tyrosine kinase 2*), la famille des protéines tyrosine kinases (PTK) cytoplasmiques riches en proline et dépourvues de récepteurs membranaires [245, 247]. Sa séquence, qui est très conservée au cours de l'évolution, comporte une très grande identité au sein des espèces mammifères et autres organismes eucaryotes plus simples [245, 255]. FAK présente une grande homologie de séquence et de structure avec PYK2, en particulier dans la nature et l'agencement de ses différents domaines [243].

La structure de FAK est constituée de trois domaines principaux : un domaine amino-terminal *FERM* (bande *Four.one-Ezrin-Radixin-Moesin*), un domaine catalytique *Kinase* et un domaine carboxy-terminal *FAT* (*Focal Adhesion Targeting domain*). D'autres zones moins "structurées" se trouvent entre les domaines FERM et Kinase (région dite de "*linker*") et entre Kinase et *FAT* (région riche en proline, souvent désignée sous le terme "*pro-rich*").

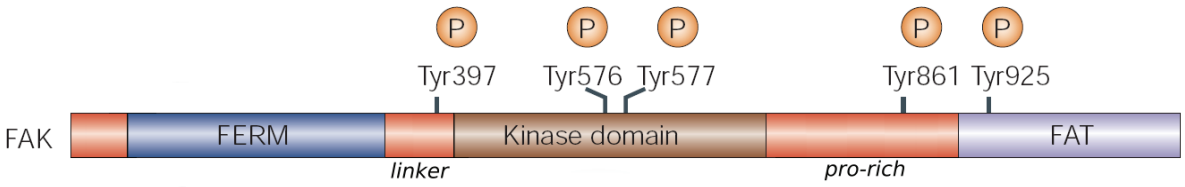


FIG. 25 – Organisation de la protéine FAK et localisation des principaux sites de phosphorylation. Modifié de (Mitra et al., 2005). [256]

Le domaine N-terminal FERM (~ 330 résidus) interagit avec les domaines cytoplasmiques des récepteurs transmembranaires, notamment avec les intégrines (sous-unité β) [240, 246]. Il peut se lier au domaine Kinase voisin, ce qui donne lieu à une forme inactivée (fermée) de FAK [257]. À l'inverse, un autre type d'interaction, de nature intermoléculaire cette fois, entre FERM et l'ensemble de la protéine FAK, a également été montré comme essentiel pour garantir l'activité optimale du domaine Kinase [258]. Ainsi, le domaine FERM est capable d'autoréguler dans les deux sens l'activité de la protéine FAK dont il fait partie.

Le domaine Kinase (~ 270 résidus) est un domaine typique à activité catalytique comportant deux résidus tyrosine qui jouent un rôle important dans l'activité de FAK (Tyr_{576} et Tyr_{577}). En effet, une fois phosphorylés, ils augmentent l'activité enzymatique de FAK [259] et permettent la création de nouveaux sites de liaison pour recruter divers partenaires [260]. La région *linker* le précédant (~ 50 résidus) contient le résidu Tyr_{397}, dont la phosphorylation est cruciale pour l'activation de FAK [261].

La partie C-terminale de FAK comprend une région riche en proline (~ 210 résidus) et le domaine FAT (~ 140 résidus). La région riche en proline peut se lier au domaine SH3 de différentes protéines du complexe d'adhésion, participant ainsi à sa réorganisation [245, 262, 263]. Comme son nom l'indique, le domaine FAT (*Focal Adhesion Targeting domain*) est le domaine indispensable pour localiser FAK au sein du complexe d'adhérence focale, par l'intermédiaire de liaisons avec la Paxilline et la Taline [264-267]. Les résultats présentés dans ce chapitre concernent uniquement le domaine FAT de FAK. Le rôle, la structure et l'intérêt de FAT seront détaillés dans la section suivante II-2.

II-1.2. Activation de la protéine FAK

Dans son état basal, la protéine FAK est inactive du fait des interactions intramoléculaires FERM/Kinase évoquées précédemment [257, 268]. La transition vers une conformation active débute par l'interaction de FAK avec différents récepteurs transmembranaires [269, 270], en particulier les intégrines [271]. Cela induit un changement conformationnel et la formation de dimères de FAK [272] sur lesquels l'autophosphorylation du résidu Tyr_{397} est autorisée [257] (cf. FIG. 26). La présence de la phosphotyrosine $pTyr_{397}$ caractérise la forme activée de FAK : une mutation de Tyr_{397} inhibe toute possibilité d'activité biologique ultérieure de la forme inactive de FAK [261].

Dans FAK activée, $pTyr_{397}$ constitue un site de liaison de forte affinité pour les domaines SH2 des protéines kinase de la famille Src [245, 246, 252]. La formation de complexes FAK-Src activé [261, 273] entraîne la phosphorylation des résidus Tyr_{576}, Tyr_{577}, Tyr_{861} et Tyr_{925}. Cela permet à FAK de recruter des partenaires additionnels [259, 260]. Les complexes FAK-Src sont alors impliqués dans de multiples voies de signalisation associées à des processus biologiques variés [243, 246, 261, 274]. Par exemple, sur la partie C-terminale de FAK, la phosphorylation de Tyr_{925} conduit au recrutement du domaine SH2 de la protéine Grb2 entraînant l'activation de la cascade de transduction Ras-MAPK par FAK [275, 276].

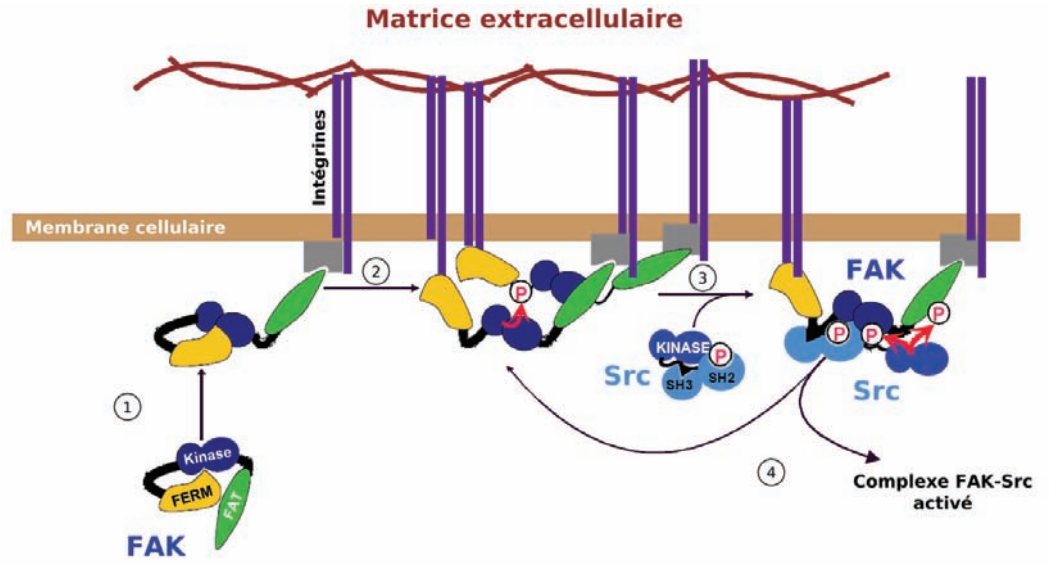


FIG. 26 – Illustration des grandes étapes du mécanisme d'activation de FAK. 1) Recrutement de FAK au sein du complexe d'adhérence focale, près des intégrines. 2) Changement conformationnel de FAK permettant sa dimérisation nécessaire à son activation (autophosphorylation de Tyr_{397}). 3) Amarrage de Src sur la phosphotyrosine $pTyr_{397}$. La formation du complexe FAK-Src permet la phosphorylation d'autres sites et l'activation de plusieurs voies de signalisation.

II-1.3. Principaux processus cellulaires régulés par FAK

Comme mentionné précédemment, FAK est au cœur d'un système régulateur impliquant de nombreuses protéines du complexe d'adhésion. À travers les cascades de signalisation qu'elle active, FAK joue un rôle important dans le contrôle de divers processus biologiques comprenant la survie, la prolifération et la migration des cellules [246, 247].

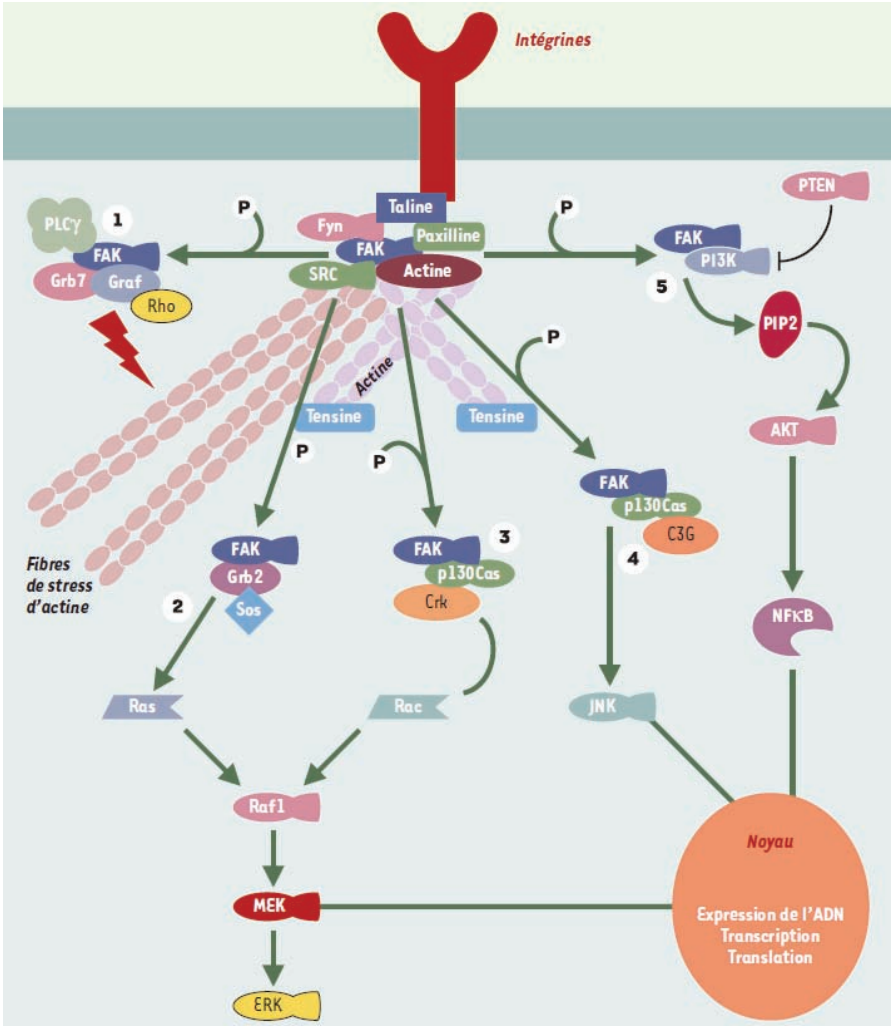


FIG. 27 – Schéma des principales voies de signalisation de l'adhérence focale, toutes associées à l'activité de FAK. Elles participent au remaniement du cytosquelette et permettent également de réguler plusieurs processus vitaux pour la cellule tels que la survie, la prolifération et la migration cellulaires. D'après (Cornillon et al., 2003). [247]

Ces processus cellulaires sont déréglés pendant le développement d'un cancer. C'est ainsi que la multiplication anarchique de cellules anormales, favorisée par le rôle anti-apoptotique de FAK [277], peut se concentrer et conduire à la formation de tumeurs. Dans de nombreux cas, ces tumeurs peuvent avoir un caractère invasif et migrer dans l'organisme ; se développent alors des métastases.

La figure et le tableau ci-après résument les principales protéines activées par FAK qui participent à ces processus biologiques dans les contextes de cellule saine et de cellule cancéreuse.

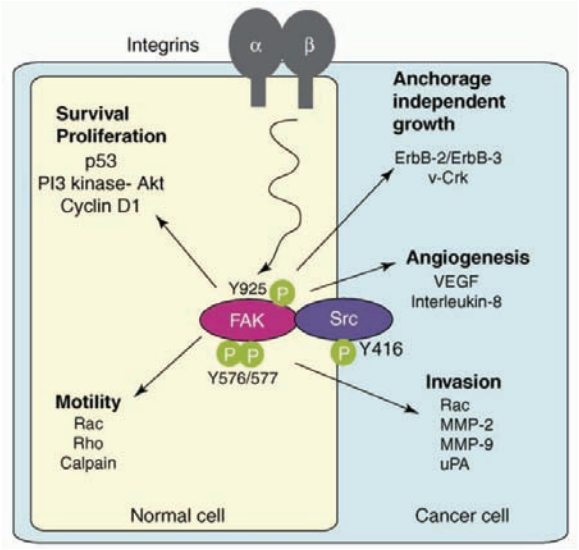


FIG. 28 – Schéma simplifié des interactions protéiques de FAK au cours des principaux processus cellulaires dans lesquels elle joue un rôle capital. En jaune : la survie, la prolifération et la migration, au sein d'une cellule saine. En bleu et dans l'ordre : la protection contre l'anoïkis, l'angiogénèse† et l'invasion, au sein d'une cellule cancéreuse. D'après (Mitra et al., 2006)* [278].

Processus biologique	Protéines des voies de signalisations activées par FAK	Références
Survie	p53	[279, 280]
	PI3K → PKB → Bad → GSK3	[281, 282]
	(Src) → p130Cas → Ras → JNK	[283]
Prolifération	Grb2 → MAPK → Cycline D1	[275]
	PKC/PI3K → Rb → Cycline D3	[284]
Migration	Calpaïnes	[285]
	Src/p130Cas → MAPK	[286]
	Src/p130Cas/PI3K → MAPK	[287]
	GSK3/PP1	[288]
Angiogénèse	MEK → ERK → VEGF	[289]
Invasion	MMPs	[290]
	Erb2/3	[291]

TAB. 3 – Résumé des principales protéines des voies de signalisation activées par FAK, impliquées dans la survie, la prolifération, la migration, l'angiogénèse et l'invasion. Reproduit de (van Nimwegen et al., 2007). [254]

* Anoïkis : mécanisme de mort cellulaire suite à la perte de contact de la cellule avec la matrice extracellulaire.

† Angiogénèse : mécanisme de croissance des vaisseaux sanguins, qui, dans sa forme pathologique, permet d'irriguer des tumeurs cancéreuses.

II-1.4. FAK, une cible d'intérêt thérapeutique pour le traitement du cancer

Les preuves reliant FAK à de multiples processus de cancérisation sont nombreuses [250-254]. En particulier, sa surexpression, associée à un comportement invasif, est observée dans de nombreuses variétés de cellules et de tumeurs cancéreuses, tandis que son expression est très faible dans les tissus sains et les néoplasmes bénins [250, 253, 254, 292, 293]. Le niveau d'expression de FAK dans les tissus constitue par conséquent un indicateur de choix pour le pronostic et le diagnostic du cancer. Le fait que l'inhibition de la signalisation et/ou de l'expression de FAK puisse bloquer le développement tumoral et la formation des métastases [250] illustre également à quel point cette protéine présente un grand intérêt comme cible thérapeutique pour le traitement du cancer [253, 254].

Étant donné que FAK est ubiquitairement exprimée à travers les tissus, on pourrait émettre des réserves sur la potentielle toxicité qui pourrait résulter de son inhibition. Toutefois, des études effectuées sur certaines lignées cellulaires ont montré que l'inhibition de FAK pouvait entraîner une diminution de la motilité des cellules malignes et provoquer leur apoptose, sans pour autant avoir d'effets néfastes sur les cellules saines [294-296]. Ces résultats prometteurs ont conduit plusieurs groupes de recherche à concourir au développement de thérapies anti-cancéreuses ciblant FAK spécifiquement.

Deux stratégies distinctes peuvent être envisagées pour mettre au point de nouvelles thérapies pouvant interférer avec l'action de FAK dans les processus d'invasion tumorale et de développement des métastases.

La première est de concevoir des composés compétiteurs de l'ATP pouvant inhiber le domaine catalytique Kinase de FAK afin d'empêcher son action dans l'activation des multiples voies de signalisation lui étant associées. Récemment, plusieurs groupes ont publié des études sur la mise au point de nouveaux inhibiteurs ciblant le site catalytique de FAK [297-300]. L'un d'entre eux, le PF-00562271, passé en phase clinique I, semble particulièrement prometteur au vu de sa faible toxicité associée à son potentiel anti-tumoral [301].

La seconde stratégie consiste à tirer partie du rôle central de FAK au sein du complexe d'adhérence focale en tant que protéine adaptatrice. Cela vise à développer des composés dans le but d'entraver l'accessibilité des tyrosines requises pour le recrutement des partenaires de FAK au cours de son activation, ou encore de bloquer des sites d'interactions protéine-protéine de FAK qui sont essentielles pour sa localisation dans le complexe d'adhésion focale, impérative pour son activation. A ce jour, une seule étude publiée fait l'objet de ce type de stratégie sur FAK. Se basant sur des cellules du cancer du sein, un peptide de 12 acides aminés a été montré comme efficace en bloquant l'interaction de FAK avec une autre kinase, VEGFR-3, pour entraîner spécifiquement l'apoptose des cellules malignes, sans

avoir d'effet sur les cellules saines [302]. Une partie de ce travail, à travers la conception *in silico* de peptidomimétiques ciblant le domaine FAT de FAK, est basée sur la seconde stratégie.

II-1.5. Accessibilité structurale des domaines de FAK

La structure tridimensionnelle de la protéine FAK dans son intégralité n'a pas encore été déterminée. Seules les structures des trois domaines principaux la constituant ont été résolues expérimentalement. En effet, le domaine N-terminal FERM (codes PDB : 2AEH et 2AL6) [303], le domaine catalytique Kinase (codes PDB : 1MP8 [304] et 2J0L [305]) et dernièrement l'ensemble FERM-Kinase (code PDB : 2J0J) [305] ont été obtenus par diffraction RX. Le domaine FAT, pour sa part, a été déterminé par diffraction RX et par RMN, sous sa forme libre (codes PDB : 1K40 [306], 1K04-1K05 [307] et 1PV3 [308]) ou complexé à des fragments de la Paxilline (codes PDB : 1KTM [309], 1OW6-1OW7-1OW8 [310] et 1QVX [311]).

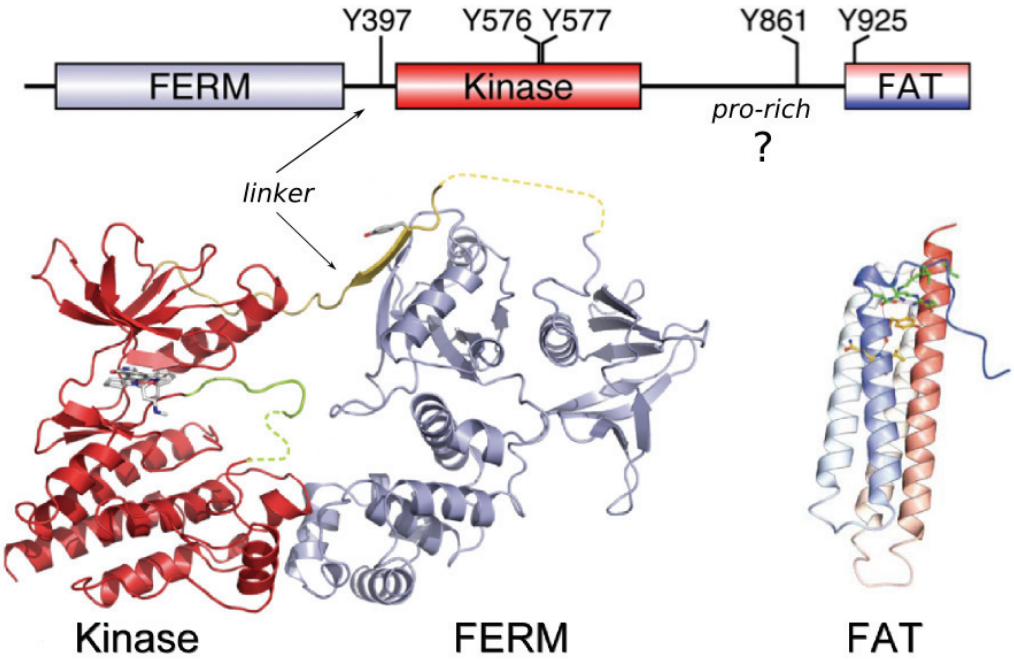


FIG. 29 – Structure de la protéine FAK. Organisation de ses domaines et aperçu de certaines structures tridimensionnelles disponibles des domaines FERM-Kinase (code PDB : 1MP8) [305] *et FAT (code PDB : 1K04)* [307]. *Adapté de (Lietha et al., 2005).* [305]

Les portions de FAK résolues expérimentalement, comme on peut le voir dans la FIG. 29, constituent à ce jour un puzzle pour lequel une seule pièce est manquante : il s'agit de la section riche en proline, reliant les domaines Kinase et FAT. Cette région de FAK est délicate à déterminer expérimentalement car très flexible et peu organisée. La construction d'un modèle complet de FAK implique donc, en premier lieu, la re-création de la structure tridimensionnelle de ce fragment, en recourant, par exemple, à des méthodes de prédiction de structures secondaires [312]. L'étape suivante consiste ensuite à relier de façon pertinente les différents fragments géométriques. Cette opération peut faire intervenir des techniques de docking protéine-protéine [166] et se baser, par exemple, sur une enveloppe de la protéine en solution obtenue par une expérience de SAXS (*Small Angle X-ray Scattering*) [313, 314]. En collaboration avec nos collègues de la biologie structurale (équipe de Stefan Arold du Centre de Biochimie Structurale de Montpellier - UMR/CNRS 5048), la construction d'un tel modèle est en cours au sein de notre laboratoire. Il est prévu, en fin de processus, de soumettre le modèle à une simulation par dynamique moléculaire afin de tester sa robustesse et, au cas où celle-ci s'avérerait satisfaisante, procéder éventuellement à des optimisations locales. Enfin, la dernière étape de validation du modèle consistera à déterminer une série de mutations pertinentes, par l'identification d'une liste de résidus prédits par le modèle pour avoir un rôle crucial dans l'activité de la protéine. Des expériences de mutagenèse dirigée permettront alors de confronter le modèle - via ces prédictions - à la réalité expérimentale.

II-1.6. Objectifs et positionnement de nos travaux

La description de la protéine FAK a montré combien différents angles d'étude existent pour mieux comprendre son fonctionnement au sein de la cellule et pour essayer de développer de nouvelles thérapies anti-cancéreuses la ciblant. En effet, FAK comporte divers domaines sur lesquels il est possible d'intervenir pour la mise au point de composés anti-métastatiques. Pour ce travail, nous nous sommes intéressés au domaine FAT, d'une part pour son intérêt comme cible thérapeutique importante, et d'autre part parce que l'étude d'un tel domaine pouvait mener à de nouvelles réflexions sur les développements méthodologiques nécessaires afin d'élargir le champ d'application de la plateforme VSM-G.

Dans la section suivante, nous dresserons donc une description détaillée du domaine FAT. Nous présenterons ensuite les résultats de nos études portant sur la dynamique de FAT et sur la conception *in silico* de petites molécules peptidomimétiques ciblant ce domaine.

II-2. Fonctions et caractérisation du domaine FAT de FAK

II-2.1. Rôles biologiques

Le domaine FAT (*Focal Adhesion Targeting domain*) joue un rôle crucial dans la localisation de la protéine FAK au sein du complexe d'adhérence focale. A travers sa liaison avec certaines protéines liées aux intégrines et l'état de phosphorylation de sa tyrosine 925 qui constitue une sorte d'interrupteur, il régule la formation et le renouvellement du complexe. Le rôle du domaine FAT conditionne l'activation de FAK et donc la capacité de cette kinase à enclencher les cascades de signalisation contrôlant divers processus biologiques, lesquels sont suractivés dans les phénomènes de croissance tumorale et de développement métastatique.

Localisation de FAK au sein du complexe d'adhérence focale

Le domaine FAT est la région C-terminale de FAK nécessaire et suffisante pour permettre son attachement au sein du complexe d'adhérence focale, à proximité des intégrines [315, 316]. Il s'agit d'une condition nécessaire à l'activation de FAK, précédant sa participation à diverses cascades de signalisation (voir FIG. 26) [243, 246]. L'attachement de FAK au complexe d'adhérence focale s'effectue au moyen de liaisons entre FAT et différentes protéines partenaires, notamment la Paxilline [264, 265, 267, 317, 318] et la Taline [266, 319].

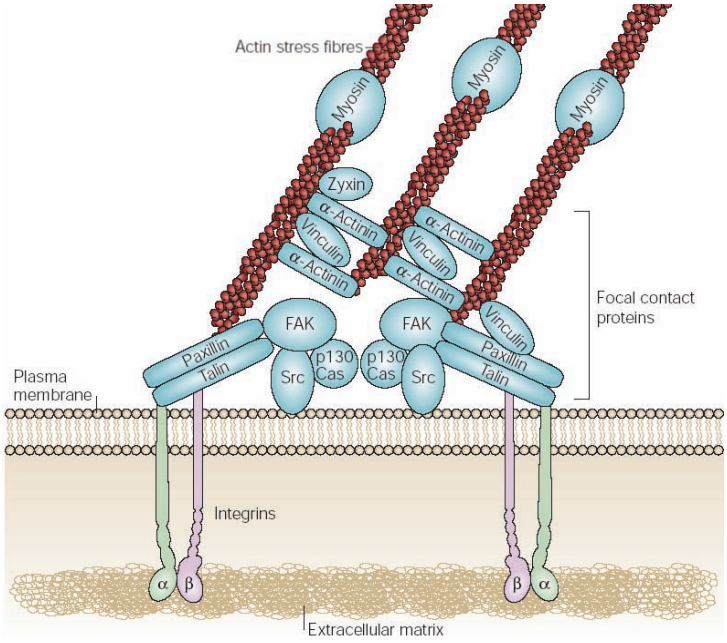


FIG. 30 – Représentation simplifiée de l'architecture du complexe d'adhérence focale. FAK est y recrutée grâce à ses interactions avec la Paxilline et la Taline, protéines liées aux intégrines. Tiré de (Mitra et al., 2005). [256]

La séquence des événements menant à l'attachement de FAK au complexe d'adhérence focale ne demeure que partiellement connue à ce jour. Le cas de la liaison entre FAK et la Paxilline en donne l'illustration. En effet, tandis que certains travaux indiquent que cette liaison semble essentielle pour la localisation de FAK dans les adhésions focales [266, 267], d'autres concluent exactement l'inverse [320]. De telles contradictions pourraient indiquer que l'association de FAK au complexe d'adhérence focale s'effectue par des mécanismes complexes faisant intervenir des partenaires multiples, parmi lesquels la Paxilline et la Taline [266].

Décrochage du complexe d'adhésion et activation de la voie Ras-MAPK

Le domaine FAT contient la tyrosine 925, qui est un site de phosphorylation par la protéine Src au cours de l'activation de FAK [275]. Une fois phosphorylée, cette tyrosine représente un site de liaison de haute affinité pour recruter le domaine SH2 de la protéine adaptatrice Grb2 [276]. La liaison entre Grb2 et FAK est à l'origine d'un des mécanismes d'activation de la voie de signalisation Ras-MAPK [275, 276]. Cette cascade de signalisation, essentielle notamment dans la croissance et la régulation cellulaire, est suractivée dans le processus de développement tumoral. Cette voie de signalisation fait donc l'objet d'une recherche accrue pour la mise au point de thérapies anti-cancéreuses dont l'intervention peut avoir lieu à diverses étapes de son déroulement [321-323].

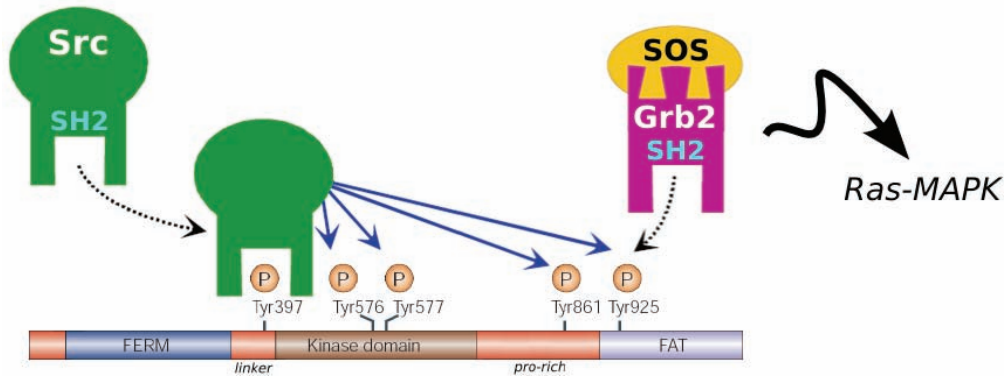


FIG. 31 – Schéma de l'activation de la cascade Ras-MAPK par FAK. Une fois phosphorylée par la kinase Src, la Tyr_{925} constitue un site de liaison spécifique pour recruter le complexe Grb2-SOS. La formation du complexe de signalisation FAK-Grb2-SOS va permettre l'activation de Ras enclenchant la cascade des MAP-kinases.

En initiant la voie de signalisation Ras-MAPK, la formation du complexe FAK-Grb2-SOS entraîne le décrochage de FAK des adhésions focales [324]. Ce processus participe ainsi à la réorganisation du cytosquelette à travers la dynamique de structuration/déstructuration du complexe d'adhérence focale.
Bien que la position de Tyr_{925} au sein du domaine FAT (qui sera détaillée dans la partie II.2.2), la rende relativement accessible et exposée, les interactions du domaine FAT avec certaines protéines telles que la Paxilline peuvent, par gêne stérique, limiter la capacité de ce site à être phosphorylé

efficacement par Src et ensuite de recruter la protéine Grb2. De telles associations protéine-protéine peuvent limiter ou bloquer l'accessibilité de Tyr_{925} et ainsi l'activation, entre autres, de la voie Ras-MAPK. Cela peut expliquer pourquoi, en comparaison des autres sites de phosphorylation de FAK, Tyr_{925} soit faiblement phosphorylée par Src *in vivo* [325]. Il a, par contre, été démontré que la phosphorylation de Tyr_{925}, joue un rôle crucial dans le déclenchement de l'angiogénèse et donc dans la progression des tumeurs cancéreuses, via la cascade FAK → Grb2 → Ras → MAPK [289]. Par conséquent, l'accessibilité de Tyr_{925} représente un des leviers de régulation de l'activité de FAK, sur lequel il est possible d'agir afin de bloquer la croissance tumorale et le développement métastatique, dans le contexte de cellules malignes, en provoquant leur apoptose.

II-2.2. Structure tridimensionnelle

FAT constitue un domaine d'environ 140 acides aminés, en position C-terminale de FAK. Comme évoqué précédemment, la structure du domaine FAT est disponible dans la PDB sous forme libre [306-308] et complexée à des fragments de la Paxilline [309-311]. A ce jour, aucune structure de FAT co-complexée, même partiellement, avec une autre protéine que la Paxilline n'est disponible.

Les structures obtenues par RMN et par diffraction RX ont, à une exception près (que nous évoquerons à la fin de cette partie), la même architecture : la structure secondaire se présente sous la forme d'un fagot compact (et très bien conservé à travers les espèces) de quatre hélices-α, reliées entre elles par de courtes boucles (coudes-β) [306-311]. Il comporte un cœur hydrophobe qui assure la stabilité d'une telle architecture [306]. Le résidu tyrosine 925, dont le rôle vient d'être présenté, est positionné en début de la première hélice (H1) (cf. FIG. 32).

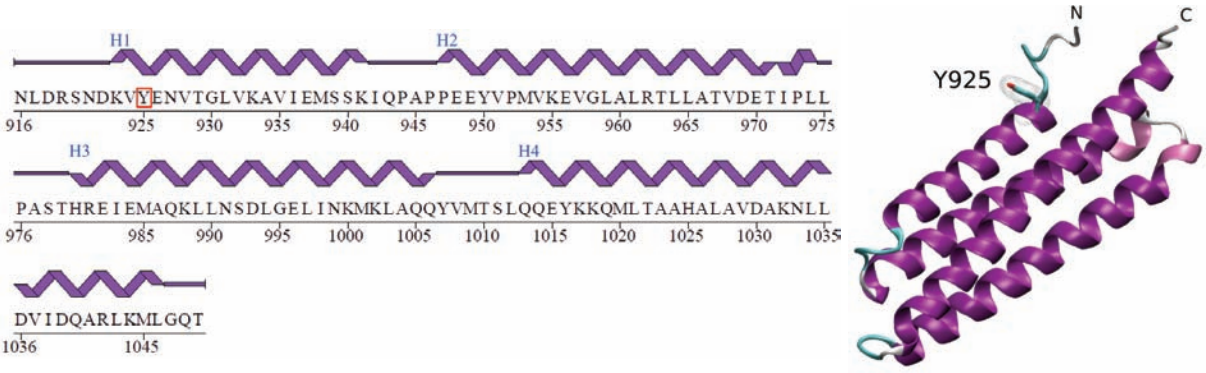


FIG. 32 – Structure du domaine FAT (code PDB : 10W7). (a) Représentation de la séquence et de la structure secondaire (image générée par le serveur PDBsum [326]*). (b) Structure tridimensionnelle en fagot d'hélices sur laquelle est indiquée la position de* Tyr_{925}.

Déterminants structuraux de l'interaction FAT-Paxilline

L'intégrité de la structure de FAT est nécessaire pour lier la Paxilline. En effet, des études ont montré que des mutations ponctuelles sur certains résidus formant le cœur hydrophobe de FAT conduisent à l'incapacité de FAT à lier la Paxilline, ce qui est probablement du au fait de la désorganisation de sa structure en fagot [267].

Le domaine FAT comporte deux sites de liaison à la Paxilline identifiés entre les hélices H1-H4 et H2-H3 et situés sur deux faces opposées du fagot d'hélices. Il s'agit de zones majoritairement hydrophobes et bordées par quelques résidus chargés et connues sous le nom de *patchs hydrophobes*. Les fragments de Paxilline impliqués dans la liaison avec FAT sont, quant à eux, deux régions riches en leucine appelées *motifs LD*[*] [327, 328]. Ces motifs LD (LD2 et LD4) sont observés partiellement sur les structures expérimentales (cf. FIG. 33), sous la forme de peptides amphiphiles hélicoïdaux [309-311].

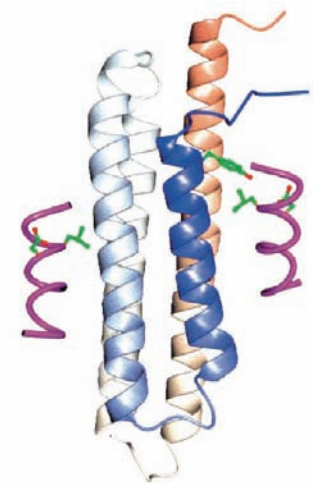


FIG. 33 – Structure de FAT en complexe avec deux motifs LD de la Paxilline. Les deux domaines LD (en mauve) se lient sur les deux régions hydrophobes de FAT entre les hélices H1-H4 et H2-H3. La suite de résidus L et D caractérisant leur séquence, ainsi que la Tyr_{925} de FAT sont montrés en bâtons verts. D'après (Hoellerer et al., 2004) [310].

La liaison simultanée de LD2 et LD4 a été montrée comme nécessaire à la formation d'un complexe FAT-Paxilline stable [329]. Par conséquent, une des stratégies en vue de moduler la localisation de FAK au sein du complexe d'adhérence focale pourrait consister à cibler les sites de liaison de FAT à la Paxilline pour empêcher la formation du complexe FAK-Paxilline [317, 330].

[*] Les motifs LD sont des courtes séquences répétitives comportant la suite consensuelle "**LD**XLLXXL" et constituant des interfaces pour lier de manière spécifique certaines protéines du complexe d'adhésion focale.

Déterminants structuraux nécessaires à la phosphorylation de Tyr_{925}

La façon dont la tyrosine 925 de FAT se fait phosphoryler par Src et la séquence des événements qui y conduisent sont encore largement sujettes à débat et plusieurs questions restent sans réponse :
- Comment les événements "phosphorylation de Tyr_{925} par Src" et "liaison de FAT à la Paxilline" sont-ils liés et avec quelle chronologie ?
- Une réorganisation du fagot (ou du moins de l'hélice H1) est-elle requise pour rendre Tyr_{925} accessible au site catalytique de Src, ce qui rendrait possible sa phosphorylation ? Si oui, cela est-il dû à l'ouverture de H1 par rapport au reste du fagot ou à la désorganisation de cette hélice ?

En ce qui concerne la première question, la structure de FAT en complexe avec la Paxilline montre que la position du motif LD localisé entre les hélices H1 et H4 obstrue en grande partie l'accessibilité de la tyrosine clé Tyr_{925} (cf. FIG. 33). Dans une telle conformation du domaine FAT, la phosphorylation par Src de ce résidu et le recrutement ultérieur de la protéine Grb2 par la phosphotyrosine semblent donc difficiles. Certains résultats montrent d'ailleurs que la liaison de la Paxilline avec FAT et la phosphorylation du résidu Tyr_{925} sont vraisemblablement deux événements mutuellement exclusifs (cf. FIG. 34) [308-310, 331].

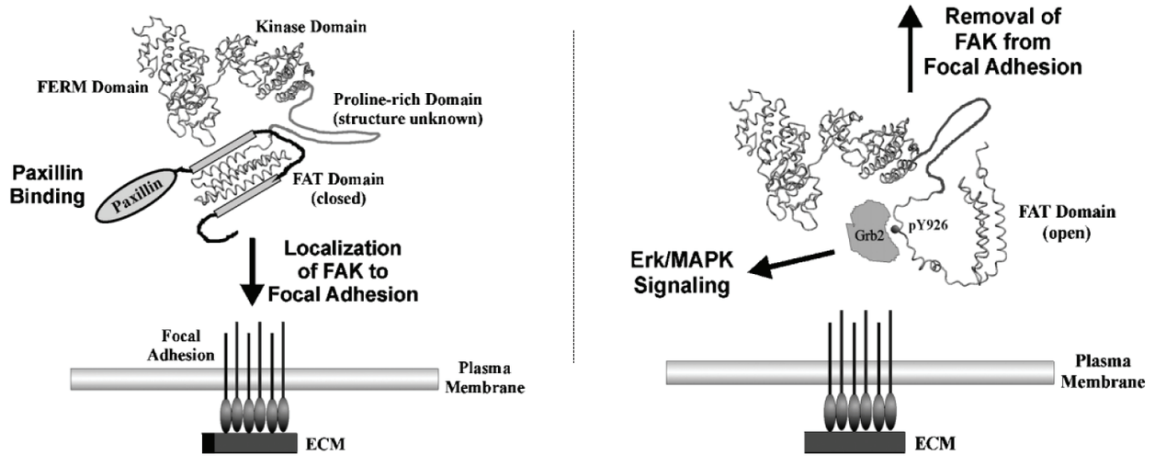


FIG. 34 – Modèle proposé illustrant le rôle biologique des formes ouverte et fermée du domaine FAT. Tiré de (Dixon et al., 2004). [331]

Quant à la désorganisation du fagot par son hélice H1, des études structurales indiquent que les kinases du type de Src et la protéine Grb2 se lient à des substrats dont les structures secondaires sont respectivement en brin-β [332, 333] et en coude-β [334]. Il est donc proposé que l'hélice H1 portant Tyr_{925} doive se déformer afin d'adopter, même partiellement, une conformation ouverte ou étendue. Celle-ci pourrait permettre, parmi d'autres hypothèses sur sa fonction [267], la phosphorylation puis l'association avec Grb2 [307, 309, 331]. Une telle dynamique de H1 est envisagée dans plusieurs publications [308, 331].

Cependant, on ne peut exclure la possibilité que des changements conformationnels moins drastiques que ceux cités ci-dessus puissent également conduire à un positionnement propice à la phosphorylation de Tyr_{925} dans le site catalytique de Src. Un ancrage de Src par son domaine SH3 sur le domaine *pro-rich* de FAK (en amont de FAT), conduisant à un changement conformationnel de la jonction *pro-rich*/FAT, pourrait positionner favorablement le début de l'hélice H1 de FAT et ainsi rendre possible la phosphorylation de Tyr_{925} [335].

Déterminants structuraux nécessaires à la désolidarisation de H1 du reste du fagot

Tandis que la plupart des structures expérimentales de FAT montrent un domaine en fagot compact, une autre conformation bien distincte a également été résolue [307]. Celle-ci se présente sous la forme d'un dimère dans lequel l'hélice H1 d'un monomère, capable de se dissocier du fagot, s'associe avec les hélices H2-H3-H4 de l'autre monomère et *vice-versa* (cf. FIG. 35b). Bien que cette forme semble stable *in vitro*, sa pertinence dans les conditions physiologiques semble discutable et serait par conséquent un artefact dû, sans doute, aux conditions de cristallisation [307]. Si l'existence d'une telle forme dimérique est peu probable *in vivo*, elle montre tout de même qu'un tel changement conformationnel pourrait avoir lieu, impliquant la boucle en coude-β reliant les hélices H1 et H2. Des études par RMN indiquent que cette hypothèse est plausible car cette région charnière est relativement flexible en solution [308, 331]. La boucle H1-H2 serait alors cruciale dans la désolidarisation de l'hélice H1 par rapport au reste du fagot. Nous allons à présent tester cette hypothèse.

a) **b)**

FIG. 35 – Les différentes conformations adoptées par le domaine FAT. a) Superposition des structures disponibles de FAT (codes PDB : 1OW6, 1KTM, 1QVX, 1PV3, 1K04 et 1K40). Les fagots se superposent relativement bien, excepté la structure 1K40 pour laquelle l'hélice H1 est désolidarisée des autres hélices. b) Structure 1K40 cristallisée sous forme d'un dimère avec échange d'hélices, d'après (Arold et al., 2002) [307].

II-3. Influence de la région charnière entre H1-H2 du domaine FAT dans la régulation de la protéine FAK

Ce travail, qui vise à étudier la dynamique du domaine FAT dans la régulation de l'activité de la protéine FAK, a été effectué en collaboration avec nos collègues biologistes de Paris (équipe de Jean-Antoine Girault, Université Pierre et Marie Curie, Paris VI - UMR/INSERM 536) et biophysiciens de Montpellier (équipe de Stefan Arold du Centre de Biochimie Structurale de Montpellier - UMR/CNRS 5048).

Comme nous l'avons vu dans la précédente section, la flexibilité du domaine FAT semble jouer un rôle déterminant dans la localisation intracellulaire et l'activité de FAK à travers : 1) ses liaisons avec les protéines du complexe d'adhésion focale (notamment la Paxilline) et 2) la régulation de la phosphorylation de sa tyrosine 925. D'après certaines études, la flexibilité de la région charnière entre les hélices H1 et H2 de FAT pourrait être responsable de l'ouverture de l'hélice H1 par rapport au reste du fagot [307, 308, 331, 336]. Ce changement conformationnel conduirait alors à rendre plus accessible Tyr_{925}.

Dans le but de tester cette proposition, nos collègues expérimentateurs ont étudié des formes mutantes de FAK dans lesquelles la séquence de la boucle entre les hélices H1 et H2 a été modifiée afin de favoriser ou de limiter l'ouverture de l'hélice H1. Pour cela, avaient été considérés des acides aminés pouvant en augmenter la flexibilité ou au contraire la réduire, cette modulation pouvant aussi se faire en allongeant ou en raccourcissant cette boucle.

L'hypothèse de travail est la suivante : l'introduction de résidus proline, associée à un raccourcissement de la boucle et à la réduction de la flexibilité de celle-ci, provoque, par un effet de "ressort tendu", une ouverture plus facile de H1. Au contraire, l'introduction de résidus glycine et le maintien de la longueur de la boucle augmentent la flexibilité de celle-ci, conduisant à une tension moindre, insuffisante pour modifier la conformation originale du fagot de quatre hélices. Ainsi, quatre mutants de FAK ont été considérés : trois supposés comporter, pour le domaine FAT, une boucle H1-H2 plus ou moins souple et un autre présentant, au contraire, une boucle plus tendue. Le comportement de ces divers mutants par rapport à la forme sauvage été étudié *in vivo*.

En parallèle, nous avons mené une série de simulations par dynamique moléculaire à la fois sur la forme sauvage (WT), sur la forme présentant une boucle tendue (T-FAT) et sur une des formes flexibles (R-FAT (G_3)). Notre but était de caractériser précisément, au niveau moléculaire, la

réorganisation du domaine et de pouvoir décrire les effets dynamiques globaux occasionnés par les mutations.

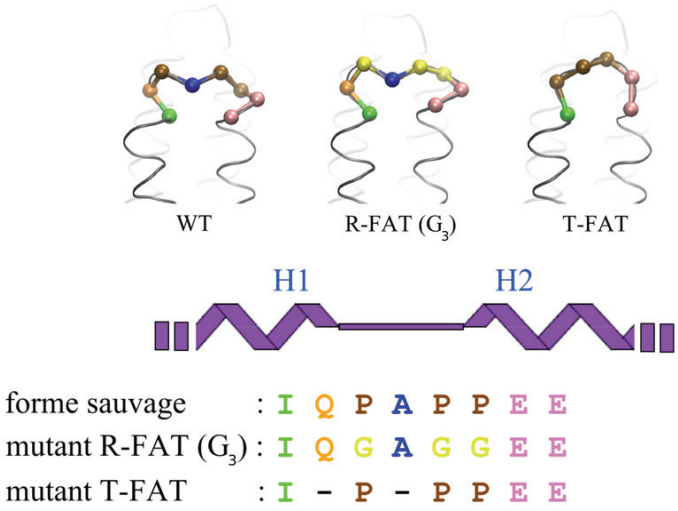


FIG. 36 – Séquence de la boucle reliant les hélices H1 et H2 des formes du domaine FAT qui ont été simulées par dynamique moléculaire. En comparaison à la forme sauvage (WT), les mutants R-FAT et T-FAT ont été imaginés pour comporter une boucle plus souple ou plus tendue respectivement.

Ce travail a conduit à la rédaction de l'article "*Conformation of the Focal Adhesion Targeting domain is critical for FAK interactions in cells*" qui est en révision.

Les résultats expérimentaux démontrent clairement l'effet de la flexibilité de la boucle entre H1 et H2 sur la dynamique de FAT. Cette boucle joue donc un rôle très important aussi bien dans la localisation de FAK au sein du complexe d'adhérence focale que dans la modulation de la phosphorylation de Tyr_{925} :

- L'augmentation de la flexibilité (mutants R-FAT) conduit à une baisse de la phosphorylation et au maintien de la liaison à la Paxilline.
- La diminution de la flexibilité (mutant T-FAT) a pour effet une accentuation de la phosphorylation et l'absence de liaison à la Paxilline.

L'interprétation de ces résultats, présentée dans l'article en terme d'ouverture ou de maintien de l'hélice H1, est faite au travers de l'hypothèse énoncée auparavant concernant l' "effet ressort" : l'augmentation de la flexibilité conduisant à une diminution de cet effet et au maintien de l'hélice H1 dans le fagot ; la diminution de la flexibilité provoquant l'effet inverse et l'ouverture de H1.

Cependant, nos résultats de simulation par dynamique moléculaire montrent que, dans nos conditions de simulation, les trois protéines considérées (WT, R-FAT (G_3) et T-FAT) conservent l'arrangement en fagot et aucune ouverture de H1 n'a eu lieu. Néanmoins, selon les mutants et par rapport à la forme sauvage, des modifications au niveau du début de l'hélice H1 ont été observées. Ainsi, l'observation de la trajectoire du mutant T-FAT montre que cette extrémité N-terminale laisse la tyrosine 925 plus accessible que dans la forme sauvage. A l'opposé, pour le mutant R-FAT, la Tyr_{925} est la plupart du temps masquée par l'extrémité N-terminale de FAT au cours de la simulation ; phénomène induit par la réorganisation d'un pont salin impliquant deux résidus des hélices H1 et H4. Si on corrèle cette accessibilité de Tyr_{925} avec les niveaux de phosphorylation mesurés *in vivo*, nos résultats sont en bon accord avec l'expérience.

Dans l'état actuel de nos résultats, nous ne pouvons pas expliquer les différences de comportement de FAT observées *in vivo* vis-à-vis de la Paxilline, car nous n'observons pas de réorganisation majeure du fagot. À quoi cela est-il dû ? Diverses explications sont possibles, par exemple : soit nos simulations n'ont pas été assez longues pour observer l'ouverture de l'hélice H1, soit un autre phénomène que l'ouverture de H1 doit être considéré. En effet, des essais d'ancrage manuel de la kinase Src sur un modèle tridimensionnel de FAK complet montrent que l'ouverture de H1 ne semble pas indispensable pour présenter Tyr_{925} au site catalytique de Src (travail de modélisation entrepris dans notre laboratoire par Bernard Maigret – cf. partie II-1.5). En revanche, des modifications de l'extrémité N-terminale de H1 semblent pour cela nécessaires ; une telle hypothèse a déjà été proposée [331, 337]. Des travaux supplémentaires de modélisation seront nécessaires pour clarifier la situation, en employant, par exemple, la dynamique moléculaire forcée "*Steered Molecular Dynamics*" (SMD) [338]. Une telle technique a déjà été utilisée dans une étude portant sur le dépliement de la structure secondaire de FAT dans sa forme sauvage [339].

II-4. Article #4

"Conformation of the Focal Adhesion Targeting domain is critical for FAK interactions in cells", *en révision.*

ARTICLE #4

"Conformation of the Focal Adhesion Targeting domain is critical for FAK interactions in cells"

CONFORMATION OF THE FOCAL ADHESION TARGETING DOMAIN IS CRITICAL FOR FAK INTERACTIONS IN CELLS

Gress Kadaré [1,2,3], Said el Messari [1,2,3], Karen Brami-Cherrier [1,2,3], Marie-Claude Boutterin [1,2,3], Alexandre Beautrait [4], Bernard Maigret [4], Stefan T. Arold [5,6], and Jean-Antoine Girault [1,2,3]

[1] *INSERM, U839, F-75005, Paris, France;*
[2] *Université Pierre et Marie Curie (UPMC-Paris 6), F-75005, Paris;*
[3] *Institut du Fer a Moulin, F-75005, Paris;*
[4] *Nancy Université, LORIA, Orpailleur team, UMR 7503, F-54506, Vandœuvre-lès-Nancy Cedex, France;*
[5] *INSERM, Unité 554, Montpellier, France;*
[6] *Université de Montpellier, CNRS, UMR 5048, Centre de Biochimie Structurale, Montpellier.*

Address correspondence to:
Jean-Antoine Girault, INSERM U839, Institut du Fer à Moulin, 17 rue du Fer à Moulin, 75005 Paris, France.
Tel: +33 (0)1 45 87 61 50; Fax: +33 (0)1 45 87 61 59; e-mail: girault@fer-a-moulin.inserm.fr

ABSTRACT

Focal adhesion kinase (FAK) is a non-receptor tyrosine kinase transducing signals from integrin-based adhesion complexes and a variety of membrane receptors. FAK is enriched at focal adhesions through interactions of its focal adhesion targeting (FAT) domain with paxillin and talin. The FAT domain is formed by a bundle of 4 alpha-helices in which the first helix (H1) can dissociate from the others. Phosphorylation of Tyr-925 by Src-family kinases requires partial opening of the bundle and may release FAK from focal adhesions. The role of the FAT domain dynamics in intact cells is not known. To investigate this role, we introduced mutations in the hinge (loop 1) between H1 and H2 designed to hinder (relaxed, R-FAK: replacement of Pro-944, -946, -947 by Gly), or facilitate (tense, T-FAK: deletion of Gln-943 and Ala-945) the FAT bundle opening. The mutations did not alter the capacity of FAK to autophosphorylate on Tyr-397. In contrast, phosphorylation of Tyr-925 was decreased in R-FAK and increased in T-FAK. This difference was more pronounced following orthovanadate treatment or Fyn co-transfection. Molecular dynamics simulations disclosed some rearrangements of the bundle that may participate in these effects. Enrichment at focal adhesions was less pronounced for T-FAK than wild type or R-FAK. Paxillin binding was decreased in T-FAK and increased in R-FAK, whereas the interaction with talin was unaltered. These experiments show that the first hinge controls the properties of the FAT domain in cells and suggest that its flexibility is an essential parameter of FAK function.

KEYWORDS

Focal adhesion kinase; tyrosine kinase; focal adhesion targeting; paxillin; cell adhesion; conformational changes; tyrosine phosphorylation.

INTRODUCTION

FAK plays a major role in transducing signals downstream from integrins and other membrane receptors [1, 2]. It is involved in the control of cell growth, survival and migration [3, 4]. FAK is also important for tumor invasiveness and metastasis and is a promising target for anticancer drugs [5]. The biological importance of FAK is demonstrated by the lethality of its null mutation around midgestation [6]. Following integrin engagement FAK is recruited to focal adhesions and undergoes a series of phosphorylation reactions including autophosphorylation and phosphorylation by Src-family kinases. Several phosphorylated tyrosines have been identified in FAK. Autophosphorylation of Tyr-397, located in the linker between the N-terminus and the catalytic domain, creates a binding site for Src family kinases [7]. Subsequent phosphorylation of FAK by Src on several residues, including Tyr-576 and Tyr-577 enhances its activity [8]. On the other hand phosphorylated Tyr-925 may serve as a Grb2-binding site [9] and facilitate release of FAK from focal adhesions [10].

The central catalytic domain of FAK is flanked by an N-terminal FERM domain [11] and a C-terminal region, which encompasses two proline-rich motifs and a focal adhesion targeting (FAT) domain, essential for FAK function [12]. The FERM, kinase and FAT domains are well conserved among metazoans [13]. The structure of the FAT domain has been determined by X-ray crystallography [14, 15] and NMR [16]. It is composed of four helices that form a “right-turn” elongated bundle. The four-helix FAT domain displays two binding sites on opposite sides of the molecule that bind paxillin LD peptides in a helical conformation [17, 18].

The first helix (H1) of the purified recombinant FAT domain has the capacity to dissociate from the rest of the bundle and this can lead to the formation of stable dimers with swapped H1 helices *in vitro* [15]. A physiological relevance of these helix-swapped FAT dimers is unlikely; rather they are thought to be an *in vitro* artifact that results from the intrinsic propensity of FAT domains to expulse their first helix. The resulting conformational changes in FAT appear crucial for its cellular role since Tyr-925, located on FAT H1, is poorly accessible, and its phosphorylation and interaction with the SH2 domain of Grb2 requires a rearrangement of the bundle [15]. A FAT mutant, in which the hydrophobic core is partly destabilized, lacks paxillin binding, but is more phosphorylated on Tyr-925, is still targeted to focal adhesions and has an increased propensity to form dimers in solution [19, 20]. NMR and discrete molecular dynamics studies have shown that the hinge between H1 and H2 samples multiple conformations in solution and that H1 can loose its helical character [20, 21]. These dynamic features have been proposed to be important for FAK biological properties.

The driving force behind H1 opening is probably the P-944APP motif situated in the narrow FAT H1-H2 hinge region [15]. Indeed, experimental evidence on several proteins has established that proline-rich linker regions can act as torsion-energy loaded molecular springs, promoting structural transitions and arm-exchange dimers [22-24].

To address the role of the FAT domain dynamics in intact cells we have produced and studied mutant forms of full length FAK in which the FAT H1-H2 hinge region was modified in order to increase or decrease its propensity to open. We show that mutations which increase tension in the hinge region decrease paxillin binding without preventing focal adhesion targeting or talin binding, and facilitate phosphorylation of Tyr-925. Mutations which decrease tension in the hinge enhance paxillin binding and decrease Tyr-925 phosphorylation. Thus, our results underline the biological importance of FAT dynamics *in vivo* for FAK function.

MATERIAL AND METHODS

Antibodies and Reagents

Monoclonal (4.47) anti-FAK antibody was from Upstate Biotech and used at a dilution of 1:1000, polyclonal anti-FAK A17 (1:2000 for immunoblotting) was from Santa Cruz Biotechnology and anti-phospho-Y397 from Biosource (1:2000). Polyclonal anti-phospho- Y925-FAK antibodies (1:2000) were from Cell Signaling and monoclonal anti-phosphotyrosine, clone 4G10 (1:2000), from Upstate Biotech. Monoclonal antibodies anti-talin, clone 8D4 (1:1000) and anti-VSV (1:500, for immunofluorescence) were from Sigma. Monoclonal antibody anti-VSV clone P5D4 was from Roche Molecular Biochemicals (1:500 for immunoprecipitation). Monoclonal anti-paxillin antibody (1:1000) was from Zymed. Polyclonal affinity purified anti-VSV antibodies (1:2000 for immunoblotting) were a generous gift from M. Arpin, Curie Institute, Paris). An anti-VSV rabbit serum [25] was used for FAK-paxillin immunoprecipitation. All other reagents were from Sigma unless otherwise noted.

Cell Culture and Transfection

COS7 cells were cultured in Dulbecco's modified Eagle's medium supplemented with 10% fetal bovine serum, and transfected with 8 µg of DNA per 100 mm diameter culture dish, in the presence of polyethyleneimine as described previously [26]. Total DNA quantity was maintained constant with empty pcDNA3 plasmid. Cells were lysed 48 h after transfection. For immunofluorescence experiments CHO-K1 (American Type Culture Collection) cells were maintained in Ham's F12 medium supplemented with 7.5% fetal calf serum, 1 mM glutamine, 100 units/ml penicillin and 100 µg/ml streptomycin. Transfection was carried out with Lipofectamine 2000 (Invitrogen) according to the manufacturer's instructions.

Localization of FAK by immunofluorescence

CHO cells were rapidly rinsed in PBS, fixed in PBS containing 4% paraformaldehyde (PFA) for 10 min at RT, and then permeabilized by incubation for 10 min in PBS containing 0.1% TritonX100. After saturation with 5% bovine serum albumin for 30 min at RT, the coverslips were incubated for 120 min in 1:500 dilution of anti-VSV antibody, washed twice in PBS containing 0.1 % BSA and incubated with a 1:300 dilution of Alexa 488-conjugated polyclonal anti-mouse antibody.

F-actin was visualized directly utilizing TRITC-phalloidin (1:2000, Sigma). The coverslips were washed four times with PBS and mounted in Mowiol (Sigma). Cells were examined at the *Institut du*

Fer à Moulin cell imaging facility, with a Leica TCS SP II (Leica Microsystems, Heidelberg, Germany) confocal laser scanning device, equipped with an argon/krypton laser, on a Leica DM IRBE inverted microscope.

DNA constructs and Mutagenesis

The N-terminal tagging of full-length rat FAK (AF 020777) by VSV was realized as follows: first a 300-nucleotide (nt) fragment containing a SacII site in the 3' untranslated region of FAK was eliminated by digestion with BamHI-SmaI, filled by T4 DNA polymerase and self-ligated. Then, a new Sac II site was introduced immediately downstream from FAK ATG start codon, without affecting the primary sequence. Synthetic phosphorylated oligonucleotides corresponding respectively to VSV epitope flanked by semi-SacII sites were introduced in frame with FAK sequence in the newly created SacII site, generating plasmids pBKCMV-VSV-FAK. All constructs were verified by DNA sequencing. In all experiments we used rat FAK° without the Pro-Trp-Arg insertion in FAT domain [27, 28].

Mutations in the FAT region of FAK were produced using an XhoI (nt 2950)-SacI (nt 3430) fragment of FAK subcloned in pBlueScript.SK+ (Stratagene) as a template, corresponding pairs of oligonucleotides designed with the desired mutation, and the Quick Change mutagenesis kit from Stratagene. After validation of every mutation by sequencing, each mutated version of FAT [fragment XhoI (2950)-SacI (3430)] was cloned back to full-length VSV-tagged FAK.

Immunoprecipitations

Forty-eight hours after transfection, COS7 cells were homogenized in modified RIPA buffer (1% Triton X-100, 0.5% sodium deoxycholate, 0.1% SDS, 50 mM Tris [pH 7.4], 150 mM NaCl, 1 mM sodium orthovanadate) and Complete® (Boehringer) protease inhibitors as described [26]. For FAK-paxillin-talin co-immunoprecipitations, cells were lysed in FAK-pax buffer [29] (1% Triton X-100, 150 mM NaCl, 20 mM Tris base, 0.05% Tween-20, 1 mM NaF, 1 mM Na_3VO_4, supplemented with Complete. Lysates were precleared by incubation for 1-3 h at 4°C with 100 µl of a mixture (50% each vol/vol) Sephacryl and protein A- or G- Sepharose beads (GE Healthcare) saturated with 25 mg/ml BSA. Immunoprecipitation was carried out overnight at 4°C with 40 µl of rabbit anti-VSV serum, or 5 µl of monoclonal anti-VSV antibodies. The beads were washed three times with lysis buffer, resuspended in Laemmli loading buffer, placed at 100°C for 2 min and subjected to SDS-polyacrylamide gel electrophoresis.

Separated proteins were transferred to a nitrocellulose membrane (GE Healthcare), immunoblotted with the appropriate antibodies and visualized either with horseradish peroxidase-conjugated anti-mouse/anti-rabbit antibodies (1:4000) and the ECL detection system (GE Healthcare) or with IR-Dye 800 CW donkey anti-mouse/anti-rabbit IgG antibodies (Rockland) and detection by infrared fluorescence with an Odyssey Li-Cor scanner.

Molecular dynamics (MD) simulations

As the whole FAK structure is not known, the simulations were performed on FAT domain only. As a starting structure, we used the wild type FAT domain, PDB entry 1OW7 (16), chain A, from which we removed the co-crystallized peptide fragment of paxillin, and we introduced the R-FAK (G_3) and T-FAK mutations. The three prepared systems were next placed in 80 x 80 x 80 $Å^3$ cubic boxes of TIP3P explicit water. The MD program NAMD [30] was employed in conjunction with the CHARMM27 force field to describe the three systems that encompassed a similar number of atoms (~51,000). Each system was first energy-minimized (6,400 steps of conjugate gradients) and next equilibrated (100 ps of MD). A 40-ns MD simulation was then produced for each of them, with a snapshot structure recorded every picosecond. MD simulations were carried out in the NPT ensemble: Langevin dynamics and Langevin piston methods were applied to keep the temperature (300 K) and the pressure (1 atm) of the system fixed. The equations of motion were integrated with a 2 fs time-step, using a multiple time-steps algorithm so that short- and long-range forces were calculated every 2 and 4 time-steps respectively. The SHAKE algorithm was used to constrain chemical bonds between hydrogen and heavy atoms to their equilibrium value. Long-range electrostatic forces were taken into account using the particle-mesh Ewald approach.

RESULTS

FAT "hinge region" mutations do not alter stability and autophosphorylation of FAK in intact cells

To study the role of the conformational dynamics of FAT domain in intact cells, we mutated the residues of loop 1 "hinge region" between H1 and H2. We altered the strain introduced by the prolines of the P_{944} APP motif (**Fig. 1A**), using the strategy established by Rousseau et al. [24]: to relieve the strain introduced by the three prolines (Pro-944, Pro-946, Pro-947), each of them was replaced by a flexible glycine residue. This produced "relaxed" mutants R-FAK-GP_2: QGAPP, G_2P: QGAGP and G_3: QGAGG, in which the "closed" conformation should be stabilized (**Fig. 1A**). On the opposite, to facilitate the opening of the FAT domain, we produced a "tense" mutant of FAK (T-FAK) by deleting the two non-proline residues surrounding the proline-rich motif (**Fig. 1A**). To distinguish FAK mutants from endogenous FAK molecules, they were tagged at the N-terminus with a VSV-epitope. Immunoblotting with antibodies for VSV (**Fig 1B** upper panel), and FAK (**Fig. 1B**, middle panels), showed the correct expression levels of the mutated proteins.

We studied the autophosphorylation of the various mutated FAK by immunoblotting with a specific phospho-Y397 antibody (**Fig 1B**, lower panel). We found no difference between mutant and wild type FAK, demonstrating that the mutations did not alter the correct folding and stability of the protein or its activation through autophosphorylation.

Molecular dynamics (MD) studies reveal slight rearrangements of Tyr-925 environment in R-FAT-FAK

The use of MD simulations for studying mutations effects on the flexibility of proteins was already successfully reported [31, 32]. Here we used longer timescale MD simulations starting from the known 3-D structure of the FAT domain in which the R-FAK (G_3) and T-FAK mutations were introduced, to assess the consequences of the mutations on the structure of the FAT domain. Examination of the evolution of several geometrical and energetic parameters during the dynamics showed that, after 40 ns of stimulation, no major structural change was detected in the wild type FAT domain. During the whole MD trajectory, the 4-helice bundle was very stable and the Tyr-925 phenol side chain was accessible to the solvent as in the X-ray starting structure (**Fig. 2A**). Similar analysis performed in the R-FAT (G_3) mutant showed a slight modification of the position of helix 1 relative to helix 4, resulting in the loss of the salt bridge between Asp-922 and Arg-1042. The small shift of helix 1 away from helix 4 induced a displacement of the N-terminus (Asn-916) tail resulting in the partial masking of

Tyr-925 (**Fig. 2B**). Moreover, Tyr-925 side chain moves and is more buried inside the bundle, the phenol hydroxyl making favorable hydrogen bond either with Asp-1036 or Asp-1039 side chains. In contrast, in the T-FAT mutant, this Tyr-925 residue was statistically more accessible (**Fig. 2C**) while no significant modification of the bundle structure was observed, as in the wild type. Thus, within the short time-frame of the simulation, the modifications introduced by mutagenesis did not destabilize the FAT domain but induced in the R-FAT (G_3) mutant, slight rearrangements of the helix bundle that could modify the accessibility and position of Tyr-925 side chain. We anticipate that the detection of full destabilization would require much longer times of MD simulation which were beyond our computing capacities.

T-FAT-FAK is less enriched at focal adhesions than wild type or R-FAT-FAK

Wild type vsv-FAK was highly enriched in focal adhesions of transfected CHO cells (**Fig. 3**, upper panel). The same was observed for vsv-R-FAT-FAK mutants (**Fig. 3**, middle panel). In cells transfected with vsv-T-FAT-FAK, although VSV immunoreactivity was also detected at focal adhesions, we observed a relatively intense diffuse cytoplasmic immunofluorescence which indicated an increased proportion of cytosolic FAK (**Fig. 3**, lower panel).

Conformation of the FAT domain regulates Tyr-925 phosphorylation

Tyr-925 of FAK is phosphorylated by the Src family kinases Src and Fyn after recruitment of FAK to focal adhesions [33]. However, structural studies of the FAT domain showed that Tyr-925 is located on H1, incompatible with Src kinase or Grb SH2 interactions, which require Tyr-925 to be in a β-sheet or β-turn conformation, respectively [15, 16]. Consequently, these studies infer that phosphorylation and Grb2 association of Tyr-925 require FAT to be in an open conformation. We examined the role of FAT "hinge region" in modulating phosphorylation of Tyr-925 in full-length FAK in intact cells. First we studied the phosphorylation state of Tyr-925 in FAK R-FAT and T-FAT mutants in transfected COS7 cells. Cells were grown for 48 h either under normal culture conditions (**Fig. 4A**), or in the presence of 50 μM orthovanadate for the last 16 hours (**Fig. 4B**). The 3 R-FAT mutations (P_2G, PG_2, and G_3) led to a decreased level of phosphorylation of Tyr-925 compared to wild type FAK (**Fig. 4A, B**). In contrast, the T-FAT-FAK mutant exhibited a dramatic increase in the phosphorylation level of Tyr-925. A Y925F mutant was used as a negative control in these experiments. The faint signal of Tyr-925 FAK phosphorylation observed in this mutant indicated the level of phosphorylation of the endogenous protein (**Fig. 4B**).

The series of FAT mutants was then co-transfected with B-Fyn to determine how they might serve as substrates of phosphorylation by this kinase in intact cells. The 3 R-FAT-FAK mutants were less phosphorylated than wild type FAK in B-Fyn-transfected cells (**Fig. 4C**). In contrast, T-FAT-FAK was an excellent substrate for phosphorylation by B-Fyn (**Fig. 4C**). As above, the faint phosphorylation signal detected for Y925F mutant indicated the contribution of endogenous FAK. Similar results were obtained using NIH 3T3 cells (data not shown). These results provide strong evidence that mutations in the "hinge region" alter the accessibility of Y925 to Src-family kinases.

FAT hinge mutations alter interactions with paxillin but not with talin

The FAT domain contains binding sites for the FA proteins paxillin and talin, which are both thought to participate in the localization of FAK to FA. We investigated the association of mutated FAT with endogenous paxillin and talin by immunoprecipitation. COS7 cells were transfected with wild type or mutated vsv-FAK, alone or in combination with a plasmid coding for B-Fyn. After 48 hours, cells were lysed and the phosphorylation of Tyr-925 was analyzed by immunoblotting with a phospho-specific antibody (**Fig. 5**, upper panel). Transfected FAK proteins were immunoprecipitated with an anti-VSV serum and precipitates were immunoblotted with monoclonal antibodies directed against FAK, paxillin, and talin, as indicated (**Fig. 5**). Only slight differences in paxillin binding were observed between wild type FAK and R-FAT-FAK (G_3) or Y925F-FAK mutants (**Fig. 5**, lanes 1, 2, and 4, respectively), in the absence of cotransfection with B-Fyn. In contrast, virtually no paxillin immunoreactivity was associated with the T-FAK mutant (**Fig. 5**, lane 3).

The differences in paxillin binding were amplified when B-Fyn was coexpressed with FAK proteins (**Fig. 5**, lanes 5-8). As described above, Fyn strongly enhanced the phosphorylation of Tyr-925 in wild type and, even more so, in T-FAT-FAK, whereas R-FAT-FAK was less phosphorylated and only endogenous FAK phosphorylation was observed in Y925F-FAK-transfected cells (**Fig. 5**, lanes 5-8). The association of FAK with paxillin gave a mirror image of that obtained for Tyr-925 phosphorylation. In Fyn cotransfected cells, paxillin binding was absent in T-FAT-FAK, whereas it was slightly increased in R-FAT-FAK (G_3) and dramatically enhanced in Y925F-FAK (**Fig. 5**, lanes 5-8). Interestingly, these latter mutations revealed a very strong effect of Fyn phosphorylation on paxillin binding, which was hardly detectable in wild type FAK. In the presence of Fyn, binding of paxillin to Y925F-FAK was dramatically enhanced. Thus, it appeared that phosphorylation of Tyr-925 counteracted the increased paxillin binding promoted by Fyn.

DISCUSSION

The FAT domain of FAK is important for its targeting to focal adhesions and its signaling and regulation [12, 34]. I has been previously shown that recombinant FAT is capable of expulsing H1 from its four-helix bundle *in vitro*, and suggested that this structural transformation allows FAK Tyr-925 to become phosphorylated by Src kinases and, subsequently, to associate with Grb2 *in vivo* [15]. Here, we addressed the physiological relevance of FAT structural transitions. Our results show that mutations in the hinge between helices 1 and 2 of FAT, designed to stabilize or destabilize its bundle structure have profound influences on paxillin binding and Tyr-925 phosphorylation in intact cells. As expected, these effects are in opposite directions: mutations designed to stabilize the closed conformation of FAT impair Tyr-925 phosphorylation and increase paxillin binding, whereas the converse is true with mutations which tend to promote opening of the FAT bundle. Interestingly, MD predicted changes in the accessibility of Tyr-925 in R-FAT-FAK after only 40 ns simulation, which could already account for the observations in intact cells. Of course these observations do not preclude more significant alterations of the FAT bundle that would take longer to occur, but they indicate the sensitivity of this residue to distant structural modifications.

The T-FAT-FAK mutants, which had a markedly decreased interaction with paxillin, were still able to accumulate at focal adhesion, although their concentration in the cytoplasm was higher than in the case of wild type or R-FAT-FAK. Since T-FAT-FAK binding to paxillin was dramatically decreased, these observations support previous reports providing evidence that paxillin binding is not the sole mechanism of focal adhesion targeting of FAK [19]. Our observations that the FAT hinge mutations did not alter co-immunoprecipitation with talin indicate that binding to talin may account for the persistent enrichment at focal adhesions.

Indeed a deletion removing amino acids 853-963 of FAK (helix 1 and most of helix 2) was shown not to alter its interaction with talin [35], revealing that the integrity of the four-helix bundle was not required for talin binding. These results indicate that the talin binding site is located within 41 last amino acids of FAK and is not perturbed by the dynamics of helix 1 of FAT.

Our results also show that increased tyrosine kinase activity provided by Fyn co-transfection had two distinct effects. It dramatically increased Tyr-925 phosphorylation and prevented paxillin binding. However, in the Y925F FAK mutant an increased binding of paxillin was apparent. This may be related to a direct effect of paxillin tyrosine phosphorylation on the interaction of the two proteins. Alternatively it could result from an indirect effect, for example through phosphorylation of paxillin by ERK on Ser-83, which has been shown to increase its affinity for FAK [29, 36]. Therefore, it appears clearly that the activation of FAK and the subsequent recruitment of Src-family kinases, has two opposing effects on its interaction with paxillin: on the one hand it increases paxillin affinity for FAK,

presumably through phosphorylation of paxillin; on the other hand it decreases this interaction through phosphorylation of Tyr-925. This underlines the critical switch role of Tyr-925 phosphorylation in the organization of the molecular complexes at focal adhesions.

The critical role of Tyr-925 accessibility can be accounted for by a structural model in which H1 opening allows the helix region surrounding Tyr-925 to unfold, and adopt conformations compatible with kinase and Grb2 interactions, in agreement with a previous report [21]. FAT H1 opening not only abrogates the paxillin binding site between H1 and H4, but also appears to affect the second binding site, located between H2 and H3. Indeed, unconstrained MD simulations performed on the wild type form, starting from the open conformation of the FAT domain, showed that the three remaining helices 2-4 are maintained in a stable 3-helices bundle while rearranging themselves and therefore deconstructing the second LD motif binding site (data not shown). Our results also infer that phosphorylation of Y-925 is incompatible with LD-motif binding to H1-H4. Since previous results suggest that phospho-Y-925 does not block FAT from refolding into the four-helix bundle structure [15], it is likely that the phosphate group directly interferes with LD-motif interaction. Indeed, Tyr-925 is partially covered by the LD2 peptide of paxillin which binds between H1 and H4 of FAT domain [16-18].

Thus, our analysis provides strong evidence that structural transitions of FAT are physiologically relevant, and regulate key functions of FAK. Moreover, our study corroborates that the driving force behind the H1 opening of FAT is the P-944APP motif in the H1-H2 hinge region. The important remaining question is what triggers H1 opening? Experimental evidence and theoretical calculations (Beautrait and Maigret, unpublished observations) show that H1 opening is an energetically very unfavorable transition, requiring the rupture of a large number of hydrophobic and electrostatic interactions between H1 and the rest of the four-helix bundle. Despite the strain introduced by the proline-rich H1-H2 linker, H1 opening is an extremely rare and transient event *in vitro*, having occurred only in about 10% of isolated FAT domains after 12-24 hours [15]. Such long time spans, might suggest that the H1-opening dynamics act as a 'molecular clock' for focal adhesion turnover. On the other hand, it is tempting to speculate that factors recruited to FAK catalyze and stabilize FAT opening.

In conclusion, our study shows that the dynamics of the FAT region is an essential parameter for FAK function in intact cells, regulating its targeting, its interaction with paxillin and its phosphorylation on Tyr-925. We also provide evidence that proline residues in the linker between H1 and H2 in the FAT domain are an essential structural motif for this dynamics *in vivo*.

ACKNOWLEDGEMENTS

We thank Dr. Monique Arpin (Curie Institute, Paris) for providing anti-VSV antibodies. This work was supported in part by grants from *Agence Nationale de la Recherche* (ANR) and *Association pour la recherche contre le cancer* (ARC). AB 3rd year of PhD graduate studies year was funded by ARC.

FIGURES

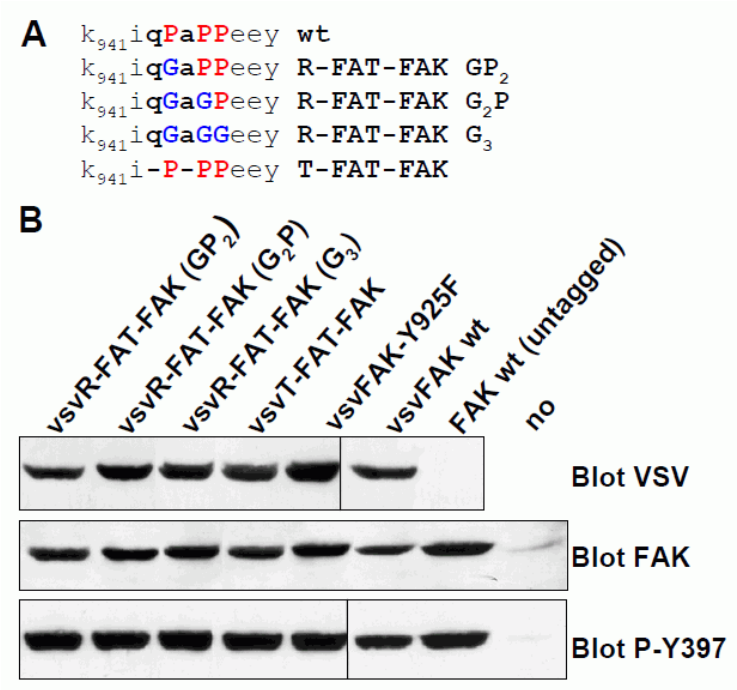


Fig. 1 - Expression of FAK with mutations in the H1-H2 hinge of the FAT domain. A: Mutations designed to release the tension (R) in the hinge region (replacement of Pro-944, Pro-946, and Pro-947 by Gly), or to increase this tension (T) (deletion of Gln-943 and Ala-945). B: Mutant proteins containing an N-terminal VSV tag were expressed by transfection in COS7 cells. Cell lysates were analyzed by immunoblotting with antibodies for VSV, FAK or phospho-Tyr-397-FAK (P-Y397).

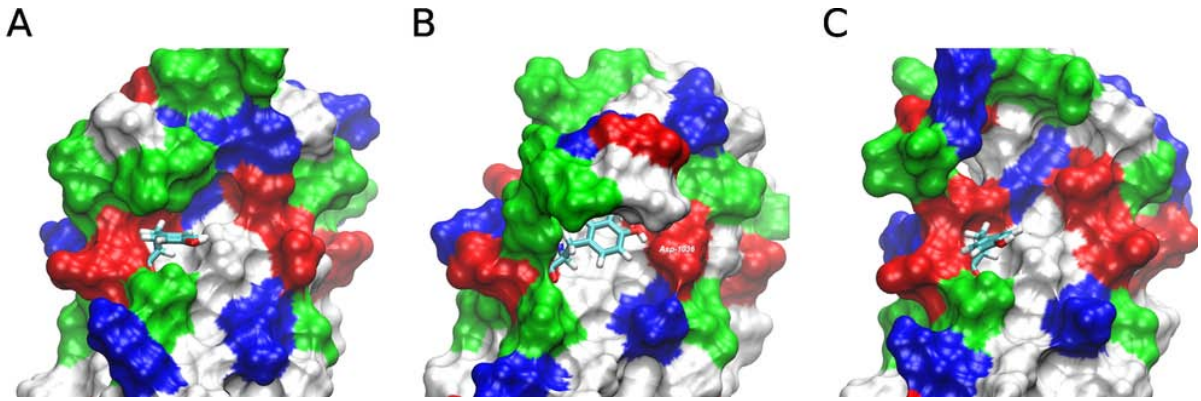


Fig. 2 - MD simulation of mutated FAK structure reveals changes in Tyr-925 accessibility. Comparison of Tyr-925 relative accessibility for the different forms of FAT domain (A: wild-type; B: R-FAT (G_3); C: T-FAT). Rendering was done with VMD [37] and MSMS plugin [38]. Tyr-925 is displayed as sticks and the rest of FAT domain as a surface colored by properties (acidic residues in red, basic in blue, hydrophilic in green, and hydrophobic in white).

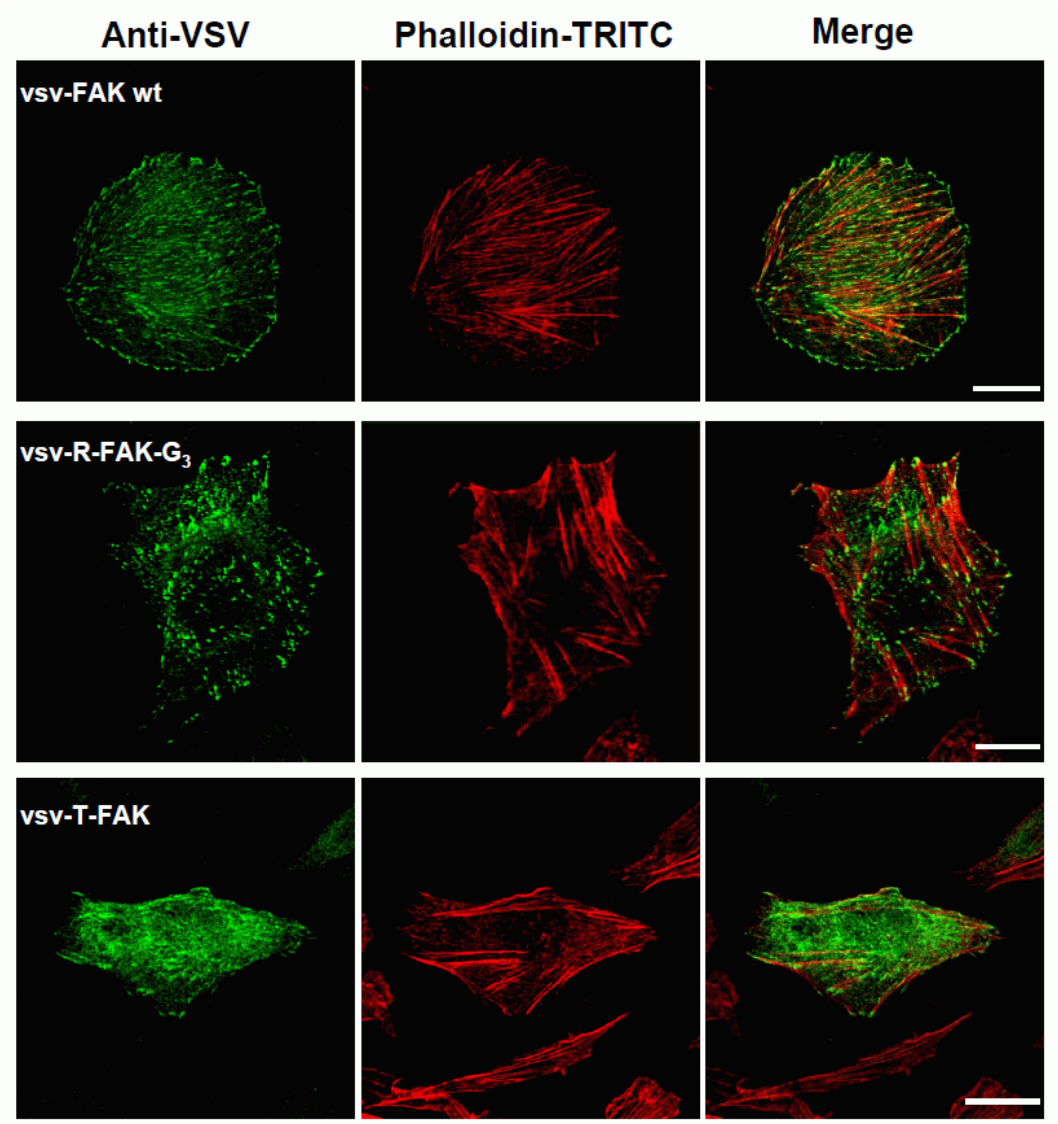


Fig. 3 - Intracellular localization of FAT mutants. CHO cells were transfected with VSV tagged wild type FAK (upper panel), R-FAT-FAK (middle panel) or T-FAT-FAK (lower panel). Localization of FAK was determined by immunofluorescence with VSV antibodies (left column) and actin filaments were visualized with TRITC-phalloidin (middle column). Wild type and R-FAT vsv-FAK immunoreactivity was virtually all in focal adhesions, whereas T-FAT FAK immunoreactivity was also clearly detectable in the cytoplasm. Scale bar: 20 µm.

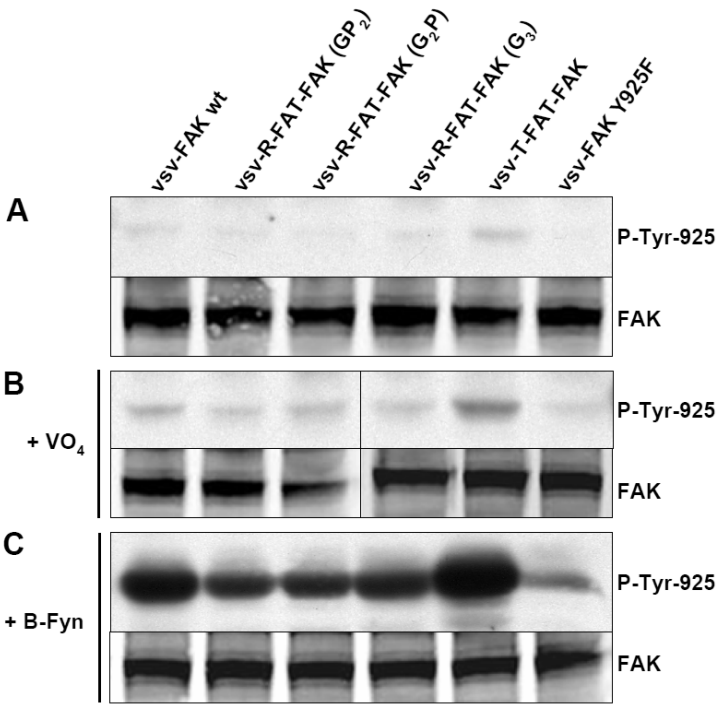


Fig. 4 - Mutation of the FAT hinge alters phosphorylation of Tyr-925. A: Various VSV-tagged FAK constructs were transfected in COS7 cells. Cells were lysed and P-Tyr-925 and FAK analyzed by immunoblotting with specific antibodies. B: Same experiment in which cells were treated for 16 h with 50 µM orthovanadate. C: Same as in A except that cells were co-transfected with FAK and B-Fyn.

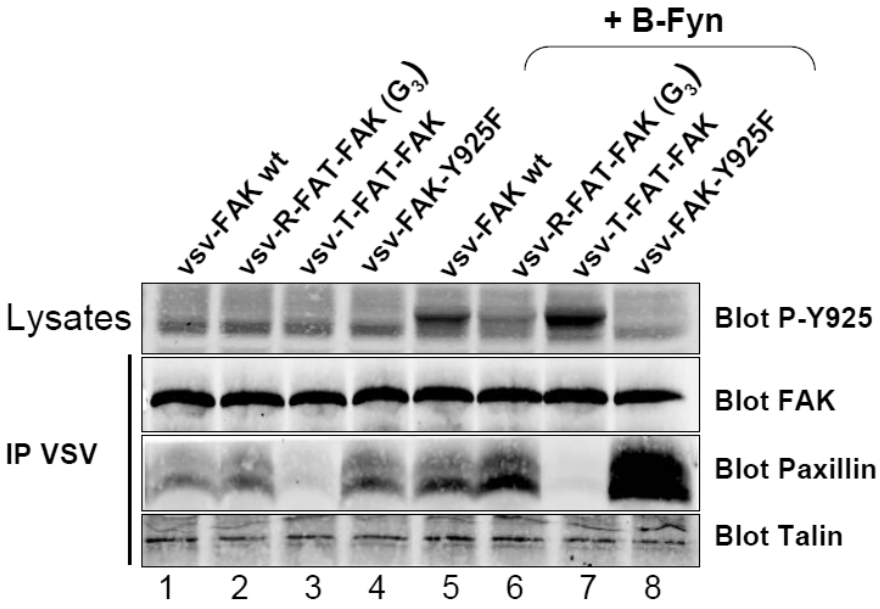


Fig. 5 - Mutation of the FAT hinge alters interaction of FAK with paxillin but not talin. Various VSV-tagged FAK constructs were transfected in COS7 cells alone or with B-Fyn, as indicated. Cells were lysed and FAK immunoprecipitated with anti-VSV antibodies. Immune precipitates were probed with antibodies for FAK, paxillin and talin as indicated. P-Tyr-925 was determined in total cell lysates (upper row).

REFERENCES

(1) Zachary, I.; Rozengurt, E. Focal adhesion kinase ($p125^{FAK}$): A point of convergence in the action of neuropeptides, integrins, and oncogenes. *Cell* **1992**, *71*, 891-894.

(2) Parsons, J. T. Focal adhesion kinase: the first ten years. *J. Cell Sci.* **2003**, *116* (Pt 8), 1409-1416.

(3) Schaller, M. D. Biochemical signals and biological responses elicited by the focal adhesion kinase. *Biochim. Biophys. Acta* **2001**, *1540* (1), 1-21.

(4) Mitra, S. K.; Schlaepfer, D. D. Integrin-regulated FAK-Src signaling in normal and cancer cells. *Curr. Opin. Cell Biol.* **2006**, *18* (5), 516-523.

(5) McLean, G. W.; Carragher, N. O.; Avizienyte, E.; Evans, J.; Brunton, V. G. The role of focal adhesion kinase in cancer: a new therapeutic opportunity. *Nat. Rev. Cancer* **2005**, *5*, 505-515.

(6) Ilic, D.; Furuta, Y.; Kanazawa, S.; Takeda, N.; Sobue, K.; Nakatsuji, N.; Nomura, S.; Fujimoto, J.; Okada, M.; Yamamoto, T.; Aizawa, S. Reduced cell motility and enhanced focal adhesion contact formation in cells from FAK-deficient mice. *Nature* **1995**, *377*, 539-544.

(7) Schaller, M. D.; Hildebrand, J. D.; Shannon, J. D.; Fox, J. W.; Vines, R. R.; Parsons, J. T. Autophosphorylation of the focal adhesion kinase, $pp125^{FAK}$, directs SH2-dependent binding of $pp60^{src}$. *Mol. Cell. Biol.* **1994**, *14*, 1680-1688.

(8) Calalb, M. B.; Polte, T. R.; Hanks, S. K. Tyrosine phosphorylation of focal adhesion kinase at sites in the catalytic domain regulates kinase activity: a role for Src family kinases. *Mol. Cell. Biol.* **1995**, *15* (2), 954-963.

(9) Schlaepfer, D. D.; Hanks, S. K.; Hunter, T.; van der Geer, P. Integrin-mediated signal transduction linked to Ras pathway by GRB2 binding to focal adhesion kinase. *Nature* **1994**, *372* (6508), 786-791.

(10) Katz, B. Z.; Romer, L.; Miyamoto, S.; Volberg, T.; Matsumoto, K.; Cukierman, E.; Geiger, B.; Yamada, K. M. Targeting membrane-localized focal adhesion kinase to focal adhesions: roles of tyrosine phosphorylation and SRC family kinases. *J. Biol. Chem.* **2003**, *278* (31), 29115-29120.

(11) Girault, J. A.; Labesse, G.; Mornon, J. P.; Callebaut, I. The N-termini of FAK and JAKs contain divergent band 4.1 domains. *Trends Biochem. Sci.* **1999**, *24* (2), 54-57.

(12) Hildebrand, J. D.; Schaller, M. D.; Parsons, J. T. Identification of sequences required for the efficient localization of the focal adhesion kinase, $pp125^{FAK}$, to cellular focal adhesions. *J. Cell Biol.* **1993**, *123*, 993-1005.

(13) Corsi, J. M.; Rouer, E.; Girault, J. A.; Enslen, H. Organization and post-transcriptional processing of focal adhesion kinase gene. *BMC Genomics* **2006**, *7*, 198.

(14) Hayashi, I.; Vuori, K.; Liddington, R. C. The focal adhesion targeting (FAT) region of focal adhesion kinase is a four-helix bundle that binds paxillin. *Nat. Struct. Mol. Biol.* **2002**, *9* (2), 101-106.

(15) Arold, S. T.; Hoellerer, M. K.; Noble, M. E. The structural basis of localization and signaling by the focal adhesion targeting domain. *Structure* **2002**, *10* (3), 319-327.

(16) Liu, G.; Guibao, C. D.; Zheng, J. Structural insight into the mechanisms of targeting and signaling of focal adhesion kinase. *Mol. Cell. Biol.* **2002**, *22* (8), 2751-2760.

(17) Hoellerer, M. K.; Noble, M. E.; Labesse, G.; Campbell, I. D.; Werner, J. M.; Arold, S. T. Molecular recognition of paxillin LD motifs by the focal adhesion targeting domain. *Structure* **2003**, *11* (10), 1207-1217.

(18) Gao, G.; Prutzman, K. C.; King, M. L.; Scheswohl, D. M.; DeRose, E. F.; London, R. E.; Schaller, M. D.; Campbell, S. L. NMR solution structure of the focal adhesion targeting domain of focal adhesion kinase in complex with a paxillin LD peptide: evidence for a two-site binding model. *J. Biol. Chem.* **2004**, *279* (9), 8441-8451.

(19) Cooley, M. A.; Broome, J. M.; Ohngemach, C.; Romer, L. H.; Schaller, M. D. Paxillin binding is not the sole determinant of focal adhesion localization or dominant-negative activity of focal adhesion kinase/focal adhesion kinase-related nonkinase. *Mol. Biol. Cell* **2000**, *11* (9), 3247-3263.

(20) Prutzman, K. C.; Gao, G.; King, M. L.; Iyer, V. V.; Mueller, G. A.; Schaller, M. D.; Campbell, S. L. The focal adhesion targeting domain of focal adhesion kinase contains a hinge region that modulates tyrosine 926 phosphorylation. *Structure* **2004**, *12* (5), 881-891.

(21) Dixon, R. D.; Chen, Y.; Ding, F.; Khare, S. D.; Prutzman, K. C.; Schaller, M. D.; Campbell, S. L.; Dokholyan, N. V. New insights into FAK signaling and localization based on detection of a FAT domain folding intermediate. *Structure* **2004**, *12* (12), 2161-2171.

(22) Bergdoll, M.; Remy, M. H.; Cagnon, C.; Masson, J. M.; Dumas, P. Proline-dependent oligomerization with arm exchange. *Structure* **1997**, *5* (3), 391-401.

(23) Bourne, Y.; Arvai, A. S.; Bernstein, S. L.; Watson, M. H.; Reed, S. I.; Endicott, J. E.; Noble, M. E.; Johnson, L. N.; Tainer, J. A. Crystal structure of the cell cycle-regulatory protein suc1 reveals a beta-hinge conformational switch. *Proc. Natl. Acad. Sci. U. S. A.* **1995**, *92* (22), 10232-10236.

(24) Rousseau, F.; Schymkowitz, J. W.; Wilkinson, H. R.; Itzhaki, L. S. Three-dimensional domain swapping in p13suc1 occurs in the unfolded state and is controlled by conserved proline residues. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98* (10), 5596-5601.

(25) Toutant, M.; Costa, A.; Studler, J. M.; Kadare, G.; Carnaud, M.; Girault, J. A. Alternative splicing controls the mechanisms of FAK autophosphorylation. *Mol. Cell. Biol.* **2002**, *22* (22), 7731-7743.

(26) Kadare, G.; Toutant, M.; Formstecher, E.; Corvol, J. C.; Carnaud, M.; Boutterin, M. C.; Girault, J. A. PIAS1-mediated sumoylation of focal adhesion kinase activates its autophosphorylation. *J. Biol. Chem.* **2003**, *278* (48), 47434-47440.

(27) Toutant, M.; Studler, J. M.; Burgaya, F.; Costa, A.; Ezan, P.; Gelman, M.; Girault, J. A. Molecular characterization of FAK neuronal isoforms. *Biochem. J.* **2000**, *348*, 119-128.

(28) Burgaya, F.; Toutant, M.; Studler, J. M.; Costa, A.; Le Bert, M.; Gelman, M.; Girault, J. A. Alternatively spliced focal adhesion kinase in rat brain with increased autophosphorylation activity. *J. Biol. Chem.* **1997**, *272*, 28720-28725.

(29) Ishibe, S.; Joly, D.; Liu, Z. X.; Cantley, L. G. Paxillin serves as an ERK-regulated scaffold for coordinating FAK and Rac activation in epithelial morphogenesis. *Mol. Cell* **2004**, *16* (2), 257-267.

(30) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **2005**, *26* (16), 1781-1802.

(31) Elmore, D. E.; Dougherty, D. A. Molecular dynamics simulations of wild-type and mutant forms of the Mycobacterium tuberculosis MscL channel. *Biophys. J.* **2001**, *81* (3), 1345-1359.

(32) Voordijk, S.; Hansson, T.; Hilvert, D.; van Gunsteren, W. F. Molecular dynamics simulations highlight mobile regions in proteins: A novel suggestion for converting a murine V(H) domain into a more tractable species. *J. Mol. Biol.* **2000**, *300* (4), 963-973.

(33) Schlaepfer, D. D.; Hunter, T. Evidence for in vivo phosphorylation of the Grb2 SH2-domain binding site on focal adhesion kinase by Src-family protein- tyrosine kinases. *Mol. Cell. Biol.* **1996**, *16*, 5623-5633.

(34) Shen, Y.; Schaller, M. D. Focal adhesion targeting: the critical determinant of FAK regulation and substrate phosphorylation. *Mol. Biol. Cell* **1999**, *10* (8), 2507-2518.

(35) Chen, H.-C.; Appeddu, P. A.; Parsons, J. T.; Hildebrand, J. D.; Schaller, M. D.; Guan, J.-L. Interaction of focal adhesion kinase with cytoskeletal protein talin. *J. Biol. Chem.* **1995**, *270*, 16995-16999.

(36) Liu, Z. X.; Yu, C. F.; Nickel, C.; Thomas, S.; Cantley, L. G. Hepatocyte growth factor induces ERK-dependent paxillin phosphorylation and regulates paxillin-focal adhesion kinase association. *J. Biol. Chem.* **2002**, *277* (12), 10452-10458.

(37) Humphrey, W.; Dalke, A.; Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graph.* **1996**, *14* (1), 33-38.

(38) Sanner, M. F.; Olson, A. J.; Spehner, J. C. Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers* **1996**, *38* (3), 305-320.

II-5. Conception *in silico* de peptidomimétiques de la Paxilline ciblant le domaine FAT

Ce projet avait un double but : il s'agissait d'une part de mettre au point des composés peptidomimétiques ciblant spécifiquement le domaine FAT et, d'autre part de traiter un problème de docking pour lequel VSM-G (cf. chapitre 3.I), dans sa version actuelle, n'était pas adapté. En effet, rappelons que le filtre géométrique employé actuellement dans VSM-G est peu efficace avec des sites actifs plats ou très ouverts. Ainsi, en dehors de l'intérêt pharmacologique du système d'étude, ce projet constitua également pour nous une occasion d'acquérir une expertise qui, par la suite, pourra s'avérer précieuse au niveau des développements méthodologiques de VSM-G.

Ce travail a été effectué dans le cadre d'une collaboration avec Mercedes Martin-Martinez et son équipe, de l'Institut de Chimie Médicinale de Madrid (Instituto de Química Médica – IQM), spécialisée dans la synthèse de composés peptidomimétiques à visée thérapeutique. Ainsi ce projet, qui combine la modélisation moléculaire et la chimie médicinale, a pour objectif de développer des composés pouvant bloquer l'invasion tumorale et le développement métastatique. Il s'agit d'une approche multidisciplinaire et rationnelle qui, nous l'espérons, pourra apporter une contribution notable à la conception de futures molécules anti-cancéreuses efficaces.

II-5.1. Objectifs

Comme nous l'avons vu dans la partie décrivant la protéine FAK, le domaine FAT joue un rôle important de régulation. En particulier, FAT intervient dans la surexpression de FAK caractérisant des cancers avancés. Plus précisément, l'état de phosphorylation du résidu Tyr_{925}, situé sur l'hélice H1 de FAT, régule la formation et le renouvellement du complexe d'adhésion focale. Une fois phosphorylé, ce résidu permet de recruter Grb2, ce qui constitue alors le point de départ d'activation de la voie de signalisation Ras-MAPK par FAK, cascade liée à l'angiogénèse pathologique [275, 289]. Il semble donc intéressant, d'un point de vue thérapeutique, de pouvoir limiter cette action, en bloquant l'accessibilité de Tyr_{925} pour que sa phosphorylation par la protéine Src ne puisse pas s'effectuer.

La stratégie à visée thérapeutique retenue consiste à concevoir de petites molécules qui pourraient se lier efficacement au domaine FAT, tout en participant au blocage de Tyr_{925} (p. ex. via une gène stérique ou bien en stabilisant une conformation de FAT dans laquelle ce résidu clé est difficilement

accessible). La surface accessible du domaine FAT ne possède pas de site d'interaction de type "clé-serrure", comme c'est le cas dans la plupart des complexes protéine-ligand, mais, au contraire, présente deux sites de liaison relativement plats et typiques des interactions protéine-protéine. La conception d'antagonistes de haute affinité pour FAT, bloquant ou limitant fortement l'accessibilité de Tyr_{925}, s'apparente à la recherche de composés non peptidiques empêchant l'association de protéines entre elles. Les diverses stratégies possibles pour concevoir de telles molécules ont fait récemment l'objet de revues [340, 341].

II-5.2. Préparation de la simulation du système FAT-LD4

Le point de départ permettant d'accomplir l'objectif annoncé précédemment est la connaissance de la structure de la séquence peptidique interagissant naturellement avec FAT. Il est alors envisageable de concevoir des antagonistes s'inspirant du/des mode(s) de liaison possible(s) de ce peptide avec le domaine FAT. Dans notre cas, nous disposions, pour cette étude de la structure expérimentale du domaine FAT complexée avec deux peptides LD2 et LD4 dérivés de domaines de la Paxilline [310]. Tyr_{925} étant portée par l'hélice H1, le site d'ancrage des peptidomimétiques à concevoir doit se trouver sur la même face du fagot et donc entre les hélices H1 et H4. D'après les études structurales, deux choix étaient possibles : soit le peptide LD2, soit le peptide LD4 [310]. Nous avons choisi arbitrairement LD4 comme système modèle pour la conception de peptidomimétiques se liant à FAT sur l'interface H1-H4 afin de perturber l'accessibilité de la chaîne phénolique de Tyr_{925}.

En partant de la structure cristallographique 1OW7 [310], nous avons préparé, en vue de conduire une simulation par dynamique moléculaire, le modèle du complexe FAT-LD4. L'étude de l'évolution de ce système au cours du temps devrait nous permettre d'identifier et d'analyser un ensemble représentatif de modes de liaison protéine-peptide mis en jeu dans l'interaction.

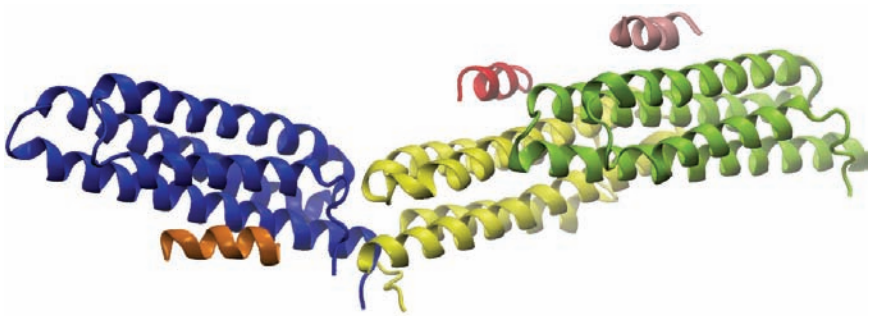


FIG. 37 – Représentation de la structure expérimentale "1OW7"
(FAT en bleu, jaune et vert ; LD4 en orange, rouge et rose)

De la structure 1OW7 de départ, seul un monomère du complexe FAT-LD4 a été retenu, à savoir les chaînes A et D, représentant respectivement les structures du domaine FAT (Asn_{916}-Thr_{1049}) et du peptide LD4 (Thr_1-Ser_{12}). Cet ensemble a ensuite été immergé dans une boîte d'eau (boite cubique de 80 Å de côté) puis neutralisé électriquement (ajout de contre-ions) et enfin protoné à pH = 7. Le système complet (FAT + LD4 + molécules de solvant et contre-ions) comporte au total ~51 100 atomes. Le programme de dynamique moléculaire NAMD [89], couplé au champ de force CHARMM22 [63], a été employé pour conduire la simulation. Le système a tout d'abord été minimisé (6 400 pas de gradients conjugués), puis équilibré pendant 500 ps et enfin une trajectoire de 10 ns a été produite. La simulation a été effectuée à température et pression constantes (300K / 1 atm) et sans contrainte particulière, excepté sur la longueur des liaisons atome lourd-hydrogène, maintenues à leur valeur d'équilibre.

II-5.3. Résultats de la simulation par dynamique moléculaire

Stabilité du système FAT-LD4

Le modèle du complexe FAT-LD4 montre une très grande stabilité le long de la simulation de 10 ns. En effet, les variations de RMSD du squelette peptidique du complexe sont très faibles (cf. FIG. 38-gauche). FAT s'avère plus stable, dans ce modèle où il est complexé avec LD4, que sous sa forme libre (étudiée elle aussi par dynamique moléculaire, cf. section II.3). La liaison FAT-LD4 stabilise en effet le fagot d'hélices de FAT, grâce à une interaction de type hélice-hélice dont l'importance a déjà été soulignée dans divers travaux structuraux [309, 310, 342]. Les éléments de structure secondaire du complexe sont également bien conservés au cours de la simulation. En particulier, l'architecture structurale en hélices-α du domaine FAT et du peptide LD4 est maintenue tout le long de la trajectoire (cf. FIG. 38-droite).

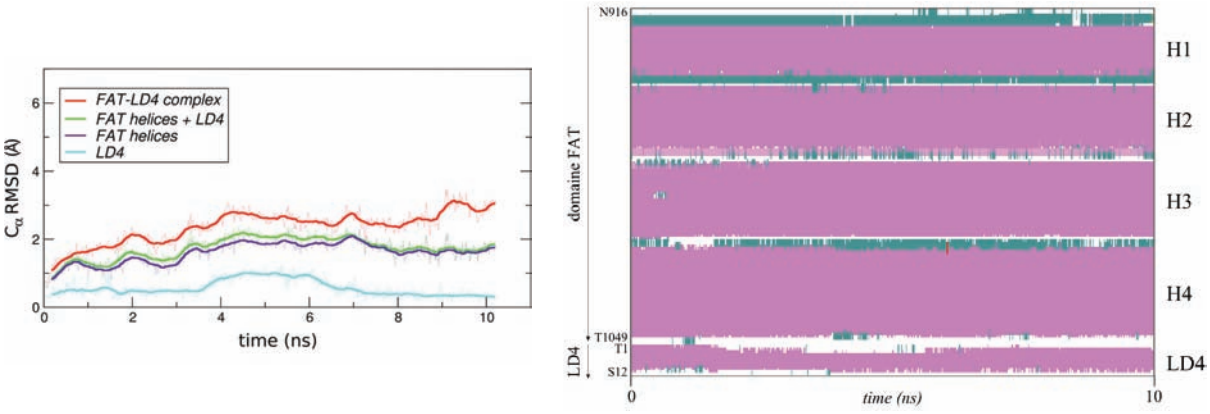


FIG. 38 – Mesures de variations structurales du complexe FAT-LD4. A gauche : évolution du RMSD (calculé sur les carbones-α) de divers sous-ensembles, en fonction du temps. A droite : évolution de la structure secondaire du complexe au cours de la simulation (Codes couleurs : hélice-α → violet ; hélice-3_{10} → rose ; coude-β → vert ; pelote → blanc).

Détails des interactions

Afin de concevoir des molécules en s'inspirant des modes de liaison de LD4 avec FAT, nous avons étudié les détails des interactions entre ces deux partenaires, mises en évidence dans la trajectoire de dynamique moléculaire. Pour cela, nous avons décomposé l'énergie d'interaction globale du système FAT-LD4 simulé[*], afin de déterminer la contribution, dans la liaison, des résidus de chaque partenaire du complexe (cf. FIG. 39).

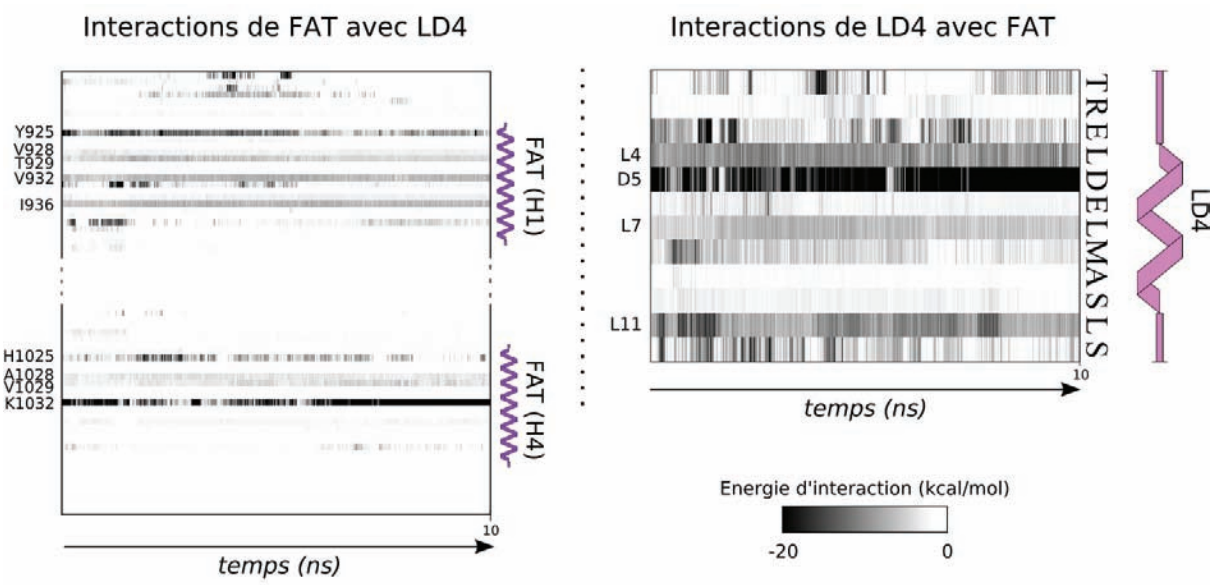


FIG. 39 – Diagrammes d'énergie d'interaction entre le domaine FAT et le peptide LD4 de la Paxilline durant la dynamique moléculaire de 10 ns. A gauche : les résidus de FAT contre l'ensemble de LD4 ; à droite : les résidus du LD4 contre l'ensemble de FAT.

L'analyse réalisée à partir de ces diagrammes a ainsi permis d'identifier aisément les résidus jouant un rôle critique dans la formation et la stabilité du complexe FAT-LD4, c'est-à-dire ceux contribuant de façon importante et constante à l'énergie d'interaction. En ce qui concerne FAT, les résidus clés sont ceux qui constituent l'essentiel du site de liaison entre les hélices H1 et H4 [310], à savoir : Tyr_{925}[†], Val_{928}, Thr_{929}, Val_{932}, Ile_{936}, His_{1025}, Ala_{1028}, Val_{1029} et, plus particulièrement, Lys_{1032}. Du côté du motif LD4 de la Paxilline, les résidus participant le plus à l'interaction FAT-LD4 sont, par ordre décroissant d'importance : $Asp_{5,}$ Leu_4, Leu_{11} et Leu_7. Cet ensemble de résidus leucine forme avec Met_8 la face hydrophobe du motif LD4 [310], orientée vers le site de liaison de FAT entre les hélices H1 et H4.

[*] Cette énergie d'interaction est la contribution enthalpique de l'énergie libre d'association, évaluée par le champ de force. Elle est calculée par la fonction "Pair interaction" de NAMD que nous avons interfacée par des scripts.

[†] L'interaction importante de Tyr_{925} avec le motif LD4 de la Paxilline illustre l'implication de ce résidu dans la liaison FAK-Paxilline. Dans pareille situation, ses degrés de liberté étant très limités, cette tyrosine n'est pas accessible pour une éventuelle phosphorylation.

Les interactions clés de la liaison entre FAT et LD4, suggérées par les diagrammes d'interaction, peuvent être identifiées précisément par une inspection visuelle de l'assemblage moléculaire. La liaison FAT-LD4 repose principalement sur un pont salin entre FAT/Lys_{1032} et LD4/Asp_5. Des contacts multiples entre le *patch hydrophobe* de FAT et les résidus hydrophobes de LD4 mentionnés précédemment complètent la liaison (cf. FIG. 40).

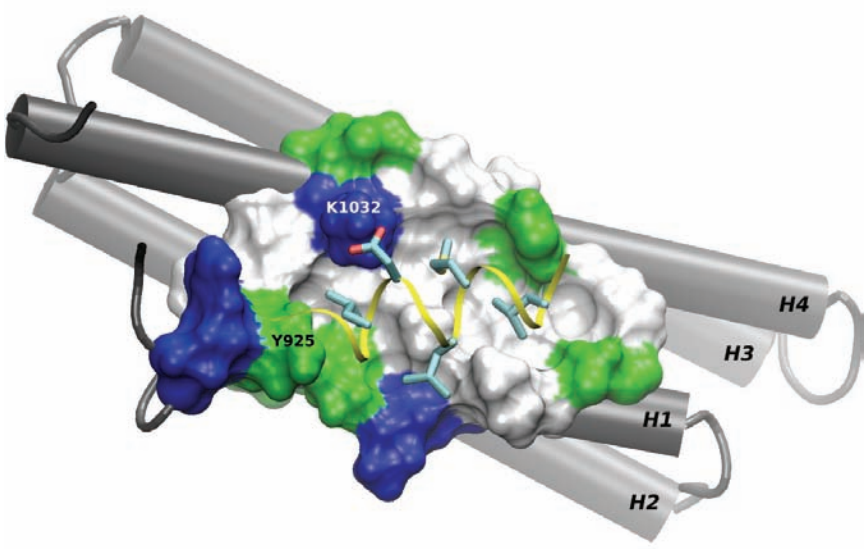


FIG. 40 – Représentation des résidus clés du peptide LD4 en interaction avec le domaine FAT. LD4 est représenté en ruban jaune, ses chaînes latérales Leu_4, Asp_5, Leu_7, Met_8 et Leu_{11} (de gauche à droite) sont en bâtons et colorés par type atomique. La surface du site de liaison, entre les hélices H1 et H4 de FAT, est colorée par propriétés (résidus apolaires en blanc ; polaires en vert ; chargés positivement en bleu).

II-5.4. Conception des molécules mimétiques

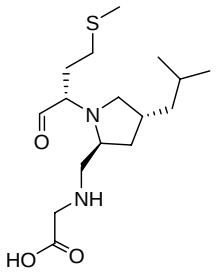
A la suite de la caractérisation des résidus clés de LD4 pour sa liaison avec FAT, nous avons imaginé, avec l'aide de nos collègues chimistes médicinaux, des molécules pouvant reproduire les caractéristiques principales du mode de liaison observé. Divers squelettes moléculaires (châssis) nous ont tout d'abord été proposés. Chaque châssis dispose d'un certain nombre de points d'ancrage sur lesquels les chimistes peuvent rattacher des fragments. Pour des raisons pratiques, nous nous sommes limités aux châssis à trois fragments. Un des fragments est défini spécifiquement pour reproduire l'interaction du résidu Asp_5 de LD4. Les deux autres fragments sont choisis arbitrairement à partir d'une liste de pseudo chaînes latérales hydrophobes suggérées par nos collègues chimistes.

En préalable aux premières synthèses, puis en parallèle aux suivantes, nous avons procédé à une série de simulations par dynamique moléculaire des complexes de FAT avec les molécules proposées. L'objectif ici était de donner aux chimistes des pistes pour guider la synthèse de composés ultérieurs. Ainsi, nous avons cherché aussi bien à évaluer la stabilité d'interaction des composés au cours du temps qu'à étudier les détails atomiques des interactions peptidomimétiques-FAT. Les résultats issus de ces simulations ont permis ensuite de proposer aux chimistes de synthétiser en priorité les molécules prédites comme étant les plus affines avec le domaine FAT. Suivant les difficultés de synthèse rencontrées, il est également arrivé que les chimistes nous proposent de tester *in silico* des composés dont la mise au point leur semblait plus aisée. Ce projet était donc basé sur un *feedback* continu entre les observations tirées des trajectoires de dynamique moléculaire et l'avancement des voies de synthèse engagées (et leurs difficultés rencontrées).

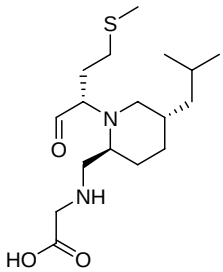
Les simulations des complexes peptidomimétiques-FAT ont été effectuées en suivant le même protocole de dynamique moléculaire que pour FAT-LD4 ; certains paramètres du champ de force qui manquaient pour décrire les petites molécules ont été ajoutés [343]. Comme point de départ de chaque trajectoire, le fragment analogue à LD4/Asp_5 des peptidomimétiques a été positionné sur le site de liaison de FAT en face de Lys_{1032}. La durée des trajectoires produites pour chaque composé était de 2 ns, faisant suite à une phase d'équilibration de 500 ps.

Détermination du châssis "idéal"

La première étape de la conception consistait à identifier un châssis moléculaire sur lequel les pseudo chaînes latérales aspartate et hydrophobes pouvaient être adéquatement disposées. Nous avons testé en premier lieu des peptidomimétiques comportant deux types de châssis différents et que les chimistes savaient synthétiser ; à savoir un "cycle à 5" (pyrrolidine) et un "cycle à 6" (pipéridine).



Mimétique #1



Mimétique #2

FIG. 41 – Premières molécules testées in silico par dynamique moléculaire.

Outre l'analyse visuelle des trajectoires (cf. FIG. 42), les cartes d'interaction de ces deux composés avec le domaine FAT ont également été générées (cf. FIG. 43) afin de surveiller l'évolution des interactions FAT/peptidomimétique au cours des simulations.

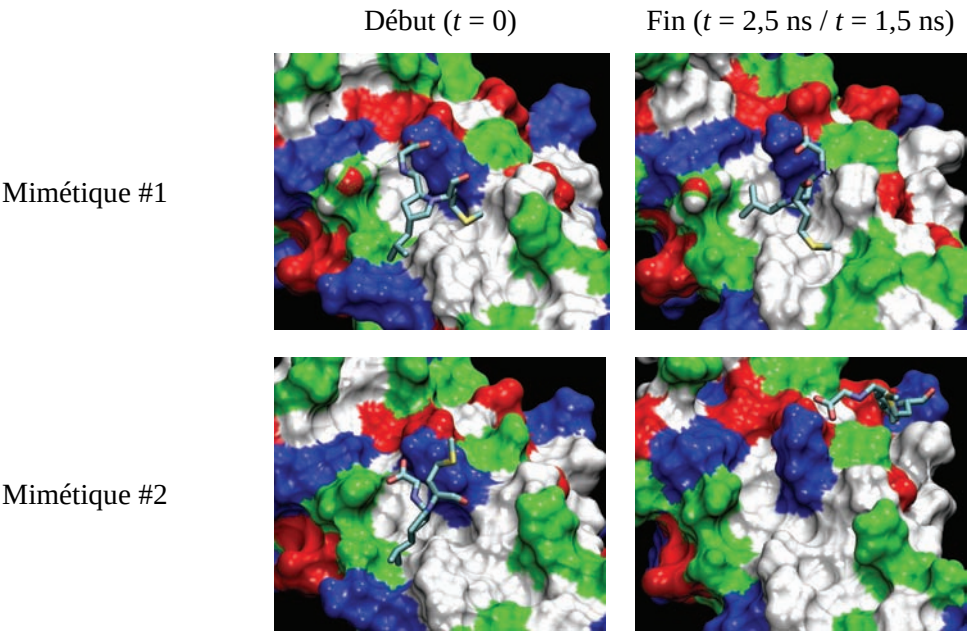


FIG. 42 – Poses des mimétiques #1 (haut) et #2 (bas) au sein du site de liaison de FAT. Les images ont été générées au début (t=0) et à la fin de leurs trajectoires (t=2,5 ns pour le composé #1 / t=1,5 ns pour le #2).

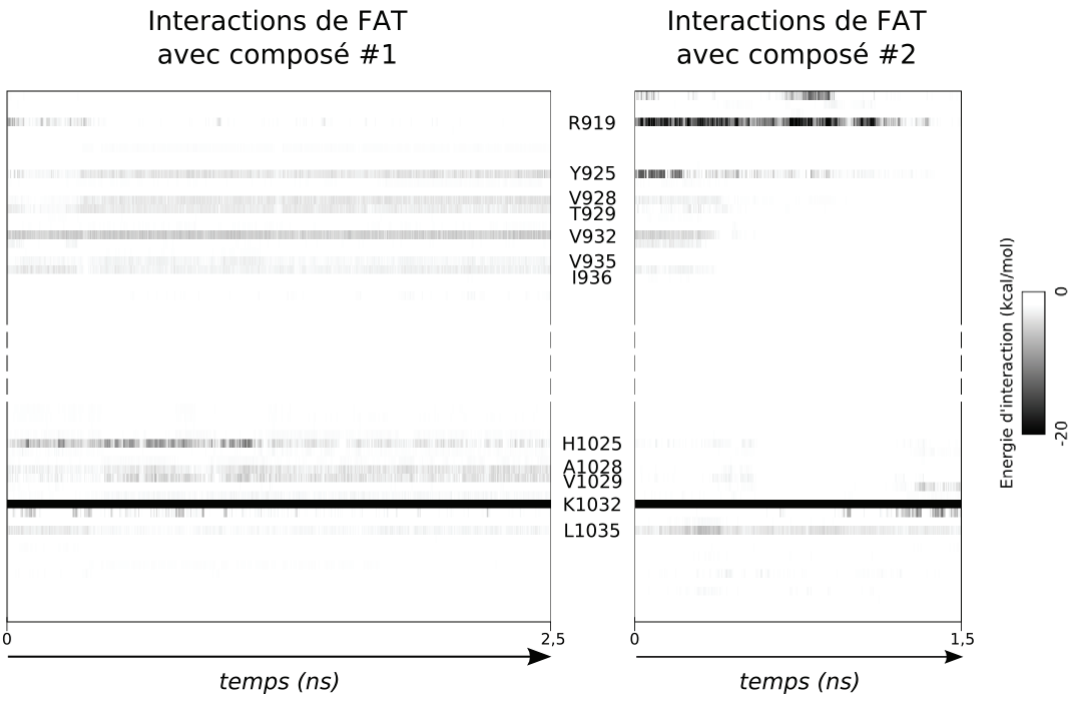


FIG. 43 – Diagrammes d'énergie d'interaction entre les résidus de FAT et les composés #1 et #2.

À l'aide de ces moyens d'analyse, nous avons constaté que les deux composés de référence présentaient des comportements très différents vis-à-vis de leur liaison avec le domaine FAT. En effet, le composé #1 ("cycle à 5") montre une très grande stabilité d'interaction avec FAT, à la différence du #2 ("cycle à 6") qui quitte rapidement le site de liaison et pour lequel la simulation a donc été stoppée à 1,5 ns.

Dès la phase d'équilibration, la molécule #1 se réoriente au sein du site de liaison, permettant alors à ses trois pseudo chaînes latérales de se repositionner de façon optimale. Ce repositionnement se traduit par un changement de mode de liaison du composé #1 avec certains résidus de FAT (cf. FIG. 43-gauche). À la suite de cet événement, les interactions sont stables jusqu'à la fin de la simulation et reproduisent le mode de liaison mis en évidence dans la structure expérimentale du complexe FAT-LD4 (cf. FIG. 39). Elles se composent d'une part du pont salin formé entre FAT/Lys_{1032} et la chaîne pseudo aspartate (~-105 $kcal.mol^{-1}$) et, d'autre part de nombreuses interactions plus faibles dont la majorité sont de nature hydrophobe. Des analyses complémentaires ont montré que les effets de solvant ne perturbaient guère ces interactions clés dans la liaison entre FAT et la molécule #1.

Le mimétique #2, quant à lui, ne se maintient pas dans une conformation stable et de basse énergie au sein du site de liaison qu'il quitte assez rapidement pour rejoindre le solvant (cf. FIG. 42). Le diagramme d'énergie d'interaction de la figure 43 met ainsi en évidence la perte d'un nombre important d'interactions au bout d'environ 500 ps. De plus, en fin de trajectoire (1,5 ns), seul le pont salin avec Lys_{1032} semble caractériser l'interaction de la molécule #2 avec FAT. L'inspection visuelle de la trajectoire est en accord avec ces observations.

D'autres simulations sur des composés avec des châssis à "cycle à 5" ou "cycle à 6" ont confirmé les observations décrites ci-dessus. La taille du cycle et donc sa flexibilité semblent être un critère déterminant dans la stabilité d'interaction des composés avec le domaine FAT. En particulier, la disposition des pseudo chaînes latérales sur les châssis "cycle à 5" correspond à une contrainte conformationnelle favorable au niveau du placement de la molécule dans le site de liaison. Par ailleurs, la nature et la topologie de ce site de liaison ont été mises en évidence (cf. FIG. 44), ce qui nous a indiqué des caractéristiques générales à rechercher concernant les fragments des mimétiques. Ainsi, nous ciblons la présence d'un groupement chargé négativement pour pouvoir former un pont salin avec Lys_{1032} et également deux autres fragments apolaires pouvant occuper au mieux les deux concavités A et B, maximisant ainsi l'effet de désolvatation du site de liaison qui est une caractéristique recherchée pour la mise au point de peptidomimétiques compétitifs.

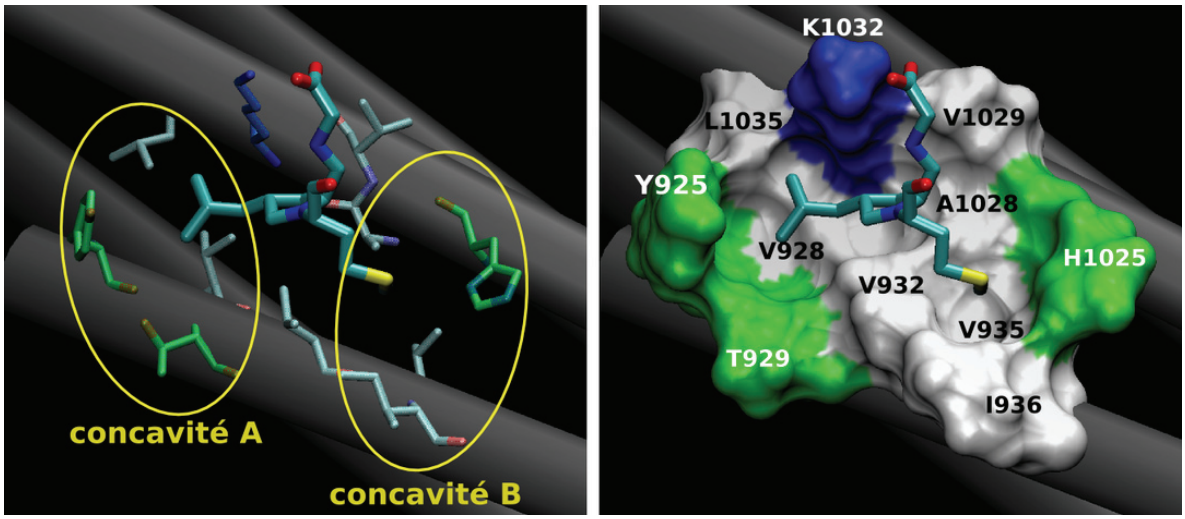


FIG. 44 – Topologie du site de liaison entre les hélices H1 et H4 de FAT, ici avec le composé #1.

Variations sur les pseudo chaînes latérales

Nous avons retenu le squelette "cycle à 5" (cf. FIG. 45) et sommes partis du principe général que, sur les trois fragments, un doit être chargé négativement afin de former le pont salin caractéristique de la liaison FAT/LD4, tandis que les deux autres doivent avoir un caractère hydrophobe. Nous avons ainsi modélisé une série de mimétiques potentiels de LD4. Ces molécules ont été testées *in silico* et, pour celles prédites comme étant les plus affines sur FAT, la synthèse organique a été engagée.

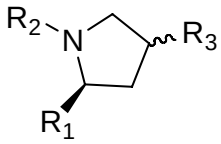


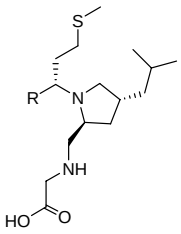
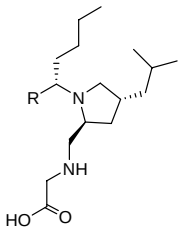
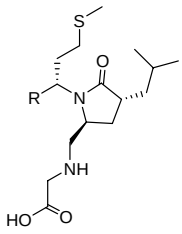
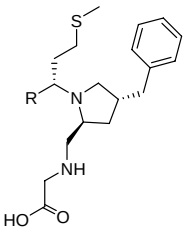
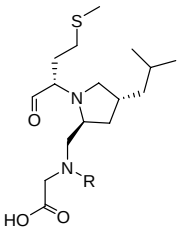
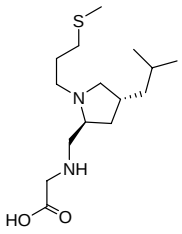
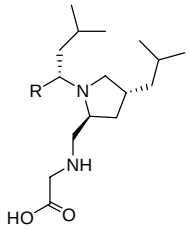
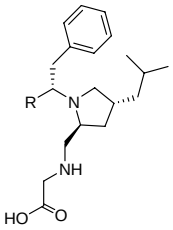
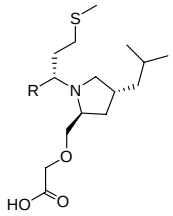
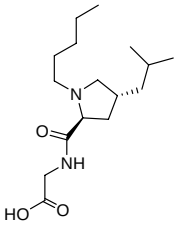
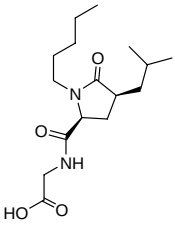
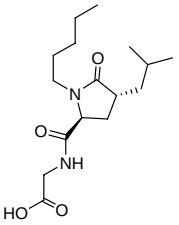
FIG. 45 – Architecture générique des molécules considérées.

L'analyse des simulations effectuées sur ces composés nous a fourni des indications claires sur le rôle des interactions de leurs pseudo chaînes latérales avec le site de liaison[*]. On a ainsi observé que :

- R_1 forme un pont salin avec Lys_{1032} et fait des contacts hydrophobes avec Val_{1029}.
- R_2 occupe la concavité B (définie auparavant, cf. FIG. 44) et fait des contacts hydrophobes essentiellement avec Val_{935}, Val_{932} et Ile_{936}. Des interactions avec His_{1025} (de type π-π, alkyle-π, ou liaisons hydrogène transitoires) peuvent avoir lieu.
- R_3 occupe la concavité A par contacts hydrophobes avec Val_{928}, L_{1035} et Tyr_{925}. Les interactions bloquant les degrés de liberté de Tyr_{925} sont du type π-π ou alkyle-π.

[*] Pour des raisons évidentes, l'observation de la trajectoire et les diagrammes d'énergie de chacune de ces molécules ne seront pas détaillés.

Le tableau 4 ci-dessous présente de manière qualitative les résultats des simulations effectuées sur chacun des composés.

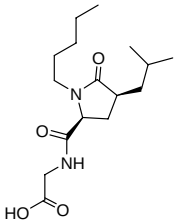
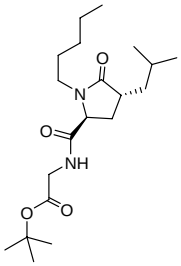
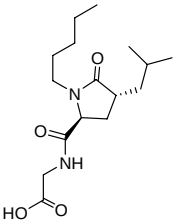
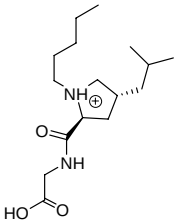
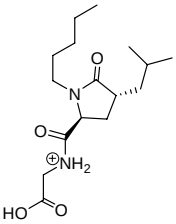
A	B	C	D	E	F
					
R = COH ☺ R = COOMe 😐 R = COO ☺ R = $CONH_2$ ☺	R = COH 😐 R = COOMe 😐 R = COO 😐 R = $CONH_2$ 😐	R = COO ☹ R = $CONH_2$ ☹	R = COH ☺ R = COOMe ☺ R = COO ☺ R = $CONH_2$ ☺	R = isobutyl ☹ R = CH_2Ph ☹	😐
G	**H**	**I**	**J**	**K**	**L**
					
R = H ☺ R = COO ☺ R = $CONH_2$ ☺	R = COH ☹ R = COO 😐 R = $CONH_2$ ☺	R = H ☺ R = COO ☺ R = $CONH_2$ ☺	😐	☺	☺

TAB. 4 – Tableau récapitulant les résultats des simulations de manière qualitative (stabilité de leurs interactions avec le site de liaison et leur ancrage dans celui-ci).

II-5.5. État des lieux

En dépit de leur optimisme initial, nos collègues chimistes ont dû faire face à de nombreux problèmes de synthèse. C'est pour cette raison qu'à ce jour, seuls cinq composés ont été synthétisés avec succès (cf. TAB. 5). Parmi eux, figurent les mimétiques K et L que nous avions déjà simulés par dynamique moléculaire (cf. TAB. 4). Ces deux molécules seront ainsi, à l'issue des tests biologiques*, des étalons qui nous permettront de confronter directement les prédictions théoriques avec l'expérience. Les trois autres composés (MT156, MT210 et MT211), dont les simulations sont prévues, sont des produits intermédiaires de synthèse.

* Des expériences de localisation par immunofluorescence (pour observer si les composés déplacent FAK des adhésions focales) et des mesures de phosphorylation sont actuellement en cours sur ces cinq composés.

	MT152 (K)	MT156	MT171 (L)	MT210	MT211
					
Simulation	☺	?	☺	?	?
Expérience	?	?	?	?	?

TAB. 5 – Résumé des cinq composés ayant été synthétisés avec succès. Le statut actuel quant à leurs résultats théoriques et expérimentaux est indiqué.

II-6. Conclusion et perspectives

FAK est une protéine au fonctionnement complexe. La place centrale qu'elle occupe au sein du complexe d'adhérence focale fait d'elle une plateforme de multiples interactions protéine-protéine impliquées dans la régulation de nombreuses voies de signalisation. Certains aspects de son rôle de régulateur restent non élucidés à ce jour et les résultats contradictoires de certains travaux montrent à quel point certains aspects des mécanismes et des phénomènes régissant son activité au sein de la cellule restent difficiles à saisir.

Comme nous avons pu le voir à travers l'étude de l'influence de la région charnière entre les hélices H1 et H2 de son domaine FAT, du point de vue du modélisateur, l'étude de FAK peut s'avérer encore plus délicate compte tenu des limitations inhérentes à certaines méthodes. Ainsi, les résultats préliminaires de nos simulations dans le cadre de l'étude de la flexibilité de la boucle H1-H2 de FAT ne confirment ni n'invalident clairement notre hypothèse de travail initiale. À ce stade, il nous semble difficile de déterminer à quel(s) niveau(x) pourrait se situer principalement la problématique. On pourrait s'intéresser aussi bien à la pertinence de l'hypothèse et du protocole de simulation employé qu'à la validité du modèle structural employé.

Dans tous les cas de figure, nous envisageons de poursuivre nos investigations afin d'essayer de comprendre le décalage entre les indications fournies par les données expérimentales et les résultats

préliminaires de nos simulations. Ceux-ci suggèrent que l'ouverture de l'hélice H1 par rapport au reste du fagot n'est pas un phénomène spontané, un tel changement conformationnel devant avoir une barrière énergétique relativement élevée. C'est pourquoi nous envisageons en premier lieu de travailler dans des conditions de simulation différentes (par exemple sous contraintes), et si possible avec un modèle plus complet (la protéine FAK en entier, ou bien une portion de celle-ci complexée avec un partenaire tel que Src ou Grb2).

La seconde application sur FAK vise à mettre au point, de façon rationnelle, des petites molécules anti-métastase ciblant le domaine FAT. On utilise pour cela la dynamique moléculaire sur la base d'une structure expérimentale (complexe FAT-LD4). Cette stratégie a d'abord servi à identifier les principales caractéristiques de l'interaction FAT-LD4, ce qui a guidé la conception initiale, avec l'aide de nos collègues chimistes médicinaux, d'une série de peptidomimétiques mimant le motif LD4 de la Paxilline. Les simulations des complexes correspondants permettent de sélectionner les meilleures molécules-candidates tout en affinant nos connaissances sur les modes de liaison du récepteur ciblé sur FAT et ainsi d'optimiser les mimétiques.

Les synthèses des molécules proposées à nos collègues chimistes se sont malheureusement avérées plus délicates qu'ils ne l'avaient prévu. Toutefois, cinq composés ayant été synthétisés avec succès sont en cours de tests biologiques. Ici, la dynamique moléculaire s'est relevée être précieuse pour discerner les différences d'affinité entre certains composés comportant des différences subtiles de groupements fonctionnels.

Bien qu'elle soit relativement coûteuse en temps de calculs, notre approche a clairement permis aux chimistes d'économiser un temps précieux dans notre démarche commune d'optimisation. Comparée à une stratégie purement expérimentale, la stratégie collaborative reposant sur une interaction entre prédictions théoriques fines et validation expérimentale est la plus efficace.

CONCLUSION GÉNÉRALE

Un des défis des stratégies modernes de recherche de médicaments est celui du criblage de chimiothèques de grande taille sur des cibles biologiques complexes. Pour le relever, le criblage virtuel est une approche prometteuse dont le développement méthodologique est en constante expansion.

La plateforme logicielle VSM-G ambitionne de regrouper au sein d'une interface cohérente et simple d'emploi la conception de protocoles de criblage virtuel efficaces malgré l'hétérogénéité des techniques associées. À l'heure actuelle, VSM-G est un prototype fonctionnel qui a déjà fait ses preuves dans plusieurs campagnes de criblage. Les résultats obtenus indiquent que les concepts mis en œuvre s'avèrent pertinents (cf. article #2) ; ce qui est particulièrement encourageant dans l'optique d'étendre et de consolider VSM-G.

VSM-G, sous sa forme présente (cf. Partie 3.I), procède selon un cheminement décrémentiel : il permet de réduire une base de composés pour l'enrichir vis-à-vis de la cible. Des méthodes innovantes sont mises en œuvre, telles que l'utilisation d'harmoniques sphériques (cf. article #1). Les protocoles employés dans la Partie 3.II se classent eux aussi dans la catégorie du *structure-based drug design*. Ils diffèrent toutefois par leur orientation incrémentielle : on génère à partir d'un petit nombre de molécules de référence des petites séries d'analogues dans un but d'optimisation (cf. Partie 3.II-5). On n'utilise cette fois-ci que la technique de dynamique moléculaire, méthode de simulation particulièrement éprouvée (cf. Partie 2.I).

VSM-G permet l'exploration rapide de chimiothèques présentant une importante diversité chimique ; il s'agit d'un criblage virtuel adapté à la découverte de touches. À l'opposé, l'approche par dynamique moléculaire constitue un criblage focalisé sur une région locale en terme de diversité et se place dans le contexte d'optimisation de touches vers des têtes de série. Ainsi, les deux stratégies employées dans ce travail de thèse s'avèrent complémentaires, en particulier dans le cadre des étapes préliminaires de la recherche pharmaceutique (cf. Partie 1.II). Concernant l'extension de VSM-G, priorité est donnée à l'intégration de la dynamique moléculaire à la fin du processus de criblage multi-étapes. La validation de cette stratégie unifiée sera effectuée sur un ensemble de systèmes biologiques plus représentatifs que les seuls LXRβ et FAK/FAT.

Nous allons à présent dresser le bilan des projets applicatifs de ce travail. En ce qui concerne la cible LXRβ, les résultats issus du criblage haut-débit entrepris avec VSM-G sont prometteurs mais malheureusement confidentiels. Si nous avons pu en présenter la teneur, nous ne disposons pas du niveau de détail suffisant qui aurait pu nous permettre d'ajuster les paramètres du protocole employé. Il aurait été particulièrement intéressant, de notre point de vue, d'adjoindre au docking rigide géométrique au moins une étape ultérieure (p. ex. de docking flexible, cf. article #2). Nous aurions alors pu valider l'approche hiérarchique du criblage implémentée dans VSM-G de façon plus rigoureuse, c'est-à-dire sur un ensemble significatif de données expérimentales. La pertinence du choix de multiples conformations de la cible pour le criblage – afin de tenir compte des effets d'*induced fit* (cf. article #3) – aurait aussi pu être mieux évaluée.

Le domaine FAT de FAK est, quant à lui, une cible de grand intérêt mais bien plus difficile à étudier par sa complexité (cf. article #4). Nous sommes toutefois parvenus à concevoir *in silico* un ensemble de composés prometteurs parmi lesquels cinq ont été synthétisés et sont en cours de test *in vivo*. Du point de vue méthodologique, nous avons par cette étude acquis une expertise qui s'avérera précieuse pour certains des développements futurs programmés en ce qui concerne VSM-G.

RÉFÉRENCES BIBLIOGRAPHIQUES

(1) Watson, J. D.; Crick, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **1953**, *171* (4356), 737-8.

(2) Watson, J. D.; Crick, F. H. Genetical implications of the structure of deoxyribonucleic acid. *Nature* **1953**, *171* (4361), 964-7.

(3) Crick, F. H.; Barnett, L.; Brenner, S.; Watts-Tobin, R. J. General nature of the genetic code for proteins. *Nature* **1961**, *192*, 1227-32.

(4) Yanofsky, C. Gene structure and protein structure. *Sci. Am.* **1967**, *216* (5), 80-94.

(5) Mitra, K.; Frank, J. Ribosome dynamics: insights from atomic structure modeling into cryo-electron microscopy maps. *Annu. Rev. Biophys. Biomol. Struct.* **2006**, *35*, 299-317.

(6) Horwich, A. L.; Fenton, W. A.; Chapman, E.; Farr, G. W. Two families of chaperonin: physiology and mechanism. *Annu. Rev. Cell Dev. Biol.* **2007**, *23*, 115-145.

(7) Anfinsen, C. B. Principles that govern the folding of protein chains. *Science* **1973**, *181* (96), 223-30.

(8) Dobson, C. M. Protein folding and misfolding. *Nature* **2003**, *426* (6968), 884-890.

(9) Onuchic, J. N.; Wolynes, P. G. Theory of protein folding. *Curr. Opin. Struct. Biol.* **2004**, *14* (1), 70-75.

(10) Dill, K. A.; Ozkan, S. B.; Weikl, T. R.; Chodera, J. D.; Voelz, V. A. The protein folding problem: when will it be solved? *Curr. Opin. Struct. Biol.* **2007**, *17* (3), 342-346.

(11) Moult, J. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr. Opin. Struct. Biol.* **2005**, *15* (3), 285-289.

(12) James, L. C.; Tawfik, D. S. Conformational diversity and protein evolution--a 60-year-old hypothesis revisited. *Trends Biochem. Sci.* **2003**, *28* (7), 361-8.

(13) Liu, H. L.; Hsu, J. P. Recent developments in structural proteomics for protein structure determination. *Proteomics* **2005**, *5* (8), 2056-2068.

(14) Uetz, P.; Finley, R. L., Jr. From protein networks to biological systems. *FEBS Lett.* **2005**, *579* (8), 1821-7.

(15) Parrish, J. R.; Gulyas, K. D.; Finley, R. L., Jr. Yeast two-hybrid contributions to interactome mapping. *Curr. Opin. Biotechnol.* **2006**, *17* (4), 387-93.

(16) Hanahan, D.; Weinberg, R. A. The hallmarks of cancer. *Cell* **2000**, *100* (1), 57-70.

(17) Devos, D.; Russell, R. B. A more complete, complexed and structured interactome. *Curr. Opin. Struct. Biol.* **2007**, *17* (3), 370-7.

(18) Gavin, A. C.; Bosche, M.; Krause, R., et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **2002**, *415* (6868), 141-7.

(19) Ho, Y.; Gruhler, A.; Heilbut, A., et al. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* **2002**, *415* (6868), 180-3.

(20) Krogan, N. J.; Cagney, G.; Yu, H., et al. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature* **2006**, *440* (7084), 637-43.

(21) Stewart, L.; Clark, R.; Behnke, C. High-throughput crystallization and structure determination in drug discovery. *Drug Discov. Today* **2002**, *7* (3), 187-196.

(22) Berman, H. M.; Westbrook, J.; Feng, Z., et al. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235-42.

(23) Stelzl, U.; Worm, U.; Lalowski, M., et al. A human protein-protein interaction network: a resource for annotating the proteome. *Cell* **2005**, *122* (6), 957-968.

(24) Gandhi, T. K.; Zhong, J.; Mathivanan, S., et al. Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nat. Genet.* **2006**, *38* (3), 285-293.

(25) Lander, E. S.; Linton, L. M.; Birren, B., et al. Initial sequencing and analysis of the human genome. *Nature* **2001**, *409* (6822), 860-921.

(26) Venter, J. C.; Adams, M. D.; Myers, E. W., et al. The sequence of the human genome. *Science* **2001**, *291* (5507), 1304-51.

(27) IHGSC. Finishing the euchromatic sequence of the human genome. *Nature* **2004**, *431* (7011), 931-45.

(28) Levy, S.; Sutton, G.; Ng, P. C., et al. The Diploid Genome Sequence of an Individual Human. *PLoS Biol.* **2007**, *5* (10), e254.

(29) Hopkins, A. L.; Groom, C. R. The druggable genome. *Nat. Rev. Drug Discovery* **2002**, *1* (9), 727-30.

(30) Russ, A. P.; Lampel, S. The druggable genome: an update. *Drug Discov. Today* **2005**, *10* (23-24), 1607-10.

(31) Paolini, G. V.; Shapland, R. H.; van Hoorn, W. P.; Mason, J. S.; Hopkins, A. L. Global mapping of pharmacological space. *Nat. Biotechnol.* **2006**, *24* (7), 805-15.

(32) Overington, J. P.; Al-Lazikani, B.; Hopkins, A. L. How many drug targets are there? *Nat. Rev. Drug Discovery* **2006**, *5* (12), 993-6.

(33) Weinshilboum, R. M.; Wang, L. Pharmacogenetics and pharmacogenomics: development, science, and translation. *Annu. Rev. Genomics Hum. Genet.* **2006**, *7*, 223-45.

(34) Loriot, M. A.; Beaune, P. La pharmacogénétique : le lien entre gènes et réponse aux médicaments. *M. S. Méd. Sci.* **2004**, *20* (6-7), 634-6.

(35) Verma, I. M.; Weitzman, M. D. Gene therapy: twenty-first century medicine. *Annu. Rev. Biochem.* **2005**, *74*, 711-38.

(36) Lindsay, M. A. Target discovery. *Nat. Rev. Drug Discovery* **2003**, *2* (10), 831-8.

(37) Lindsay, M. A. Finding new drug targets in the 21st century. *Drug Discov. Today* **2005**, *10* (23-24), 1683-7.

(38) Butcher, E. C.; Berg, E. L.; Kunkel, E. J. Systems biology in drug discovery. *Nat. Biotechnol.* **2004**, *22* (10), 1253-9.

(39) Anderson, A. C. The process of structure-based drug design. *Chem. Biol.* **2003**, *10* (9), 787-97.

(40) Bleicher, K. H.; Bohm, H. J.; Muller, K.; Alanine, A. I. Hit and lead generation: beyond high-throughput screening. *Nat. Rev. Drug Discovery* **2003**, *2* (5), 369-378.

(41) Walters, W. P.; Stahl, M. T.; Murcko, M. A. Virtual screening - an overview. *Drug Discov. Today* **1998**, *3* (4), 160-178.

(42) Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **2004**, *432* (7019), 862-865.

(43) Bajorath, J. Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discovery* **2002**, *1* (11), 882-894.

(44) Dobson, C. M. Chemical space and biology. *Nature* **2004**, *432* (7019), 824-8.

(45) Lipinski, C.; Hopkins, A. Navigating chemical space for biology and medicine. *Nature* **2004**, *432* (7019), 855-61.

(46) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews* **2001**, *46* (1-3), 3-26.

(47) DiMasi, J. A.; Hansen, R. W.; Grabowski, H. G. The price of innovation: new estimates of drug development costs. *J. Health. Econ.* **2003**, *22* (2), 151-85.

(48) Jorgensen, W. L. The many roles of computation in drug discovery. *Science* **2004**, *303* (5665), 1813-8.

(49) Tang, Y.; Zhu, W.; Chen, K.; Jiang, H. New technologies in computer-aided drug design: Toward target identification and new chemical entity discovery. *Drug Discovery Today* **2006**, *3* (3), 307-313.

(50) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature Reviews Drug Discovery* **2004**, *3* (11), 935-49.

(51) Wlodawer, A.; Vondrasek, J. Inhibitors of HIV-1 protease: a major success of structure-assisted drug design. *Annu. Rev. Biophys. Biomol. Struct.* **1998**, *27*, 249-84.

(52) Grüneberg, S.; Stubbs, M. T.; Klebe, G. Successful virtual screening for novel inhibitors of human carbonic anhydrase: strategy and experimental confirmation. *J. Med. Chem.* **2002**, *45* (17), 3588-602.

(53) Vangrevelinghe, E.; Zimmermann, K.; Schoepfer, J.; Portmann, R.; Fabbro, D.; Furet, P. Discovery of a potent and selective protein kinase CK2 inhibitor by high-throughput docking. *J. Med. Chem.* **2003**, *46* (13), 2656-62.

(54) Alvarez, J. C. High-throughput docking as a source of novel drug leads. *Curr. Opin. Chem. Biol.* **2004**, *8* (4), 365-370.

(55) Ghosh, S.; Nie, A.; An, J.; Huang, Z. Structure-based virtual screening of chemical libraries for drug discovery. *Curr. Opin. Chem. Biol.* **2006**, *10* (3), 194-202.

(56) Hillisch, A.; Pineda, L. F.; Hilgenfeld, R. Utility of homology models in the drug discovery process. *Drug Discovery Today* **2004**, *9* (15), 659-69.

(57) Hénin, J.; Maigret, B.; Tarek, M.; Escrieut, C.; Fourmy, D.; Chipot, C. Probing a model of a GPCR/ligand complex in an explicit membrane environment: the human cholecystokinin-1 receptor. *Biophysical Journal* **2006**, *90* (4), 1232-40.

(58) Karplus, M.; Kuriyan, J. Molecular dynamics and protein function. *Proceedings of the National Academy of Sciences of the United States of America* **2005**, *102* (19), 6679-85.

(59) Freddolino, P. L.; Arkhipov, A. S.; Larson, S. B.; McPherson, A.; Schulten, K. Molecular dynamics simulations of the complete satellite tobacco mosaic virus. *Structure* **2006**, *14* (3), 437-49.

(60) Born, M.; Oppenheimer, R. Zur Quantentheorie der Molekeln. *Annalen der Physik* **1927**, *389* (20), 457-484.

(61) MacKerell Jr., A. D. Empirical force fields for biological macromolecules: Overview and issues. *Journal of Computational Chemistry* **2004**, *25* (13), 1584-1604.

(62) MacKerell Jr., A. D.; Wiórkiewicz-Kuczera, J.; Karplus, M. An all-atom empirical energy function for the simulation of nucleic acids. *Journal of the American Chemical Society* **1995**, *117*, 11946-11975.

(63) MacKerell Jr., A. D.; Bashford, D.; Bellott, M., et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *Journal of Physical Chemistry* **1998**, *102*, 3586-3616.

(64) MacKerell Jr., A. D.; Benavali, N. K. All-atom empirical force field for nucleic acids: II. Application to molecular dynamics simulations of DNA and RNA in solution. *Journal of Computational Chemistry* **2000**, *21* (2), 105-120.

(65) MacKerell Jr., A. D.; Feig, M.; Brooks III, C. L. Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *Journal of Computational Chemistry* **2004**, *25* (11), 1400-1415.

(66) Kollman, P. A.; Dixon, R.; Cornell, W.; Fox, T.; Chipot, C.; Pohorille, A. The development/application of a "minimalist" organic/biochemical molecular mechanic force field using a combination of ab initio calculations and experimental data. In *Computer simulations of biomolecular systems*; Wilkinson, A.; Weiner, P.; van Gunsteren, W. F., Elsevier: 1997; 3, pp 83-96.

(67) Weiner, S. J.; Kollman, P. A.; Case, D. A., et al. A new force field for molecular mechanical simulation of nucleic acids and proteins. *Journal of the American Chemical Society* **1984**, *106*, 765-784.

(68) Weiner, S. J.; Kollman, P. A.; Nguyen, D. T.; Case, D. A. An all-atom force field for simulations of proteins and nucleic acids. *Journal of Computational Chemistry* **1986**, *7*, 230-252.

(69) Cornell, W. D.; Cieplak, P.; Bayly, C. I., et al. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society* **1995**, *117*, 5179-5197.

(70) Cheatham III, T. E.; Cieplak, P.; Kollman, P. A. A modified version of the Cornell et al. force field with improved sugar pucker phases and helical repeat. *Journal of Biomolecular Structure & Dynamics* **1999**, *16*, 845-862.

(71) Duan, Y.; Wu, C.; Chowdhury, S., et al. A point-charge force field for molecular mechanics simulations of proteins. *Journal of Computational Chemistry* **2003**, *24*, 1999-2012.

(72) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general Amber force field. *Journal of Computational Chemistry* **2004**, *25*, 1157-1174.

(73) Daura, X.; Mark, A.; van Gunsteren, W. Parametrization of aliphatic CHn united atoms of GROMOS96 force field. *Journal of computational chemistry* **1998**, *19* (5), 535-547.

(74) Oostenbrink, C.; Villa, A.; Mark, A.; van Gunsteren, W. A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. *Journal of computational chemistry* **2004**, *25* (13), 1656-1676.

(75) Jorgensen, W.; Tirado-Rives, J. The OPLS potential functions for proteins. Energy minimizations for crystals of cyclic peptides and crambin. *Journal of the American Chemical Society* **1988**, *110* (6), 1657-1666.

(76) Jorgensen, W.; Maxwell, D.; Tirado-Rives, J. Development and testing of the OPLS all-atom force-field on conformational energetics and properties of organic liquids. *Journal of the American Chemical Society* **1996**, *118* (45), 11225-11236.

(77) Lennard-Jones, J. E. Cohesion. *Proceedings of the Physical Society* **1931**, *43* (5), 461-482.

(78) Rappé, A.; Casewit, C.; Colwell, K.; Goddard III, W.; Skiff, W. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *Journal of the American Chemical Society* **1992**, *114* (25), 10024-10035.

(79) Maple, J.; Dinur, U.; Hagler, A. Derivation of force fields for molecular mechanics and dynamics from ab initio energy surfaces. *Proceedings of the National Academy of Sciences of the United States of America* **1988**, *85* (15), 5350-5354.

(80) Allinger, N.; Yuh, Y.; Lii, J. Molecular mechanics. The MM3 force field for hydrocarbons. 1. *Journal of the American Chemical Society* **1989**, *111* (23), 8551-8566.

(81) Neumaier, A. Molecular modeling of proteins and mathematical prediction of protein structure. *SIAM review* **1997**, *39* (3), 407-460.

(82) Momany, F. A.; McGuire, R. F.; Burgess, A. W.; Sheraga, H. A. Energy parameters in polypeptides. VII. Geometric parameters, partial atomic charges, nonbonded interactions, hydrogen bond interactions, and intrinsic torsional potentials for the naturally occuring amino acids. *Journal of Physical Chemistry* **1975**, *79* (22), 2361-2381.

(83) Némethy, G.; Pottle, M. S.; Sheraga, H. A. Energy parameters in polypeptides. 9. Updating of geometrical parameters, nonbonded interactions, and hydrogen bond interactions for the naturally occuring amino acids. *Journal of Physical Chemistry* **1983**, *87*, 1833-1887.

(84) Foloppe, N.; MacKerell Jr., A. D. All-atom empirical force field for nucleic acids: I. Parameter optimization based on small molecule and condensed phase macromolecular target data. *Journal of Computational Chemistry* **2000**, *21* (2), 86-104.

(85) van Gunsteren, W. F.; Berendsen, J. C. Computer simulation of molecular dynamics: Methodology, applications, and perspectives in chemistry. *Angewandte Chemie, International Edition in English* **1990**, *29*, 992-1023.

(86) Ponder, J. W.; Case, D. A. Force fields for protein simulations. *Advances in Protein Chemistry* **2003**, *66*, 27-85.

(87) Nelson, M.; Humphrey, W.; Gursoy, A.; Dalke, A.; Kalé, L.; Skeel, R.; Schulten, K. NAMD - A parallel, object-oriented molecular dynamics program. *International Journal of Supercomputer Applications and High Performance Computing* **1996**, *10*, 251-268.

(88) Kalé, L.; Skeel, R.; Bhandarkar, M., et al. NAMD2: Greater scalability for parallel molecular dynamics. *Journal of Computational Physics* **1999**, *151*, 283-312.

(89) Phillips, J. C.; Braun, R.; Wang, W., et al. Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry* **2005**, *26* (16), 1781-1802.

(90) Scott, W.; Hünenberger, P.; Tironi, I., et al. The GROMOS biomolecular simulation program package. *Journal of Physical Chemistry A* **1999**, *103* (19), 3596-3607.

(91) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry* **1983**, *4* (2), 187-217.

(92) MacKerell Jr., A. D.; Brooks, B.; Brooks III, C. L.; Nilsson, L.; Roux, B.; Won, Y.; Karplus, M. CHARMM: The energy function and its parametrization with an overview of the program. In *The Encyclopedia of computational chemistry*; Schleyer, P. V. R., et al., John Wiley & sons: Chichester, **1998**, pp 271-277.

(93) Weiner, P. K.; Kollman, P. A. AMBER: Assisted Model Building with Energy Refinement. A general program for modeling molecules and their interactions. *Journal of Computational Chemistry* **1981**, *2* (3), 287-303.

(94) Pearlman, D. A.; Case, D. A.; Caldwell, J. W., et al. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Computer Physics Communications* **1995**, *91*, 1-41.

(95) Halgren, T.; Damm, W. Polarizable force fields. *Current Opinion in Structural Biology* **2001**, *11* (2), 236-242.

(96) Caldwell, J.; Kollman, P. Cation-pi interactions: nonadditive effects are critical in their accurate representation. *Journal of the American Chemical Society* **1995**, *117* (14), 4177-4178.

(97) Rick, S.; Stuart, S. Potentials and Algorithms for Incorporating Polarizability in Computer Simulations. In *Reviews in Computational Chemistry, Volume 18*; Kenny B. Lipkowitz, D. B. B., 2003; pp 89-146.

(98) Yu, H.; van Gunsteren, W. Accounting for polarization in molecular simulation. *Computer Physics Communications* **2005**, *172* (2), 69-85.

(99) Gresh, N.; Shi, G.-B. Conformation-dependent intermolecular interaction energies of the triphosphate anion with divalent metal cations. Application to the ATP-binding site of a binuclear bacterial enzyme. A parallel quantum chemical and polarizable molecular mechanics investigation. *Journal of Computational Chemistry* **2004**, *25*, 160-168.

(100) Antony, J.; Piquemal, J.-P.; Gresh, N. Complexes of thiomandelate and captopril mercaptocarboxylate inhibitors to metallo-ß-lactamase by polarizable molecular mechanics. Validation on model binding sites by quantum chemistry. *Journal of Computational Chemistry* **2005**, *26* (11), 1131-1147.

(101) Gresh, N. Development, validation, and applications of anisotropic polarizable molecular mechanics to study ligand and drug-receptor interactions. *Current Pharmaceutical Design* **2006**, *12*, 2121-2158.

(102) Kaminski, G. A.; Stern, H. A.; Berne, B. J.; Friesner, R. A. Development of an accurate and robust polarizable molecular mechanics force field from ab initio quantum chemistry. *Journal of Physical Chemistry A* **2004**, *108* (4), 621-627.

(103) Maple, J.; Cao, Y.; Damm, W.; Halgren, T.; Kaminski, G.; Zhang, L.; Friesner, R. A polarizable force field and continuum solvation methodology for modeling of protein-ligand interactions. *Journal of Chemical Theory and Computation* **2005**, *1* (4), 694-715.

(104) Palmo, K.; Mannfors, B.; Mirkin, N.; Krimm, S. Potential energy functions: from consistent force fields to spectroscopically determined polarizable force fields. *Biopolymers* **2003**, *68* (3), 383-394.

(105) Schnieders, M.; Baker, N.; Ren, P.; Ponder, J. Polarizable atomic multipole solutes in a Poisson-Boltzmann continuum. *The Journal of Chemical Physics* **2007**, *126*, 124114.

(106) Leach, A. R. *Molecular modelling: principles and applications (2nd Edition)*. Prentice Hall: **2001**

(107) Metropolis, N.; Rosenbluth, A. E.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. Equation of state calculations by fast computing machines. *Journal of Chemical Physics* **1953**, *21*, 1087-1092.

(108) Berendsen, H.; Postma, J.; van Gunsteren, W.; DiNola, A.; Haak, J. Molecular dynamics with coupling to an external bath. *Journal of Chemical Physics* **1984**, *81*, 3684.

(109) Andersen, H. Molecular dynamics simulations at constant pressure and/or temperature. *Journal of Chemical Physics* **1980**, *72* (4), 2384-2393.

(110) Verlet, L. Computer "experiments" on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules. *Physical Review* **1967**, *159* (1), 98-103.

(111) Hockney, R. W. The potential calculation and some applications. *Methods in Computational Physics* **1970**, *9*, 136-211.

(112) Beeman, D. Some multistep methods for use in molecular dynamics simulations. *Journal of Computational Physics* **1976**, *20*, 130-139.

(113) Swope, W. C.; Anderson, H. C.; Berens, P. H.; Wilson, K. R. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules:application to small water clusters. *Journal of Chemical Physics* **1982**, *76*, 637-649.

(114) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics* **1977**, *23* (3), 327-341.

(115) Kräutler, V.; van Gunsteren, W. F.; Hünerberger, P. H. A fast SHAKE algorithm to solve distance constraint equations for small molecules in molecular dynamics simulations. *Journal of Computational Chemistry* **2001**, *22* (5), 501-508.

(116) van Gunsteren, W. F.; Berendsen, H. J. C. Algorithms for macromolecular dynamics and constraint dynamics. *Molecular Physics* **1977**, *34*, 1311-1327.

(117) Grubmüller, H.; Heller, H.; Windemuth, A.; Schulten, K. Generalized Verlet algorithm for efficient molecular dynamics simulations with long-range interactions. *Molecular Simulation* **1991**, *6*, 121-142.

(118) Tuckerman, M. E.; Berne, B. J.; Martyna, G. J. Reversible multiple time scale molecular dynamics. *Journal of Chemical Physics* **1992**, *97* (3), 1990-2001.

(119) Humphreys, D. D.; Freisner, R. A.; Berne, B. J. A multiple-time-step molecular dynamics algorithm for macromolecules. *Journal of Physical Chemistry* **1994**, *98*, 6885-6892.

(120) McCammon, J. A.; Gelin, B. R.; Karplus, M. Dynamics of folded proteins. *Nature* **1977**, *267* (5612), 585-90.

(121) Levitt, M.; Sharon, R. Accurate simulation of protein dynamics in solution. *Proceedings of the National Academy of Sciences of the United States of America* **1988**, *85* (20), 7557-61.

(122) Koehl, P. Electrostatics calculations: latest methodological advances. *Current Opinion in Structural Biology* **2006**, *16* (2), 142-151.

(123) Saito, M. Molecular dynamics simulations of proteins in solutions: Artifacts caused by the cutoff approximation. *Journal of Chemical Physics* **1994**, *101* (5), 4055-4061.

(124) Ewald, P. P. Die berechnung optischer und elektrostatischer gitterpotentiale. *Annals of Physics* **1921**, *64*, 253-287.

(125) Essmann, U.; Parera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A smooth particle mesh Ewald method. *Journal of Chemical Physics* **1995**, *103* (19), 8577-8593.

(126) Villa, E.; Balaeff, A.; Schulten, K. Structural dynamics of the lac repressor-DNA complex revealed by a multiscale simulation. *Proceedings of the National Academy of Sciences of the United States of America* **2005**, *102* (19), 6783-8.

(127) Heller, H.; Grubmüller, H.; Schulten, K. Molecular dynamics simulation on a parallel computer. *Molecular Simulation* **1990**, *5* (3), 133-165.

(128) Pande, V. S.; Baker, I.; Chapman, J., et al. Atomistic protein folding simulations on the submillisecond time scale using worldwide distributed computing. *Biopolymers* **2003**, *68* (1), 91-109.

(129) Sanbonmatsu, K. Y.; Joseph, S.; Tung, C. S. Simulating movement of tRNA into the ribosome during decoding. *Proceedings of the National Academy of Sciences of the United States of America* **2005**, *102* (44), 15854-9.

(130) Duan, Y.; Kollman, P. A. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science* **1998**, *282* (5389), 740-4.

(131) Stein, M.; Gabdoulline, R. R.; Wade, R. C. Bridging from molecular simulation to biochemical networks. *Current Opinion in Structural Biology* **2007**, *17* (2), 166-72.

(132) Brooijmans, N.; Kuntz, I. D. Molecular recognition and docking algorithms. *Annu. Rev. Biophys. Biomol. Struct.* **2003**, *32*, 335-373.

(133) Waszkowycz, B.; Perkins, T. D. J.; Sykes, R. A.; Li, J. Large-scale virtual screening for discovering leads in the postgenomic era. *IBM Systems Journal* **2001**, *40* (2), 360-376.

(134) Oprea, T. I.; Matter, H. Integrating virtual screening in lead discovery. *Current Opinion in Chemical Biology* **2004**, *8* (4), 349-358.

(135) Mestres, J. Virtual screening: a real screening complement to high-throughput screening. *Biochem Soc Trans* **2002**, *30* (4), 797-9.

(136) Stahura, F. L.; Bajorath, J. Virtual screening methods that complement HTS. *Comb Chem High Throughput Screen* **2004**, *7* (4), 259-69.

(137) Davies, J. W.; Glick, M.; Jenkins, J. L. Streamlining lead discovery by aligning in silico and high-throughput screening. *Curr. Opin. Chem. Biol.* **2006**, *10* (4), 343-351.

(138) Martin, Y. C. Diverse viewpoints on computational aspects of molecular diversity. *J. Comb. Chem.* **2001**, *3* (3), 231-250.

(139) Perez, J. J. Managing molecular diversity. *Chemical Society Reviews* **2005**, *34* (2), 143-152.

(140) Monge, A.; Arrault, A.; Marot, C.; Morin-Allory, L. Managing, profiling and analyzing a library of 2.6 million compounds gathered from 32 chemical providers. *Mol. Diversity* **2006**, *10* (3), 389-403.

(141) Kazius, J.; McGuire, R.; Bursi, R. Derivation and validation of toxicophores for mutagenicity prediction. *Journal of Medicinal Chemistry* **2005**, *48* (1), 312-320.

(142) Liao, Q.; Yao, J.; Yuan, S. Prediction of mutagenic toxicity by combination of Recursive Partitioning and Support Vector Machines. *Molecular Diversity* **2007**, *11* (2), 59-72.

(143) Rishton, G. M. Reactive compounds and in vitro false positives in HTS. *Drug discovery today* **1997**, *2* (9), 382-384.

(144) Oprea, T. I. Property distribution of drug-related chemical databases. *J Comput Aided Mol Des* **2000**, *14* (3), 251-264.

(145) Veber, D. F.; Johnson, S. R.; Cheng, H. Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular properties that influence the oral bioavailability of drug candidates. *Journal of Medicinal Chemistry* **2002**, *45* (12), 2615-2623.

(146) Walters, W. P.; Murcko, M. A. Prediction of 'drug-likeness'. *Advanced Drug Delivery Reviews* **2002**, *54* (3), 255-271.

(147) Muegge, I. Selection criteria for drug-like compounds. *Medicinal Research Reviews* **2003**, *23* (3), 302-321.

(148) Oprea, T. I. Current trends in lead discovery: are we looking for the appropriate properties? *Molecular Diversity* **2002**, *5* (4), 199-208.

(149) Hann, M. M.; Oprea, T. I. Pursuing the leadlikeness concept in pharmaceutical research. *Current Opinion in Chemical Biology* **2004**, *8* (3), 255-263.

(150) Lyne, P. D. Structure-based virtual screening: an overview. *Drug Discovery Today* **2002**, *7* (20), 1047-1055.

(151) Lengauer, T.; Lemmen, C.; Rarey, M.; Zimmermann, M. Novel technologies for virtual screening. *Drug Discov. Today* **2004**, *9* (1), 27-34.

(152) Muegge, I.; Oloff, S. Advances in virtual screening. *Drug Discovery Today* **2006**, *3* (4), 405-411.

(153) Mason, J. S.; Good, A. C.; Martin, E. J. 3-D pharmacophores in drug discovery. *Curr. Pharm. Des.* **2001**, *7* (7), 567-597.

(154) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*. ed.; Wiley-VCH New York: 2000; 'Vol.' p.

(155) Abagyan, R.; Totrov, M. High-throughput docking and lead generation. *Curr. Opin. Chem. Biol.* **2001**, *5*, 375-382.

(156) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discovery* **2004**, *3* (11), 935-949.

(157) Rester, U. Dock around the Clock - Current Status of Small Molecule Docking and Scoring. *QSAR Comb. Sci.* **2006**, *25* (7), 605-615.

(158) Bursulaya, B. D.; Totrov, M.; Abagyan, R.; Brooks III, C. L. Comparative study of several algorithms for flexible ligand docking. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 755-763.

(159) Sousa, S. F.; Fernandes, P. A.; Ramos, M. J. Protein-ligand docking: current status and future challenges. *Proteins: Struct., Funct., Bioinf.* **2006**, *65* (1), 15-26.

(160) Kontoyianni, M.; McClellan, L. M.; Sokol, G. S. Evaluation of docking performance: comparative data on docking algorithms. *J. Med. Chem.* **2004**, *47* (3), 558-565.

(161) Perola, E.; Walters, W. P.; Charifson, P. S. A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins: Struct., Funct., Bioinf.* **2004**, *56* (2), 235-49.

(162) Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins: Struct., Funct., Bioinf.* **2004**, *57*, 225-242.

(163) Cummings, M. D.; DesJarlais, R. L.; Gibbs, A. C.; Mohan, V.; Jaeger, E. P. Comparison of automated docking programs as virtual screening tools. *J. Med. Chem.* **2005**, *48* (4), 962-976.

(164) Zhou, Z.; Felts, A. K.; Friesner, R. A.; Levy, R. M. Comparative Performance of Several Flexible Docking Programs and Scoring Functions: Enrichment Studies for a Diverse Set of Pharmaceutically Relevant Targets. *J. Chem. Inf. Model.* **2007**, *47* (4), 1599-1608.

(165) Leroux, V.; Maigret, B. Should structure-based virtual screening techniques be used more extensively in modern drug discovery? *Comp. App. Chem.* **2007**, *24* (1), 1-10.

(166) Mendez, R.; Leplae, R.; Lensink, M. F.; Wodak, S. J. Assessment of CAPRI predictions in rounds 3-5 shows progress in docking procedures. *Proteins: Struct., Funct., Bioinf.* **2005**, *60* (2), 150-169.

(167) Gohlke, H.; Klebe, G. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angew. Chem., Int. Ed. Engl.* **2002**, *41* (15), 2644-2676.

(168) Ritchie, D. W.; Kemp, G. J. L. Fast computation, rotation, and comparison of low resolution spherical harmonic molecular surfaces. *J. Comput. Chem.* **1999**, *20* (4), 383-395.

(169) Cai, W.; Shao, X.; Maigret, B. Protein-ligand recognition using spherical harmonic molecular surfaces: towards a fast and efficient filter for large virtual throughput screening. *J. Mol. Graph. Model.* **2002**, (4), 313-328.

(170) Knegtel, R. M.; Kuntz, I. D.; Oshiro, C. M. Molecular docking to ensembles of protein structures. *J. Mol. Biol.* **1997**, *266* (2), 424-440.

(171) Osterberg, F.; Morris, G. M.; Sanner, M. F.; Olson, A. J.; Goodsell, D. S. Automated docking to multiple target structures: incorporation of protein mobility and structural water heterogeneity in AutoDock. *Proteins: Struct., Funct., Genet.* **2002**, *46* (1), 34-40.

(172) Claussen, H.; Buning, C.; Rarey, M.; Lengauer, T. FlexE: efficient molecular docking considering protein structure variations. *J. Mol. Biol.* **2001**, *308* (2), 377-395.

(173) Barril, X.; Morley, S. D. Unveiling the full potential of flexible receptor docking using multiple crystallographic structures. *J. Med. Chem.* **2005**, *48* (13), 4432-4443.

(174) Wong, C. F.; Kua, J.; Zhang, Y.; Straatsma, T. P.; McCammon, J. A. Molecular docking of balanol to dynamics snapshots of protein kinase A. *Proteins: Struct., Funct., Bioinf.* **2005**, *61* (4), 850-858.

(175) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **1982**, *161* (2), 269-288.

(176) Friesner, R. A.; Banks, J. L.; Murphy, R. B., et al. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47* (7), 1739-1749.

(177) OpenEye Science Software: Santa Fe, NM, USA. http://www.eyesopen.com

(178) Accelrys Inc., San Diego, CA, USA. http://www.accelrys.com/products/catalyst/

(179) Miller, M. D.; Kearsley, S. K.; Underwood, D. J.; Sheridan, R. P. FLOG: a system to select 'quasi-flexible' ligands complementary to a receptor of known three-dimensional structure. *J. Comput.-Aided Mol. Des.* **1994**, *8* (2), 153-174.

(180) Pang, Y. P.; Perola, E.; Xu, K.; Prendergast, F. G. EUDOC: a computer program for identification of drug interaction sites in macromolecules and drug leads from chemical databases. *J. Comput. Chem.* **2001**, *22* (15), 1750-1771.

(181) Nicklaus, M. C.; Wang, S.; Driscoll, J. S.; Milne, G. W. Conformational changes of small molecules binding to proteins. *Bioorg. Med. Chem.* **1995**, *3* (4), 411-428.

(182) Ewing, T. J.; Makino, S.; Skillman, A. G.; Kuntz, I. D. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *J. Comput.-Aided Mol. Des.* **2001**, *15* (5), 411-428.

(183) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261* (3), 470-489.

(184) Welch, W.; Ruppert, J.; Jain, A. N. Hammerhead: fast, fully automated docking of flexible ligands to protein binding sites. *Chem. Biol.* **1996**, *3* (6), 449-462.

(185) Mitchell, M. *An introduction to genetic algorithms*. MIT Press, **1996**.

(186) Glover, F.; Laguna, M. *Tabu Search*. Springer, **1997**.

(187) Abagyan, R.; Totrov, M.; Kuznetsov, D. ICM - A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation. *J. Comput. Chem.* **1994**, *15* (5), 488-506.

(188) McMartin, C.; Bohacek, R. S. QXP: powerful, rapid computer algorithms for structure-based drug design. *J. Comput.-Aided Mol. Des.* **1997**, *11* (4), 333-344.

(189) Liu, M.; Wang, S. MCDOCK: a Monte Carlo simulation approach to the molecular docking problem. *J. Comput.-Aided Mol. Des.* **1999**, *13* (5), 435-451.

(190) Morris, G.; Goodsell, D.; Halliday, R.; Huey, R.; Hart, W.; Belew, R.; Olson, A. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **1998**, *19* (14), 1639-1662.

(191) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267* (3), 727-748.

(192) Taylor, J. S.; Burnett, R. M. DARWIN: a program for docking flexible molecules. *Proteins: Struct., Funct., Genet.* **2000**, *41* (2), 173-191.

(193) Baxter, C. A.; Murray, C. W.; Clark, D. E.; Westhead, D. R.; Eldridge, M. D. Flexible docking using Tabu search and an empirical estimate of binding affinity. *Proteins: Struct., Funct., Genet.* **1998**, *33* (3), 367-382.

(194) Koshland Jr., D. The key–lock theory and the induced fit theory. *Angew. Chem., Int. Ed. Engl.* **1994**, *33* (23-24), 2375-2378.

(195) Erickson, J. A.; Jalaie, M.; Robertson, D. H.; Lewis, R. A.; Vieth, M. Lessons in molecular recognition: the effects of ligand and protein flexibility on molecular docking accuracy. *J. Med. Chem.* **2004**, *47* (1), 45-55.

(196) Teague, S. J. Implications of protein flexibility for drug discovery. *Nat. Rev. Drug Discovery* **2003**, *2* (7), 527-541.

(197) Ferrari, A. M.; Wei, B. Q.; Costantino, L.; Shoichet, B. K. Soft docking and multiple receptor conformations in virtual screening. *J. Med. Chem.* **2004**, *47* (21), 5076-5084.

(198) Leach, A. R. Ligand docking to proteins with discrete side-chain flexibility. *J. Mol. Biol.* **1994**, *235* (1), 345-356.

(199) Kallblad, P.; Dean, P. M. Efficient conformational sampling of local side-chain flexibility. *J. Mol. Biol.* **2003**, *326* (5), 1651-1665.

(200) Meiler, J.; Baker, D. ROSETTALIGAND: protein-small molecule docking with full side-chain flexibility. *Proteins: Struct., Funct., Bioinf.* **2006**, *65* (3), 538-548.

(201) Korb, O.; Stützle, T.; Exner, T. *PLANTS: Application of Ant Colony Optimization to Structure-Based Drug Design*. ed.; Springer Berlin / Heidelberg: 2006; 'Vol.' p 247-258.

(202) Chen, H. M.; Liu, B. F.; Huang, H. L.; Hwang, S. F.; Ho, S. Y. SODOCK: swarm optimization for highly flexible protein-ligand docking. *J. Comput. Chem.* **2007**, *28* (2), 612-623.

(203) Banks, A.; Vincent, J.; Anyakoha, C. A review of particle swarm optimization. Part I: background and development. *Nat. Comput.* **2007**, *6* (4), 467-484.

(204) Kollman, P. Free energy calculations: Applications to chemical and biochemical phenomena. *Chem. Rev.* **1993**, *93* (7), 2395-2417.

(205) Simonson, T.; Archontis, G.; Karplus, M. Free energy simulations come of age: protein-ligand recognition. *Acc. Chem. Res.* **2002**, *35* (6), 430-437.

(206) Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks III, C. L. Assessing scoring functions for protein-ligand interactions. *J. Med. Chem.* **2004**, *47*, 3032-3047.

(207) Schulz-Gasch, T.; Stahl, M. Scoring functions for protein-ligand interactions: a critical perspective. *Drug Discov. Today* **2004**, *1* (3), 231-239.

(208) Wang, R.; Lu, Y.; Fang, X.; Wang, S. An extensive test of 14 scoring functions using the PDBbind refined set of 800 protein-ligand complexes. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (6), 2114-2125.

(209) Kramer, B.; Rarey, M.; Lengauer, T. Evaluation of the FlexX incremental construction algorithm for protein-ligand docking. *Proteins: Struct., Funct., Genet.* **1999**, *37* (2), 228-241.

(210) Goodsell, D. S.; Morris, G. M.; Olson, A. J. Automated docking of flexible ligands: applications of AutoDock. *J. Mol. Recognit.* **1996**, *9* (1), 1-5.

(211) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput.-Aided Mol. Des.* **1997**, *11* (5), 425-445.

(212) Gehlhaar, D. K.; Verkhivker, G. M.; Rejto, P. A.; Sherman, C. J.; Fogel, D. B.; Fogel, L. J.; Freer, S. T. Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming. *Chemical Biology* **1995**, *2* (5), 317-324.

(213) Krammer, A.; Kirchhoff, P. D.; Jiang, X.; Venkatachalam, C. M.; Waldman, M. LigScore: a novel scoring function for predicting binding affinities. *Journal of Molecular Graphics and Modelling* **2005**, *23* (5), 395-407.

(214) Muegge, I.; Martin, Y. C. A General and Fast Scoring Function for Protein-Ligand Interactions: A Simplified Potential Approach. *J. Med. Chem.* **1999**, *42* (5), 791-804.

(215) DeWitte, R. S.; Shakhnovich, E. I. SMoG: de Novo Design Method Based on Simple, Fast, and Accurate Free Energy Estimates. 1. Methodology and Supporting Evidence. *J. Am. Chem. Soc.* **1996**, *118* (47), 11733-11744.

(216) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol.* **2000**, *295* (2), 337-356.

(217) Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J. Med. Chem.* **1999**, *42* (25), 5100-5109.

(218) Feher, M. Consensus scoring for protein-ligand interactions. *Drug Discov. Today* **2006**, *11* (9-10), 421-428.

(219) Venkatachalam, C. M.; Jiang, X.; Oldfield, T.; Waldman, M. LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. *J. Mol. Graph. Model.* **2003**, *21* (4), 289-307.

(220) McGann, M. R.; Almond, H. R.; Nicholls, A.; Grant, J. A.; Brown, F. K. Gaussian docking functions. *Biopolymers* **2003**, *68* (1), 76-90.

(221) Cai, W.; Xu, J.; Shao, X.; Leroux, V.; Beautrait, A.; Maigret, B. SHEF: a vHTS geometrical filter using coefficients of spherical harmonics molecular surfaces. *Journal of Molecular Modeling* **2007**, *submitted*.

(222) Kuhn, B.; Gerber, P.; Schulz-Gasch, T.; Stahl, M. Validation and use of the MM-PBSA approach for drug discovery. *J. Med. Chem.* **2005**, *48* (12), 4040-4048.

(223) Cole, J. C.; Murray, C. W.; Nissink, J. W.; Taylor, R. D.; Taylor, R. Comparing protein-ligand docking programs is difficult. *Proteins: Struct., Funct., Bioinf.* **2005**, *60* (3), 325-332.

(224) Leroux, V.; Maigret, M. Should structure-based virtual screening techniques be used more extensively in modern drug discovery? *Computers and Applied Chemistry* **2007**, *24* (1), 1-10.

(225) Laederach, A.; Reilly, P. J. Specific empirical free energy function for automated docking of carbohydrates to proteins. *J. Comput. Chem.* **2003**, *24* (14), 1748-1757.

(226) Corbeil, C. R.; Englebienne, P.; Moitessier, N. Docking ligands into flexible and solvated macromolecules. 1. Development and validation of FITTED 1.0. *J. Chem. Inf. Model.* **2007**, *47* (2), 435-449.

(227) Zhao, Y.; Sanner, M. F. FLIPDock: docking flexible ligands into flexible receptors. *Proteins: Struct., Funct., Bioinf.* **2007**, *68* (3), 726-737.

(228) Birkholtz, L. M.; Bastien, O.; Wells, G., et al. Integration and mining of malaria molecular, functional and pharmacological data: how far are we from a chemogenomic knowledge space? *Malar. J.* **2006**, *5*, 110.

(229) Zauhar, R. J.; Moyna, G.; Tian, L.; Li, Z.; Welsh, W. J. Shape signatures: a new approach to computer-aided ligand- and receptor-based drug design. *J. Med. Chem.* **2003**, *46* (26), 5674-5690.

(230) Rush, T. S., 3rd; Grant, J. A.; Mosyak, L.; Nicholls, A. A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. *J. Med. Chem.* **2005**, *48* (5), 1489-1495.

(231) Mavridis, L.; Hudson, B. D.; Ritchie, D. W. Toward high throughput 3D virtual screening using spherical harmonic molecular surface representations. *J. Chem. Inf. Model.* **2007**, *47*, 1787-1796.

(232) Morris, R. J. An evaluation of spherical designs for molecular-like surfaces. *J. Mol. Graph. Model.* **2006**, *24* (5), 356-361.

(233) Cappello, F.; Caron, E.; Dayde, M., et al. In *Grid'5000: a large scale and highly reconfigurable grid experimental testbed*, The 6th IEEE/ACM International Workshop on Grid Computing, **2005**; pp 8.

(234) Kubinyi, H. QSAR and 3D QSAR in drug design part 1: Methodology. *Drug Discov. Today* **1997**, *2* (11), 457-467.

(235) Kubinyi, H. QSAR and 3D QSAR in drug design part 2: Applications and problems. *Drug Discov. Today* **1997**, *2* (12), 538-546.

(236) Deng, Z.; Chuaqui, C.; Singh, J. Structural interaction fingerprint (SIFt): a novel method for analyzing three-dimensional protein-ligand binding interactions. *J. Med. Chem.* **2004**, *47* (2), 337-344.

(237) Marcou, G.; Rognan, D. Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. *J. Chem. Inf. Model.* **2007**, *47* (1), 195-207.

(238) Zamir, E.; Geiger, B. Molecular complexity and dynamics of cell-matrix adhesions. *J. Cell Sci.* **2001**, *114* (Pt 20), 3583-3590.

(239) Lo, S. H. Focal adhesions: what's new inside. *Dev. Biol.* **2006**, *294* (2), 280-291.

(240) Schaller, M. D.; Borgman, C. A.; Cobb, B. S.; Vines, R. R.; Reynolds, A. B.; Parsons, J. T. pp125FAK a structurally distinctive protein-tyrosine kinase associated with focal adhesions. *Proceedings of the National Academy of Sciences of the United States of America* **1992**, *89* (11), 5192-5196.

(241) Hanks, S. K.; Calalb, M. B.; Harper, M. C.; Patel, S. K. Focal adhesion protein-tyrosine kinase phosphorylated in response to cell attachment to fibronectin. *Proceedings of the National Academy of Sciences of the United States of America* **1992**, *89* (18), 8487-8491.

(242) Zachary, I.; Rozengurt, E. Focal adhesion kinase ($p125^{FAK}$): A point of convergence in the action of neuropeptides, integrins, and oncogenes. *Cell* **1992**, *71*, 891-894.

(243) Schlaepfer, D. D.; Hauck, C. R.; Sieg, D. J. Signaling through focal adhesion kinase. *Prog. Biophys. Mol. Bio.* **1999**, *71* (3-4), 435-78.

(244) Giancotti, F. G.; Ruoslahti, E. Integrin signaling. *Science* **1999**, *285* (5430), 1028-1032.

(245) Parsons, J. T. Focal adhesion kinase: the first ten years. *Journal of Cell Science* **2003**, *116* (Pt 8), 1409-1416.

(246) Schaller, M. D. Biochemical signals and biological responses elicited by the focal adhesion kinase. *Biochim. Biophys. Acta* **2001**, *1540* (1), 1-21.

(247) Cornillon, J.; Campos, L.; Guyotat, D. Focal adhesion kinase (FAK), a multifunctional protein. *M. S. Méd. Sci.* **2003**, *19* (6-7), 743-752.

(248) Mitra, S. K.; Hanson, D. A.; Schlaepfer, D. D. Focal adhesion kinase: in command and control of cell motility. *Nat Rev Mol Cell Biol* **2005**, *6* (1), 56-68.

(249) Schlessinger, J. SH2/SH3 signaling proteins. *Curr. Opin. Genet. Dev.* **1994**, *4* (1), 25-30.

(250) Gabarra-Niecko, V.; Schaller, M. D.; Dunty, J. M. FAK regulates biological processes important for the pathogenesis of cancer. *Cancer and Metastasis Reviews* **2003**, *22* (4), 359-374.

(251) Hecker, T. P.; Gladson, C. L. Focal adhesion kinase in cancer. *Front. Biosci.* **2003**, *8*, s705-714.

(252) Schlaepfer, D. D.; Mitra, S. K.; Ilic, D. Control of motile and invasive cell phenotypes by focal adhesion kinase. *Biochimica et Biophysica Acta* **2004**, *1692* (2-3), 77-102.

(253) McLean, G. W.; Carragher, N. O.; Avizienyte, E.; Evans, J.; Brunton, V. G.; Frame, M. C. The role of focal-adhesion kinase in cancer - a new therapeutic opportunity. *Nature Reviews Cancer* **2005**, *5* (7), 505-515.

(254) van Nimwegen, M. J.; van de Water, B. Focal adhesion kinase: a potential target in cancer therapy. *Biochemical Pharmacology* **2007**, *73* (5), 597-609.

(255) Jones, R. J.; Brunton, V. G.; Frame, M. C. Adhesion-linked kinases in cancer; emphasis on src, focal adhesion kinase and PI 3-kinase. *Eur. J. Cancer* **2000**, *36* (13 Spec No), 1595-1606.

(256) Mitra, S. K.; Hanson, D. A.; Schlaepfer, D. D. Focal adhesion kinase: in command and control of cell motility. *Nat. Rev. Mol. Cell Biol.* **2005**, *6* (1), 56-68.

(257) Cooper, L. A.; Shen, T. L.; Guan, J. L. Regulation of focal adhesion kinase by its amino-terminal domain through an autoinhibitory interaction. *Molecular and Cellular Biology* **2003**, *23* (22), 8030-8041.

(258) Dunty, J. M.; Gabarra-Niecko, V.; King, M. L.; Ceccarelli, D. F.; Eck, M. J.; Schaller, M. D. FERM domain interaction promotes FAK signaling. *Mol. Cell. Biol.* **2004**, *24* (12), 5353-5368.

(259) Calalb, M. B.; Polte, T. R.; Hanks, S. K. Tyrosine phosphorylation of focal adhesion kinase at sites in the catalytic domain regulates kinase activity: a role for Src family kinases. *Mol. Cell. Biol.* **1995**, *15* (2), 954-963.

(260) Owen, J. D.; Ruest, P. J.; Fry, D. W.; Hanks, S. K. Induced focal adhesion kinase (FAK) expression in FAK-null cells enhances cell spreading and migration requiring both auto- and activation loop phosphorylation sites and inhibits adhesion-dependent tyrosine phosphorylation of Pyk2. *Mol. Cell. Biol.* **1999**, *19* (7), 4806-4818.

(261) Schaller, M. D.; Hildebrand, J. D.; Shannon, J. D.; Fox, J. W.; Vines, R. R.; Parsons, J. T. Autophosphorylation of the focal adhesion kinase, pp125FAK, directs SH2-dependent binding of pp60src. *Mol. Cell. Biol.* **1994**, *14* (3), 1680-1688.

(262) Harte, M. T.; Hildebrand, J. D.; Burnham, M. R.; Bouton, A. H.; Parsons, J. T. p130Cas, a substrate associated with v-Src and v-Crk, localizes to focal adhesions and binds to focal adhesion kinase. *J. Biol. Chem.* **1996**, *271* (23), 13649-13655.

(263) Hildebrand, J. D.; Taylor, J. M.; Parsons, J. T. An SH3 domain-containing GTPase-activating protein for Rho and Cdc42 associates with focal adhesion kinase. *Mol. Cell. Biol.* **1996**, *16* (6), 3169-3178.

(264) Turner, C. E.; Miller, J. T. Primary sequence of paxillin contains putative SH2 and SH3 domain binding motifs and multiple LIM domains: identification of a vinculin and pp125Fak-binding region. *J. Cell Sci.* **1994**, *107 (Pt 6)*, 1583-1591.

(265) Hildebrand, J. D.; Schaller, M. D.; Parsons, J. T. Paxillin, a tyrosine phosphorylated focal adhesion-associated protein binds to the carboxyl terminal domain of focal adhesion kinase. *Mol. Biol. Cell* **1995**, *6* (6), 637-647.

(266) Chen, H. C.; Appeddu, P. A.; Parsons, J. T.; Hildebrand, J. D.; Schaller, M. D.; Guan, J. L. Interaction of focal adhesion kinase with cytoskeletal protein talin. *J. Biol. Chem.* **1995**, *270* (28), 16995-16999.

(267) Tachibana, K.; Sato, T.; D'Avirro, N.; Morimoto, C. Direct association of pp125FAK with paxillin, the focal adhesion-targeting mechanism of pp125FAK. *J. Exp. Med.* **1995**, *182* (4), 1089-1099.

(268) Cohen, L. A.; Guan, J. L. Residues within the first subdomain of the FERM-like domain in focal adhesion kinase are important in its regulation. *Journal of Biological Chemistry* **2005**, *280* (9), 8197-8207.

(269) Sieg, D. J.; Hauck, C. R.; Ilic, D.; Klingbeil, C. K.; Schaefer, E.; Damsky, C. H.; Schlaepfer, D. D. FAK integrates growth-factor and integrin signals to promote cell migration. *Nat. Cell Biol.* **2000**, *2* (5), 249-256.

(270) Streblow, D. N.; Vomaske, J.; Smith, P., et al. Human cytomegalovirus chemokine receptor US28-induced smooth muscle cell migration is mediated by focal adhesion kinase and Src. *J. Biol. Chem.* **2003**, *278* (50), 50456-50465.

(271) Kornberg, L.; Earp, H. S.; Parsons, J. T.; Schaller, M.; Juliano, R. L. Cell adhesion or integrin clustering increases phosphorylation of a focal adhesion-associated tyrosine kinase. *J. Biol. Chem.* **1992**, *267* (33), 23439-23442.

(272) Toutant, M.; Costa, A.; Studler, J. M.; Kadare, G.; Carnaud, M.; Girault, J. A. Alternative splicing controls the mechanisms of FAK autophosphorylation. *Mol. Cell. Biol.* **2002**, *22* (22), 7731-7743.

(273) Xing, Z.; Chen, H. C.; Nowlen, J. K.; Taylor, S. J.; Shalloway, D.; Guan, J. L. Direct interaction of v-Src with the focal adhesion kinase mediated by the Src SH2 domain. *Mol. Biol. Cell* **1994**, *5* (4), 413-421.

(274) Abbi, S.; Guan, J. L. Focal adhesion kinase: protein interactions and cellular functions. *Histol. Histopathol.* **2002**, *17* (4), 1163-1171.

(275) Schlaepfer, D. D.; Hanks, S. K.; Hunter, T.; van der Geer, P. Integrin-mediated signal transduction linked to Ras pathway by GRB2 binding to focal adhesion kinase. *Nature* **1994**, *372* (6508), 786-791.

(276) Schlaepfer, D. D.; Hunter, T. Evidence for in vivo phosphorylation of the Grb2 SH2-domain binding site on focal adhesion kinase by Src-family protein-tyrosine kinases. *Mol. Cell. Biol.* **1996**, *16* (10), 5623-5633.

(277) Frisch, S. M.; Vuori, K.; Ruoslahti, E.; Chan-Hui, P. Y. Control of adhesion-dependent cell survival by focal adhesion kinase. *J. Cell Biol.* **1996**, *134* (3), 793-799.

(278) Mitra, S. K.; Schlaepfer, D. D. Integrin-regulated FAK-Src signaling in normal and cancer cells. *Curr. Opin. Cell Biol.* **2006**, *18* (5), 516-523.

(279) Ilic, D.; Almeida, E. A.; Schlaepfer, D. D.; Dazin, P.; Aizawa, S.; Damsky, C. H. Extracellular matrix survival signals transduced by focal adhesion kinase suppress p53-mediated apoptosis. *J. Cell Biol.* **1998**, *143* (2), 547-560.

(280) Golubovskaya, V. M.; Finch, R.; Cance, W. G. Direct interaction of the N-terminal domain of focal adhesion kinase with the N-terminal transactivation domain of p53. *J. Biol. Chem.* **2005**, *280* (26), 25008-25021.

(281) Datta, S. R.; Dudek, H.; Tao, X.; Masters, S.; Fu, H.; Gotoh, Y.; Greenberg, M. E. Akt phosphorylation of BAD couples survival signals to the cell-intrinsic death machinery. *Cell* **1997**, *91* (2), 231-241.

(282) Sonoda, Y.; Watanabe, S.; Matsumoto, Y.; Aizu-Yokota, E.; Kasahara, T. FAK is the upstream signal protein of the phosphatidylinositol 3-kinase-Akt survival pathway in hydrogen peroxide-induced apoptosis of a human glioblastoma cell line. *J. Biol. Chem.* **1999**, *274* (15), 10566-10570.

(283) Almeida, E. A.; Ilic, D.; Han, Q., et al. Matrix survival signaling: from fibronectin via focal adhesion kinase to c-Jun NH(2)-terminal kinase. *J. Cell Biol.* **2000**, *149* (3), 741-754.

(284) Yamamoto, D.; Sonoda, Y.; Hasegawa, M.; Funakoshi-Tago, M.; Aizu-Yokota, E.; Kasahara, T. FAK overexpression upregulates cyclin D3 and enhances cell proliferation via the PKC and PI3-kinase-Akt pathways. *Cell. Signal.* **2003**, *15* (6), 575-583.

(285) Giannone, G.; Ronde, P.; Gaire, M.; Beaudouin, J.; Haiech, J.; Ellenberg, J.; Takeda, K. Calcium rises locally trigger focal adhesion disassembly and enhance residency of focal adhesion kinase at focal adhesions. *J. Biol. Chem.* **2004**, *279* (27), 28715-28723.

(286) Cary, L. A.; Han, D. C.; Polte, T. R.; Hanks, S. K.; Guan, J. L. Identification of p130Cas as a mediator of focal adhesion kinase-promoted cell migration. *J. Cell Biol.* **1998**, *140* (1), 211-221.

(287) Reiske, H. R.; Kao, S. C.; Cary, L. A.; Guan, J. L.; Lai, J. F.; Chen, H. C. Requirement of phosphatidylinositol 3-kinase in focal adhesion kinase-promoted cell migration. *J. Biol. Chem.* **1999**, *274* (18), 12361-12366.

(288) Bianchi, M.; De Lucchini, S.; Marin, O.; Turner, D. L.; Hanks, S. K.; Villa-Moruzzi, E. Regulation of FAK Ser-722 phosphorylation and kinase activity by GSK3 and PP1 during cell spreading and migration. *Biochem. J.* **2005**, *391* (Pt 2), 359-370.

(289) Mitra, S. K.; Mikolon, D.; Molina, J. E., et al. Intrinsic FAK activity and Y925 phosphorylation facilitate an angiogenic switch in tumors. *Oncogene* **2006**, *25* (44), 5969-5984.

(290) Hauck, C. R.; Sieg, D. J.; Hsia, D. A.; Loftus, J. C.; Gaarde, W. A.; Monia, B. P.; Schlaepfer, D. D. Inhibition of focal adhesion kinase expression or activity disrupts epidermal growth factor-stimulated signaling promoting the migration of invasive human carcinoma cells. *Cancer Res.* **2001**, *61* (19), 7079-7090.

(291) Benlimame, N.; He, Q.; Jie, S., et al. FAK signaling is critical for ErbB-2/ErbB-3 receptor cooperation for oncogenic transformation and invasion. *J. Cell Biol.* **2005**, *171* (3), 505-516.

(292) Agochiya, M.; Brunton, V. G.; Owens, D. W.; Parkinson, E. K.; Paraskeva, C.; Keith, W. N.; Frame, M. C. Increased dosage and amplification of the focal adhesion kinase gene in human cancer cells. *Oncogene* **1999**, *18* (41), 5646-5653.

(293) Maung, K.; Easty, D. J.; Hill, S. P.; Bennett, D. C. Requirement for focal adhesion kinase in tumor cell adhesion. *Oncogene* **1999**, *18* (48), 6824-8.

(294) Xu, L. H.; Owens, L. V.; Sturge, G. C.; Yang, X.; Liu, E. T.; Craven, R. J.; Cance, W. G. Attenuation of the expression of the focal adhesion kinase induces apoptosis in tumor cells. *Cell Growth Differ.* **1996**, *7* (4), 413-418.

(295) Xu, L. H.; Yang, X.; Bradham, C. A.; Brenner, D. A.; Baldwin, A. S., Jr.; Craven, R. J.; Cance, W. G. The focal adhesion kinase suppresses transformation-associated, anchorage-independent apoptosis in human breast cancer cells. Involvement of death receptor-related signaling pathways. *J. Biol. Chem.* **2000**, *275* (39), 30597-30604.

(296) Beviglia, L.; Golubovskaya, V.; Xu, L.; Yang, X.; Craven, R. J.; Cance, W. G. Focal adhesion kinase N-terminus in breast carcinoma cells induces rounding, detachment and apoptosis. *Biochem. J.* **2003**, *373* (Pt 1), 201-210.

(297) Choi, H. S.; Wang, Z.; Richmond, W., et al. Design and synthesis of 7H-pyrrolo[2,3-d]pyrimidines as focal adhesion kinase inhibitors. Part 1. *Bioorg. Med. Chem. Lett.* **2006**, *16* (8), 2173-2176.

(298) Choi, H. S.; Wang, Z.; Richmond, W., et al. Design and synthesis of 7H-pyrrolo[2,3-d]pyrimidines as focal adhesion kinase inhibitors. Part 2. *Bioorg. Med. Chem. Lett.* **2006**, *16* (10), 2689-2692.

(299) Shi, Q.; Hjelmeland, A. B.; Keir, S. T., et al. A novel low-molecular weight inhibitor of focal adhesion kinase, TAE226, inhibits glioma growth. *Mol. Carcinog.* **2007**, *46* (6), 488-496.

(300) Slack-Davis, J. K.; Martin, K. H.; Tilghman, R. W., et al. Cellular characterization of a novel focal adhesion kinase inhibitor. *J. Biol. Chem.* **2007**, *282* (20), 14845-14852.

(301) Siu, L.; Burris, H.; Mileshkin, L., et al. Phase 1 study of a focal adhesion kinase (FAK) inhibitor PF-00562271 in patients (pts) with advanced solid tumors. *J. Clin. Oncol.* **2007**, *25* (18_suppl), 3527.

(302) Garces, C. A.; Kurenova, E. V.; Golubovskaya, V. M.; Cance, W. G. Vascular endothelial growth factor receptor-3 and focal adhesion kinase bind and suppress apoptosis in breast cancer cells. *Cancer Res.* **2006**, *66* (3), 1446-1454.

(303) Ceccarelli, D. F.; Song, H. K.; Poy, F.; Schaller, M. D.; Eck, M. J. Crystal structure of the FERM domain of focal adhesion kinase. *Journal of Biological Chemistry* **2006**, *281* (1), 252-259.

(304) Nowakowski, J.; Cronin, C. N.; McRee, D. E., et al. Structures of the cancer-related Aurora-A, FAK, and EphA2 protein kinases from nanovolume crystallography. *Structure* **2002**, *10* (12), 1659-1667.

(305) Lietha, D.; Cai, X.; Ceccarelli, D. F.; Li, Y.; Schaller, M. D.; Eck, M. J. Structural basis for the autoinhibition of focal adhesion kinase. *Cell* **2007**, *129* (6), 1177-1187.

(306) Hayashi, I.; Vuori, K.; Liddington, R. C. The focal adhesion targeting (FAT) region of focal adhesion kinase is a four-helix bundle that binds paxillin. *Nature Structural & Molecular Biology* **2002**, *9* (2), 101-106.

(307) Arold, S. T.; Hoellerer, M. K.; Noble, M. E. The structural basis of localization and signaling by the focal adhesion targeting domain. *Structure* **2002**, *10* (3), 319-327.

(308) Prutzman, K. C.; Gao, G.; King, M. L.; Iyer, V. V.; Mueller, G. A.; Schaller, M. D.; Campbell, S. L. The focal adhesion targeting domain of focal adhesion kinase contains a hinge region that modulates tyrosine 926 phosphorylation. *Structure* **2004**, *12* (5), 881-891.

(309) Liu, G.; Guibao, C. D.; Zheng, J. Structural insight into the mechanisms of targeting and signaling of focal adhesion kinase. *Molecular and Cellular Biology* **2002**, *22* (8), 2751-2760.

(310) Hoellerer, M. K.; Noble, M. E.; Labesse, G.; Campbell, I. D.; Werner, J. M.; Arold, S. T. Molecular recognition of paxillin LD motifs by the focal adhesion targeting domain. *Structure* **2003**, *11* (10), 1207-17.

(311) Gao, G.; Prutzman, K. C.; King, M. L., et al. NMR solution structure of the focal adhesion targeting domain of focal adhesion kinase in complex with a paxillin LD peptide: evidence for a two-site binding model. *J Biol Chem* **2004**, *279* (9), 8441-51.

(312) Rohl, C. A.; Strauss, C. E.; Misura, K. M.; Baker, D. Protein structure prediction using Rosetta. *Methods enzymol.* **2004**, *383*, 66-93.

(313) Tsutakawa, S. E.; Hura, G. L.; Frankel, K. A.; Cooper, P. K.; Tainer, J. A. Structural analysis of flexible proteins in solution by small angle X-ray scattering combined with crystallography. *J Struct Biol* **2007**, *158* (2), 214-223.

(314) Gherardi, E.; Sandin, S.; Petoukhov, M. V., et al. Structural basis of hepatocyte growth factor/scatter factor and MET signalling. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103* (11), 4046-4051.

(315) Hildebrand, J. D.; Schaller, M. D.; Parsons, J. T. Identification of sequences required for the efficient localization of the focal adhesion kinase, pp125FAK, to cellular focal adhesions. *J. Cell Biol.* **1993**, *123* (4), 993-1005.

(316) Shen, Y.; Schaller, M. D. Focal adhesion targeting: the critical determinant of FAK regulation and substrate phosphorylation. *Mol. Biol. Cell* **1999**, *10* (8), 2507-2518.

(317) Turner, C. E. Paxillin and focal adhesion signalling. *Nat. Cell Biol.* **2000**, *2* (12), E231-236.

(318) Brown, M. C.; Turner, C. E. Paxillin: adapting to change. *Physiol Rev* **2004**, *84* (4), 1315-1339.

(319) Nayal, A.; Webb, D. J.; Horwitz, A. F. Talin: an emerging focal point of adhesion dynamics. *Curr. Opin. Cell Biol.* **2004**, *16* (1), 94-98.

(320) Cooley, M. A.; Broome, J. M.; Ohngemach, C.; Romer, L. H.; Schaller, M. D. Paxillin binding is not the sole determinant of focal adhesion localization or dominant-negative activity of focal adhesion kinase/focal adhesion kinase-related nonkinase. *Mol. Biol. Cell* **2000**, *11* (9), 3247-3263.

(321) Downward, J. Targeting RAS signalling pathways in cancer therapy. *Nat. Rev. Cancer* **2003**, *3* (1), 11-22.

(322) Molina, J. R.; Adjei, A. A. The Ras/Raf/MAPK pathway. *J. Thorac. Oncol.* **2006**, *1* (1), 7-9.

(323) Dhillon, A. S.; Hagan, S.; Rath, O.; Kolch, W. MAP kinase signalling pathways in cancer. *Oncogene* **2007**, *26* (22), 3279-3290.

(324) Katz, B. Z.; Romer, L.; Miyamoto, S., et al. Targeting membrane-localized focal adhesion kinase to focal adhesions: roles of tyrosine phosphorylation and SRC family kinases. *J. Biol. Chem.* **2003**, *278* (31), 29115-29120.

(325) Nakamura, K.; Yano, H.; Schaefer, E.; Sabe, H. Different modes and qualities of tyrosine phosphorylation of Fak and Pyk2 during epithelial-mesenchymal transdifferentiation and cell migration: analysis of specific phosphorylation events using site-directed antibodies. *Oncogene* **2001**, *20* (21), 2626-2635.

(326) Laskowski, R. A.; Chistyakov, V. V.; Thornton, J. M. PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids. *Nucl. Acids Res.* **2005**, *33* (suppl_1), D266-268.

(327) Brown, M. C.; Perrotta, J. A.; Turner, C. E. Identification of LIM3 as the principal determinant of paxillin focal adhesion localization and characterization of a novel motif on paxillin directing vinculin and focal adhesion kinase binding. *J. Cell Biol.* **1996**, *135* (4), 1109-1123.

(328) Tumbarello, D. A.; Brown, M. C.; Turner, C. E. The paxillin LD motifs. *FEBS Lett.* **2002**, *513* (1), 114-118.

(329) Thomas, J. W.; Cooley, M. A.; Broome, J. M.; Salgia, R.; Griffin, J. D.; Lombardo, C. R.; Schaller, M. D. The role of focal adhesion kinase binding in the regulation of tyrosine phosphorylation of paxillin. *J. Biol. Chem.* **1999**, *274* (51), 36684-36692.

(330) Bertolucci, C. M.; Guibao, C. D.; Zheng, J. Structural features of the focal adhesion kinase-paxillin complex give insight into the dynamics of focal adhesion assembly. *Protein Sci.* **2005**, *14* (3), 644-652.

(331) Dixon, R. D.; Chen, Y.; Ding, F., et al. New insights into FAK signaling and localization based on detection of a FAT domain folding intermediate. *Structure* **2004**, *12* (12), 2161-2171.

(332) Hubbard, S. R. Crystal structure of the activated insulin receptor tyrosine kinase in complex with peptide substrate and ATP analog. *EMBO J.* **1997**, *16* (18), 5572-5581.

(333) Brown, N. R.; Noble, M. E.; Endicott, J. A.; Johnson, L. N. The structural basis for specificity of substrate and recruitment peptides for cyclin-dependent kinases. *Nat. Cell Biol.* **1999**, *1* (7), 438-443.

(334) Kuriyan, J.; Cowburn, D. Modular peptide recognition domains in eukaryotic signaling. *Annu. Rev. Biophys. Biomol. Struct.* **1997**, *26*, 259-288.

(335) Mak, P.; He, Z.; Kurosaki, T. Identification of amino acid residues required for a specific interaction between Src-tyrosine kinase and proline-rich region of phosphatidylinositol-3' kinase. *FEBS Lett.* **1996**, *397* (2-3), 183-185.

(336) Ding, F.; Prutzman, K. C.; Campbell, S. L.; Dokholyan, N. V. Topological determinants of protein domain swapping. *Structure* **2006**, *14* (1), 5-14.

(337) Zhou, Z.; Feng, H.; Bai, Y. Detection of a hidden folding intermediate in the focal adhesion target domain: Implications for its function and folding. *Proteins* **2006**, *65* (2), 259-265.

(338) Isralewitz, B.; Baudry, J.; Gullingsrud, J.; Kosztin, D.; Schulten, K. Steered molecular dynamics investigations of protein function. *J. Mol. Graph. Model.* **2001**, *19* (1), 13-25.

(339) Mofrad, M. R.; Golji, J.; Abdul Rahim, N. A.; Kamm, R. D. Force-induced unfolding of the focal adhesion targeting domain and the influence of paxillin binding. *Mech. Chem. Biosyst.* **2004**, *1* (4), 253-265.

(340) Arkin, M. R.; Wells, J. A. Small-molecule inhibitors of protein-protein interactions: progressing towards the dream. *Nat. Rev. Drug Discovery* **2004**, *3* (4), 301-317.

(341) Fletcher, S.; Hamilton, A. D. Protein surface recognition and proteomimetics: mimics of protein surface structure and function. *Curr. Opin. Chem. Biol.* **2005**, *9* (6), 632-638.

(342) Gao, G.; Prutzman, K. C.; King, M. L., et al. NMR solution structure of the focal adhesion targeting domain of focal adhesion kinase in complex with a paxillin LD peptide: evidence for a two-site binding model. *J. Biol. Chem.* **2004**, *279* (9), 8441-8451.

(343) Momany, F. A.; Rone, R. Validation of the general purpose QUANTA 3.2/CHARMm force field. *J. Comput. Chem.* **1992**, *13* (7), 888-900.

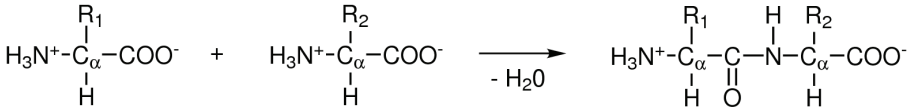
ANNEXE

LA STRUCTURE DES PROTÉINES

1. Les acides aminés et la liaison peptidique

Les propriétés spécifiques des protéines (fonctions enzymatiques et mécaniques, stabilité thermique,…) sont assurées par leur conformation native, c'est-à-dire par le repliement qu'elles adoptent au sein de la cellule.

Une protéine est un polymère dont les unités monomériques sont des acides aminés (ou résidus). L'enchaînement des acides aminés est réalisé par l'élimination d'une molécule d'eau pour former une liaison peptidique :



Formation d'une liaison peptidique

Le groupement chimique R, ou « chaîne latérale », est porté par un carbone Cα. Ce groupe R appartient à une liste de 20 acides aminés possibles (voir FIG.3b). Mis à part pour la glycine, le Cα est un carbone asymétrique. Pour les acides aminés naturels, la configuration stéréochimique de ce centre chiral est L.

La distance C-N dans la liaison peptidique est de 1.325 Å. Ce caractère de double liaison conduit à placer les quatre atomes C, O, N et H dans un même plan. La jonction de l'acide aminé en position *i-1* avec celui en position *i* crée un plan où se trouvent les atomes O_{i-1}, C_{i-1}, N_i et H_i. Les seuls degrés de liberté qui subsistent correspondent aux rotations autour de deux liaisons N_i-Cα_i et Cα- C_i, qui sont définies respectivement par les angles ϕ et ψ.

2. Organisation hiérarchique

La structure des protéines est organisée de manière hiérarchique et comporte 4 niveaux de structuration :

La structure primaire

L'ordre de succession des acides aminés constitue la structure primaire ou séquence de la protéine. Cette séquence est donnée par convention dans le sens allant de l'extrémité N-terminale à l'extrémité C-terminale. A ces extrémités, les protéines présentent généralement les groupes chargés NH_3^+ et COO^-.

La structure secondaire

L'existence de structures secondaires vient du fait que les repliements énergétiquement favorables de la chaîne peptidique sont limités et que seules certaines conformations sont possibles. Ce type de structure est agencé de manière « locale » et stabilisé par des liaisons hydrogène principalement dans le squelette peptidique. Les éléments de structure secondaire les plus courants dans les protéines sont les hélices-α, les hélices 3_{10}, les coudes-β et les feuillets-β.

- **Les hélices α :**

Dans les chaînes peptidiques, c'est en grande majorité l'hélice-α droite (tournant dans le sens horaire) qui est présente. L'hélice a pour angles caractéristiques ϕ = -57° et ψ = -47° ; elle contient 3,6 résidus par tour ; son pas est de 5,4 Å. La structure hélicoïdale est stabilisée par des liaisons hydrogène, dirigées parallèlement à l'axe de l'hélice entre le groupement C=O de l'acide aminé en position *i* et le groupement N-H de celui en position *i*+4. Les chaînes latérales des acides aminés pointent toutes vers l'extérieur de l'hélice et vers le bas quand l'hélice est orientée verticalement, l'extrémité N-terminale pointant vers le bas.

- **Les hélices-3_{10} et les coudes-β :**

Les hélices-3_{10} sont fréquemment retrouvées à l'extrémité C-terminale d'une hélice-α. Les liaisons hydrogène stabilisantes ne sont plus parallèles à l'axe et relient les acides aminés *i* et *i*+3. L'indice *10* attribué au nom de l'hélice désigne le nombre d'atomes impliqués dans le cycle qui réunit les groupes C=O (*i*) et N-H (*i*+3) qui participent tous les deux à une même liaison hydrogène. Ce type de conformation est peu fréquent et sa longueur dépasse rarement 1 à 2 tours.

Les coudes-β sont une autre forme possible de la structure peptidique où interviennent des liaisons hydrogène entre les résidus *i* et *i*+3. Ce sont les six valeurs des angles ϕ et ψ qui les différencient des hélices-3_{10}. Les coudes-β relient deux structures secondaires répétitives (hélices ou feuillets).

- **Les feuillets-β :**

Le brin-β, répétition de plusieurs acides aminés en conformation β (ϕ = -139° à -119°, ψ = 113° à 135°), est une conformation où la chaîne peptidique est étendue. Les dipôles de la liaison peptidique sont alignés dans le même sens. Dans les feuillets-β, les brins-β peuvent être parallèles ou anti-parallèles. Dans les deux cas, il y a formation d'une structure plissée en forme d'accordéon, les chaînes latérales étant alternativement d'un côté ou de l'autre du brin. Ces structures sont stabilisées par des liaisons hydrogène entre les groupements C=O et N-H de brins adjacents.

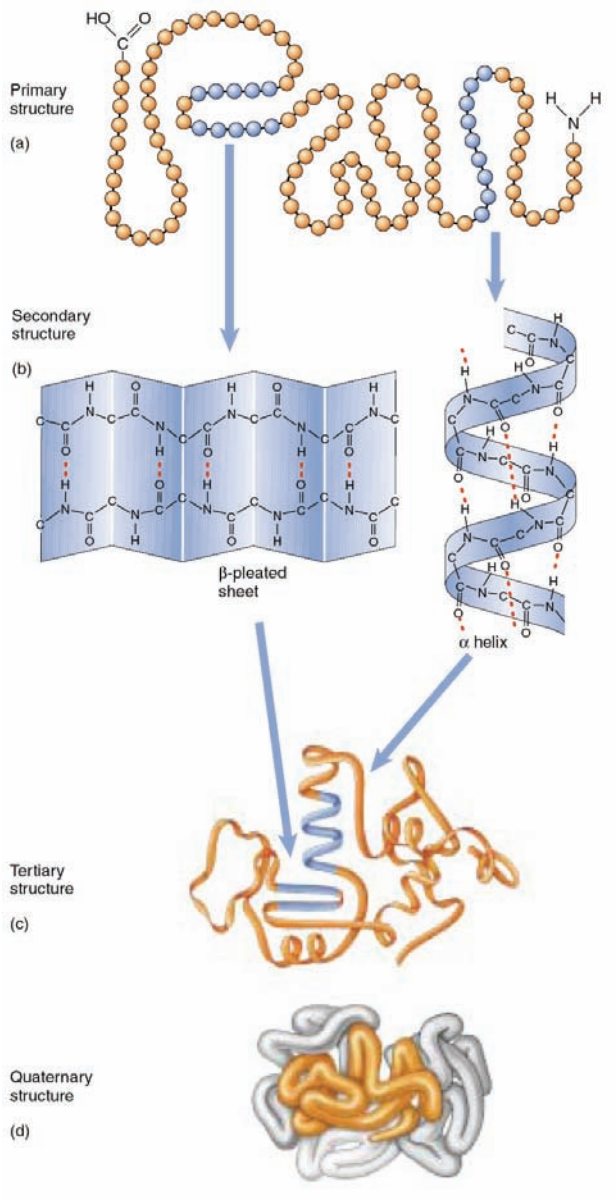
La structure tertiaire

La structure tertiaire d'une protéine est sa disposition tridimensionnelle : c'est le résultat de l'agencement des structures secondaires et de l'organisation spatiale des chaînes latérales. Un certain nombre d'interactions guident l'organisation de la structure tertiaire et la stabilisent :

- Les ponts disulfures sont des liaisons covalentes entre deux cystéines distantes dans la structure primaire mais proches dans la structure tertiaire.
- Les ponts salins se forment entre deux acides aminés ionisés de charges complémentaires.
- Les liaisons hydrogène contribuent à la stabilisation des structures tertiaires. Elles impliquent le squelette peptidique et les chaînes latérales.
- Les interactions hydrophobes ont lieu entre résidus non polaires. Ces derniers, dépourvus de groupements chargés ou d'atomes aptes à former des liaisons hydrogène, sont prédisposés à s'associer entre eux. Ils ont tendance à fuir le milieu aqueux environnant ; ce qui leur permet de se regrouper au coeur de la protéine et de réduire leur surface de contact avec le solvant (effet hydrophobe).
- Les interactions de van der Waals se manifestent entre tous les atomes de la protéine. Elles sont répulsives à courte distance (inférieure à la somme des rayons de van der Waals) et attractives à longue distance.

La structure quaternaire

La structure quaternaire résulte de l'association de deux ou plusieurs chaînes polypeptidiques (structure à plusieurs sous-unités ou monomères). Les protéines ne présentent pas toutes une structure quaternaire. Généralement, chaque polypeptide se replie plus ou moins indépendamment et les sous-unités repliées s'associent alors entre elles pour former un complexe multimérique.



Les 4 niveaux de structure d'une protéine : (a) structure primaire (séquence), (b) structure secondaire illustrée par un feuillet-β anti-parallèle et une hélice-α, (c) structure tertiaire, (d) structure quaternaire. D'après (Raven and Johnson, Biology 6th Edition)

Résumé

Les travaux présentés dans ce mémoire se situent dans le cadre général de la recherche de nouveaux médicaments par le biais de techniques informatiques.

La première partie de ce document est centrée autour du développement de la plateforme logicielle VSM-G (Virtual Screening Manager for Grids). Le but poursuivi par ce projet est de fournir un outil convivial et simple d'utilisation afin de conduire des études de criblage virtuel à haut-débit. Le cœur de VSM-G repose sur une stratégie multi-étapes de filtres successifs permettant le traitement efficace de chimiothèques de grande taille. Deux filtres ont été utilisés pour ce travail et implémentés dans VSM-G : un programme innovant d'estimation rapide de complémentarité géométrique entre molécules-candidates et site actif (SHEF) précéde un algorithme de docking flexible plus conventionnel (GOLD). Les avantages de cette méthodologie, associée à la prise en charge de multiples conformations de la cible étudiée (le récepteur nucléaire LXRβ), sont présentés tout d'abord par une étude de preuve de concept, puis à travers une campagne de criblage virtuel à grande échelle.

L'autre partie de ces travaux, exclusivement applicative, concerne l'étude du domaine FAT de la kinase d'adhérence focale FAK. FAK est une cible d'intérêt pharmaceutique particulièrement intéressante, car clairement impliquée dans divers processus de développement cancéreux. Le but de cette étude est double : il s'agit tout d'abord de mieux comprendre le mode de fonctionnement du domaine FAT de FAK à travers une étude biophysique pour en évaluer la flexibilité ; et ensuite concevoir *in silico* des petites molécules peptidomimétiques permettant de moduler son activité, ce qui pourrait limiter une progression tumorale.

Mots-clés : mise au point de nouveaux médicaments ; modélisation moléculaire ; criblage virtuel haut-débit ; dynamique moléculaire ; docking ; VSM-G ; LXRβ ; FAK ; domaine FAT ; conception de peptidomimétiques

Abstract

The work presented here deals with drug discovery by means of computational techniques.

The first part is focused around the development of the VSM-G (Virtual Screening Manager for Grids) software platform. This project aims to provide a user-friendly and easy-to-use tool for performing high throughput virtual screening experiments. The core of VSM-G is a multiple-step screening strategy in which several filters are organized sequentially as to tackle large chemical libraries efficiently. Two filters were used for this study and implemented into VSM-G: a new and fast ligand-active site geometrical complementarity estimation program (SHEF) precedes a conventional flexible docking tool (GOLD). We describe the advantages of such an approach, associated with the use of multiple target conformations for the LXRβ nuclear receptor, by presenting a proof-of-concept study. A high-throughput virtual screening campaign is then performed.

The second part of this work, exclusively applicative, deals with the study of the FAT domain of the focal adhesion kinase (FAK). FAK is an important pharmaceutical target due to its involvement in the development of various forms of cancer. The first goal is to gain knowledge regarding FAT flexibility and active state structural properties. The second objective is to design *in silico* peptidomimetic compounds targeting FAT and therefore potentially modulate FAK activity during tumour progression.

Keywords: drug design; molecular modelling; high-throughput virtual screening; molecular dynamics; docking; VSM-G; LXRβ; FAK; FAT domain; peptidomimetics design