Simulation numérique et approche orientée connaissance pour la découverte de nouvelles molécules thérapeutiques

Leo Ghemtio

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

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

Thèse

présentée pour l'obtention du titre de

Docteur de l'Université Henri Poincaré

en Chimie Informatique et Théorique

par **GHEMTIO WAFO Léo Aymar**

SIMULATION NUMERIQUE ET APPROCHE ORIENTEE CONNAISSANCE POUR LA DECOUVERTE DE NOUVELLES MOLECULES THERAPEUTIQUES

Thèse dirigée par Dr MAIGRET Bernard

Thèse présentée et soutenue à Nancy le 7 Mai 2010

Membres du Jury :

Rapporteurs :

M. Morin-Allory Luc	Professeur, Université d'Orléans, Orléans luc.morin-allory@univ-orleans.fr
M. Varnek Alexandre	Professeur, Université de Strasbourg, Strasbourg varnek@chimie.u-strasbg.fr

Examinateurs :

M. Canet Daniel	Professeur, Université H. Poincaré, Nancy daniel.canet@rmn.uhp-nancy.fr
Mme Devignes Marie-Dominique	DR CNRS, LORIA, Nancy marie-dominique.devignes@loria.fr
M. Maigret Bernard	DR CNRS, LORIA, Nancy (Directeur de thèse) bernard.maigret@loria.fr
Mme Ouwe Missi Oukem Odile	DR, CIRCB, Cameroun ooukem@yahoo.fr
M. Ritchie David	DR Chaire d'excellence, Aberdeen dave.ritchie@loria.fr

Résumé

L'innovation thérapeutique progresse traditionnellement par la combinaison du criblage expérimental et de la modélisation moléculaire. En pratique, cette dernière approche est souvent limitée par la pénurie de données expérimentales, particulièrement les informations structurales et biologiques. Aujourd'hui, la situation a complètement changé avec le séquençage à haut débit du génome humain et les avancées réalisées dans la détermination des structures tridimensionnelles des protéines. Cette détermination permet d'avoir accès à une grande quantité de données pouvant servir à la recherche de nouveaux traitements pour un grand nombre de maladies. À cet égard, les approches informatiques permettant de développer des programmes de criblage virtuel à haut débit offrent une alternative ou un complément aux méthodes expérimentales qui font gagner du temps et de l'argent dans la découverte de nouveaux traitements.

Appliqué aux grandes bases de données moléculaires, le criblage virtuel à haut débit permet de limiter le criblage expérimental en fournissant, pour chaque cible biologique visée, des molécules potentiellement intéressantes au moyen de méthodes informatiques adaptées. Cependant, la plupart de ces approches souffrent des mêmes limitations. Le coût et la durée des temps de calcul pour évaluer la fixation d'une collection de molécules à une cible, qui est considérable dans le contexte du haut débit, ainsi que la précision des résultats obtenus sont les défis les plus évidents dans le domaine. Le besoin de gérer une grande quantité de données hétérogènes est aussi particulièrement crucial.

Pour surmonter les limitations actuelles du criblage virtuel à haut débit et ainsi optimiser les premières étapes du processus de découverte de nouveaux médicaments, j'ai mis en place une méthodologie innovante permettant, d'une part, de gérer une masse importante de données hétérogènes et d'en extraire des connaissances et, d'autre part, de distribuer les calculs nécessaires sur les grilles de calcul comportant plusieurs milliers de processeurs, le tout intégré à un protocole de criblage virtuel en plusieurs étapes. L'objectif est la prise en compte, sous forme de contraintes, des connaissances sur le problème posé afin d'optimiser la précision des résultats et les coûts en termes de temps et d'argent du criblage virtuel.

Les approches méthodologiques développées ont été appliquées avec succès à une étude concernant le problème de résistance du VIH aux antiviraux, projet soutenu par la fondation Bill et Melinda Gates dans le cadre d'un projet de collaboration avec le CIRCB au Cameroun.

Title

NUMERIC SIMULATION AND KNOWLEDGE-ORIENTED APPROACH FOR THE DISCOVERY OF NEW THERAPEUTIC MOLECULES

Abstract

Therapeutic innovation has traditionally benefited from the combination of experimental screening and molecular modelling. In practice, however, the latter is often limited by the shortage of structural and biological information. Today, the situation has completely changed with the high-throughput sequencing of the human genome, and the advances realized in the three-dimensional determination of the structures of proteins. This gives access to an enormous amount of data which can be used to search for new treatments for a large number of diseases. In this respect, computational approaches have been used for

high-throughput virtual screening (HTVS) and offer an alternative or a complement to the experimental methods, which allow more time for the discovery of new treatments.

HTVS methods can be used on large molecular databases, to greatly reduce the experimental time and cost by supplying potentially interesting molecules for every desired aimed biological target. However, most of these approaches suffer the same limitations. One of these is the cost and the computing time required for estimating the binding of all the molecules from a large data bank to a target, which can be considerable in the context of the high-throughput. Also, the accuracy of the results obtained is another very evident challenge in the domain. The need to manage a large amount of heterogeneous data is also particularly crucial.

To try to surmount the current limitations of HTVS and to optimize the first stages of the drug discovery process, I set up an innovative methodology presenting two advantages. Firstly, it allows to manage an important mass of heterogeneous data and to extract knowledge from it. Secondly, it allows distributing the necessary calculations on a grid computing platform that contains several thousand of processors. The whole methodology is integrated into a multiple-step virtual screening funnel. The purpose is the consideration, in the form of constraints, of the knowledge available about the problem posed in order to optimize the accuracy of the results and the costs in terms of time and money at various stages of high-throughput virtual screening.

The methodological approaches that I developed were successfully applied to study the problem of HIV resistance to antiviral therapy. This project was supported by the Bill and Melinda Gates Foundation within the framework of a project of collaboration with the CIRCB in Cameroon.

Mots clés
Criblage virtuel à haut débit, base de données, grille de calculs, extraction de connaissances.

Keywords
Virtual high throughput screening, database, grid computing, knowledge extraction.

Intitulé et adresse du laboratoire de rattachement où la thèse a été préparée:
LORIA - Campus Scientifique - BP 239 - 54506 Vandoeuvre-lès-Nancy Cedex
Téléphone +33 3 83 59 30 00 - Télécopie +33 3 83 27 83 19 - +33 3 83 41 30 79

REMERCIEMENTS

La reconnaissance est la mémoire du cœur
« Hans Christian Andersen »

A Marietta (Fungy) et Chloé,

A ma famille, (au sens large du terme),

A ceux qui me sont chers, qui m'ont toujours encouragé et supporté,

Et à ceux et à celles, qui ont cru en moi, je leur dédie ce travail.

REMERCIEMENTS

Tant de personnes ont contribué à faire de ces trois années une période si agréable et enrichissante que je vais essayer de faire bref.

Ma reconnaissance va en premier lieu au Dr Bernard Maigret. Travailler avec lui à été est une expérience vraiment enrichissante dans la mesure où il parvient à diriger efficacement les recherches en développant un sens et un goût de l'autonomie et de la rigueur scientifique. Je pense avoir énormément appris à son contact, et je lui en suis réellement reconnaissant ; sans oublier que c'est en grande partie grâce à lui que je me suis engagé dans cette thèse! Je remercie le Dr Marie-Dominique Devignes et le Dr Malika Smaïl-Tabbone pour m'avoir co-encadré et avoir grandement contribué à mon travail de thèse. Je remercie le Dr Odile Ouwe Missi Oukem, le Dr Michel Souchet, et le Dr Appolinaire Djikeng pour leur participation et leur grande implication à mes travaux de thèse.

Je remercie ensuite les membres du jury, Pr Daniel Canet, Pr Alexande Varnek, Pr Luc Morin-Allory, Dr Marie-Dominique Devignes, Dr Bernard Maigret, Dr Odile Ouwe Missi Oukem, et le Dr David Ritchie pour s'être intéressés à ces travaux et avoir pris le temps de lire en détail mon manuscrit de thèse. Merci en particulier au Dr Odile Ouwe Missi Oukem pour avoir fait un si long voyage afin d'être présente lors de la soutenance.

Mes remerciements vont ensuite à l'ensemble de l'équipe Orpailleur et à tout le personnel du LORIA. Merci en particulier au Dr Amedeo Napoli pour m'avoir permis de réaliser ma thèse au sein de son équipe. J'ai une pensée particulière pour Yesmine Asses, Mathieu Chavent, Alexandre Beautrait, Vincent Leroux, et tous ceux qui ont partagé mon bureau ...et toutes les difficultés que j'ai pu rencontrer durant ces trois années de thèse. Une pensée particulière à Yesmine Asses, que je remercie grandement pour tout son soutient.

Mais tout cela ne serait rien sans le support constant de ma famille: un grand merci à mes parents Wafo Bernard et Kaptué Regine, mon grand-père Fotué Bernard, mon oncle Fotso Réné et à toute ma grande famille pour m'avoir toujours soutenu durant mes études ; je vous avais bien dit que ça s'arrêterait un jour!

Sans oublier bien sûr tous mes amis, avec qui j'ai partagé bon nombre de moments agréables au cours de ces trois dernières années...et bien plus encore...et ce malgré mon caractère pas toujours facile. Les lister serait bien trop long, aussi soyez sûrs, très chers amis, que je n'oublie personne.

Sommaire

Liste des abréviations

3D: tridimensionnelle

ADN: Acide désoxyribonucléique

ADME/Tox : Absorption, Distribution, Métabolisme, Excrétion et Toxicité

ARN: Acide Ribonucléique

CAS: Chemical Abstracts Service

CASP: Critical Assessment of Structure Prediction

CATS: Chemically Advanced Template Search

CATS3D: Chemically Advanced Template 3D Search

CIRCB: Centre International de Référence Chantal Biya

CEHD: Criblage Expérimental à Haut Débit

CMR: Center for Medicines Research International

CVHD: Criblage Virtuel à Haut Débit

JCVI: J. Craig Venter Institute

KDD: Knowledge Discovery from Databases

PDB : Protein Data Bank

QSAR: Quantitative Structure-Activité

RMN : Résonance Magnétique Nucléaire

SIDA: Syndrome d'Immunodéficience Acquise

SOSA: Selective Optimization of Side Activities

SCONPs: Structural Classification of Natural Products

VIH: Virus de l'Immunodéficience Humaine

VSM: Virtual Screening Manager

VSM-G: Virtual Screening Manager for Computational Grids

« La découverte est le fruit de longues méditations. »

Lavoisier

Introduction

Sommaire

I. L'innovation thérapeutique

Les découvertes thérapeutiques évoquent instantanément le souvenir de Pasteur et la première vaccination contre la rage de Joseph Meister ou celles d'Alexander Fleming dont l'oubli chanceux ouvrait en 1928 l'ère des antibiotiques. Cette réalité n'est que partielle car elle fait abstraction des équipes associées, des travaux lents et structurés des scientifiques, des échanges, des essais et des échecs qui constituent la véritable chaîne de l'innovation thérapeutique. Celle-ci n'est pas moins riche en progrès que par le passé mais elle est moins fournie en anecdotes héroïques, ceci étant certainement dû aux nouvelles méthodes de recherche qui nécessitent le maillage de multiples technologies, compétences et disciplines scientifiques. La découverte de nouveaux médicaments, de nouvelles formulations et modes de délivrance et la mise au point des analogues par modification moléculaire constituent un défi majeur pour l'industrie pharmaceutique autant aujourd'hui que par le passé. En témoignent les résistances croissantes des microorganismes pathogènes aux antibactériens, aux antifongiques et aux antiviraux, la persistance ou la résurgence de maladies infectieuses et parasitaires à l'origine de millions de morts chaque année, l'émergence de nouveaux risques infectieux, les ravages des cancers et des maladies cardiovasculaires, la montée des maladies neurodégénératives, l'absence de traitements pour des milliers de maladies plus ou moins rares. De surcroît, les effets secondaires indésirables des médicaments commercialisés sont courants, indépendamment de leur mauvaise prescription ou de leur mauvais usage, en raison notamment des interactions médicamenteuses ou parce que la toxicité du produit est jugée acceptable au regard du bénéfice apporté (thérapies du sida, de certains cancers, etc.). Or ces effets indésirables sont un frein à l'observance des traitements par les patients et donc à leur efficacité. De plus, le développement de beaucoup de médicaments est abandonné très tardivement, en phases cliniques II et III, soit au bout de cinq à onze ans de développement, pour des raisons de manque d'efficacité et de toxicité, pourtant prévisible. Tout ceci constitue autant de bonnes raisons de renforcer l'innovation thérapeutique, notamment médicamenteuse. La forte attrition (lorsque des composés sont retirés du développement) peut être reliée au défi de prédire tôt et de manière optimale la sûreté et l'efficacité au cours du processus de recherche et développement, c'est-à-dire de s'assurer que les médicaments qui vont réussir soient identifiés précocement avec un taux de certitude plus grand. Réduire l'attrition aux stades aval de développement permettra d'optimiser les coûts et d'augmenter le nombre de candidats médicaments prometteurs entrant en essais cliniques. En clair, des outils prédictifs qui viendraient affiner en amont le profil d'effets indésirables et d'efficacité d'une

nouvelle molécule renforceraient les indices cliniques, parfois jugés insuffisamment probants, et faciliteraient la décision de passer ou non à la phase suivante en fonction du rapport bénéfices apportés/risques induits Ainsi, une bonne partie des solutions pourrait provenir de l'innovation sur les outils et procédés de la recherche médicamenteuse. Ceci pourrait notamment être obtenu d'une part par l'amélioration de la gestion de connaissances (en vue de la sélection des cibles et de l'optimisation de molécules thérapeutiques), et d'autre part par l'optimisation des outils de modélisation, des bases de données, des plates-formes bioinformatiques et des méthodes de prédiction.

II. L'innovation thérapeutique à l'ère de la post-génomique

Le génome humain est maintenant séquencé et annoté.[1-4] Un des grands espoirs lié à cet accomplissement est la découverte de nouvelles protéines à fort potentiel thérapeutique. Parmi les 30 000 à 40 000 gènes humains codant pour des protéines, il a été estimé que 3 000 d'entre eux codent pour des cibles thérapeutiques : des protéines qui peuvent à la fois être liées à certaines pathologies ("cibles thérapeutiques") et être ciblées par de petites molécules ayant des propriétés caractérisant un médicament ("molécules médicaments ").[5, 6] Le nombre de cibles thérapeutiques actuellement exploitées par l'industrie pharmaceutique ne représente qu'une partie mineure de cet espace pharmacologique.[7] En effet, une étude récente synthétisant toutes les précédentes estimations fixe le nombre des cibles visées par les médicaments du marché à 500.[8] 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 indications pour mettre en œuvre des traitements pharmaceutiques individualisés. Ainsi, la discipline récente, connue sous le nom de pharmacogénomique (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.[9] 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.[10] 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 révélée 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 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.[11, 12] Le travail de cette thèse se situe dans ce dernier domaine.

II.1. Le protéome comme cible de l'innovation thérapeutique

Le matériel héréditaire est codé par les acides nucléiques qui sont le support de l'information génétique de la cellule et constituent le génome. Les acides nucléiques s'observent sous deux formes polymériques : l'acide désoxyribonucléique (ADN) et l'acide ribonucléique (ARN). L'ADN et l'ARN ont des fonctions différentes par leur structure. Dans le noyau de la cellule, le gène codant une protéine est transcrit de l'ADN en ARN. Celui-ci est exporté du noyau vers le cytoplasme où son message sera ensuite déchiffré pour synthétiser une future protéine.[13] L'ADN dans le génome humain est contenu dans 24 chromosomes distincts. Sa structure fut déterminée par Watson, Crick et Franklin en 1953.[14, 15] Quelques types d'anomalies chromosomiques majeures, y compris le manque ou des copies supplémentaires et des translocations, peuvent être détectés par examen microscopique. La plupart des changements dans l'ADN sont cependant plus subtils et exigent une analyse plus approfondie de la molécule d'ADN. Chaque chromosome contient plusieurs gènes, qui constituent les unités physiques et fonctionnelles de base de l'hérédité et codent les instructions pour la synthèse des protéines (Figure 1).[16] Les protéines sont des polymères constitués à partir des vingt acides aminés naturels.

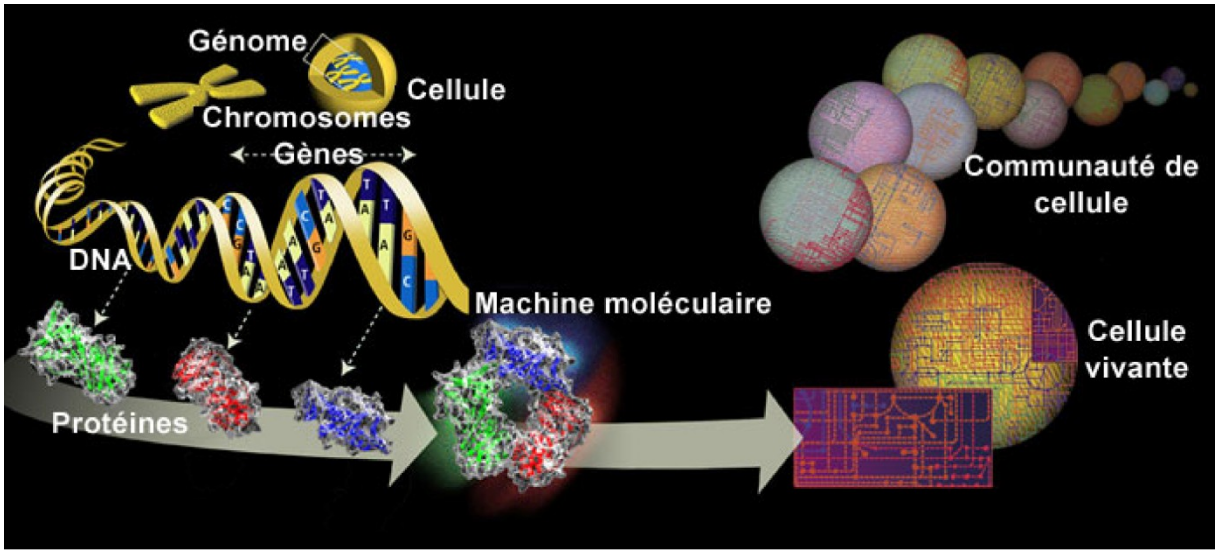


Figure 1 - De la cellule à la machinerie protéique et interactomique. Les protéines sont synthétisées à partir de l'information génétique encodée dans la cellule et elles participent au fonctionnement cellulaire seul ou en complexe. Image d'U.S.A. Department of Energy Genome Programs. http://genomics.energy.gov

Le protéome est l'ensemble des protéines exprimées dans une cellule, une partie d'une cellule (membranes, organites) ou un groupe de cellules (organe, organisme, groupe d'organismes) dans des conditions données et à un moment donné. Dans le milieu biologique où la protéine exerce son activité, elle a une structure spécifique stable avec une énergie interne minimale qui détermine sa fonction biologique. Depuis les années 1970, le repliement des protéines a été le sujet de nombreuses études. Certains principes de base de son mécanisme ont été dégagés et des résultats statistiques obtenus.[17, 18] Le but ultime de ces recherches était de pouvoir 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.[19] 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 pour l'amélioration des techniques bioinformatiques utilisées dans la prédiction du repliement des protéines.[19, 20]

II.2. L'interactome comme cible de l'innovation thérapeutique

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[21, 22] (Figure 1). 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.[23-25] On peut citer en exemple la levure Saccharomyces cerevisae, dans laquelle plus de 90 % des protéines ont été examinées et leurs interactions caractérisées, faisant d'elle le premier interactome presque entièrement identifié (Figure 2).[26] 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.[27-29]

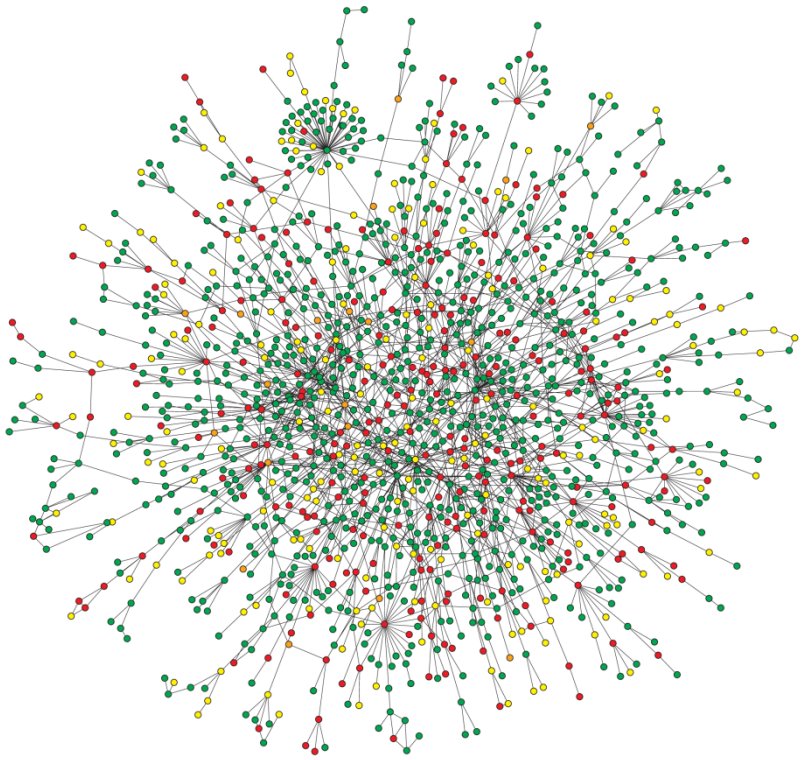


Figure 2 – Réseaux d'interactions des protéines de la levure Saccharomyces cerevisiae.[30] *Dans ce diagramme, la couleur du nœud indique l'effet sur le phénotype d'enlever la protéine correspondante (rouge = mortel, vert = non-mortel, orange = ralentit la croissance, jaune = inconnu).*

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 décrit la liste et la nature des interactions possibles qui en découlent.

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 d'un 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.[31-33] 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.[27, 29] Face aux perspectives encore vastes qui s'ouvrent à nous, le parcours actuel de la recherche thérapeutique s'inscrit dans un contexte de multiplication des connaissances interdisciplinaires et d'association de technologies de pointe appliquées à la thérapeutique humaine, afin d'explorer des pathologies complexes ou de remettre en perspective, à travers les connaissances nouvelles, la vision de maladies plus « simples ». Le processus de la recherche conjugue donc une compréhension approfondie des mécanismes pathologiques à travers la génomique, la protéomique et l'interactome, la biologie et la chimie, en s'appuyant sur l'informatique et la robotique afin de gérer et d'analyser d'innombrables données.

III. Le parcours de l'innovation thérapeutique

III.1. La recherche

III.1.1. Identification et validation des cibles thérapeutiques

Schématiquement, la recherche s'initie à partir de l'identification d'une cible thérapeutique.[34] Le mécanisme physiopathologique est disséqué au niveau le plus fin possible afin de discerner les éléments de dysfonctionnement. Longtemps, le processus de reconnaissance des cibles a reposé essentiellement sur des constats empiriques. Aujourd'hui, génomique, bioinformatique et protéomique permettent d'identifier en amont gènes ou protéines impliqués dans les maladies et susceptibles de devenir des cibles thérapeutiques.[35, 36] Enzymes, protéines, récepteurs : leur action et leur lien de causalité avec la maladie une fois connus, on cherchera soit à bloquer ou à augmenter leur action, soit à combler leur déficit, soit à remplir la fonction qu'ils ne peuvent plus exécuter. Ces connaissances permettent de rationaliser très tôt le processus de recherche de nouveaux médicaments et d'élargir considérablement le champ de cette recherche. On estime en effet couramment aujourd'hui que l'ensemble des médicaments disponibles ne ciblerait, par leur action, que 500 produits géniques.[8] Quand on sait que le génome humain comporte environ 35 000 gènes distincts et environ 200 000 protéines et que

nombre d'entre elles sont très certainement impliquées dans le développement des pathologies, on mesure pleinement l'immensité du champ exploratoire qui reste à couvrir.[5-7]

III.1.2. Identification des composés prometteurs

La cible étant validée, l'étape suivante consiste classiquement à tester l'action de dizaines de milliers, voire de millions de molécules (on parle alors de CEHD), sur cette cible.[37] Le CEHD automatisé permet de tester en parallèle par systèmes robotisés (Figure 3) un grand nombre de molécules sur une cible biologique (extraits, cellules, organismes). Pour chaque molécule de la collection, le test permettant de mesurer un effet sur une cible biologique est mis en œuvre et un signal correspondant est mesuré. C'est sur la base de ce signal qu'un choix est effectué pour retenir des molécules intéressantes. Les molécules ainsi « criblées » peuvent provenir de collections connues de la chimie traditionnelle ou sont issues de la chimie combinatoire. Les «bibliothèques» de molécules ainsi constituées varient de 1 000 à 970 000 000 composés.[38-41] Les techniques de criblage automatisé permettent seules d'analyser rapidement ces gigantesques bibliothèques.

Figure 3 – Aperçu d'une plateforme de criblage haut-débit robotisé et d'une plaque 96 puits.

Quelques « touches » (hits) sortiront de ce tri. Environ 1% des molécules criblées démontre un niveau d'activité satisfaisant. Il s'agit de composés interagissant significativement plus que la moyenne des autres composés testés sur la cible visée.[42] On vérifie ensuite si cette action correspond effectivement à l'effet recherché et si cet effet est suffisamment sélectif (s'il affecte précisément la cible). Ces molécules deviennent des « têtes de série » (leads), ce qui ne signifie pas qu'elles soient directement des « candidats médicaments ». Pour nombre d'entre elles, il conviendra d'abord d'optimiser certaines de leurs caractéristiques, dont souvent leur efficacité potentielle sur la cible. Ce processus conduit à transformer

progressivement la structure de la molécule. Mais il convient aussi de contrôler dès ce stade la capacité du composé à devenir effectivement un médicament. On vérifie alors ses qualités pharmacologiques: solubilité, passage de la barrière digestive, biodisponibilité, spécificité suffisante et absence de toxicité. L'ensemble requiert plusieurs années et les échecs sont nombreux.

III.2. Les phases précliniques et cliniques

Pour suivre le long chemin rationnel et habituel de l'innovation (Figure 4), la molécule optimisée sera testée in vivo, sur des animaux, afin de prouver que son activité thérapeutique est réelle, et les effets indésirables suffisamment limités pour pouvoir lancer le développement du désormais «candidat médicament». Le développement réel du médicament obéit ensuite, non seulement à des impératifs scientifiques rigoureux mais aussi à un cadre réglementaire extrêmement précis qui aboutit aux procédures et aux modalités d'évaluation du dossier d'autorisation de mise sur le marché. Commence alors la recherche clinique qui recouvre les phases d'étude humaine. Seul un médicament sur quinze molécules évaluées lors des différentes étapes atteindra ce stade. Ces études se déroulent en trois phases. 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. Cette phase permet d'évaluer les grandes lignes du profil de tolérance du produit et de son activité pharmacologique. La phase II se déroule en général en milieu hospitalier, sur un groupe de malades durant 2 à 3 ans. Il s'agit ici de vérifier que le rapport bénéfice/tolérance est favorable et au moins équivalent au traitement existant et qu'il n'entraîne pas des effets secondaires importants. La dose optimale, c'est-à-dire celle pour laquelle l'effet thérapeutique est le meilleur pour le moins d'effets secondaires, est établie. La phase III est la phase réelle d'essai thérapeutique. Elle est conduite chez les patients atteints de la maladie à traiter. Les règles méthodologiques sont très précises et l'essai doit être mené en comparaison entre deux groupes, en « double aveugle », l'un sous traitement, l'autre sous placebo, afin de diminuer la part de subjectivité de l'évaluation. Les essais peuvent concerner plusieurs centaines à plusieurs milliers de patients. Durant cette dernière étape, la forme galénique définitive est mise au point et les études d'efficacité thérapeutique seront complétées par celles nécessaires à la qualité pharmaceutique du produit. Une autorisation de mise sur le marché est délivrée en cas de succès et le suivi post-commercialisation constitue la phase IV.

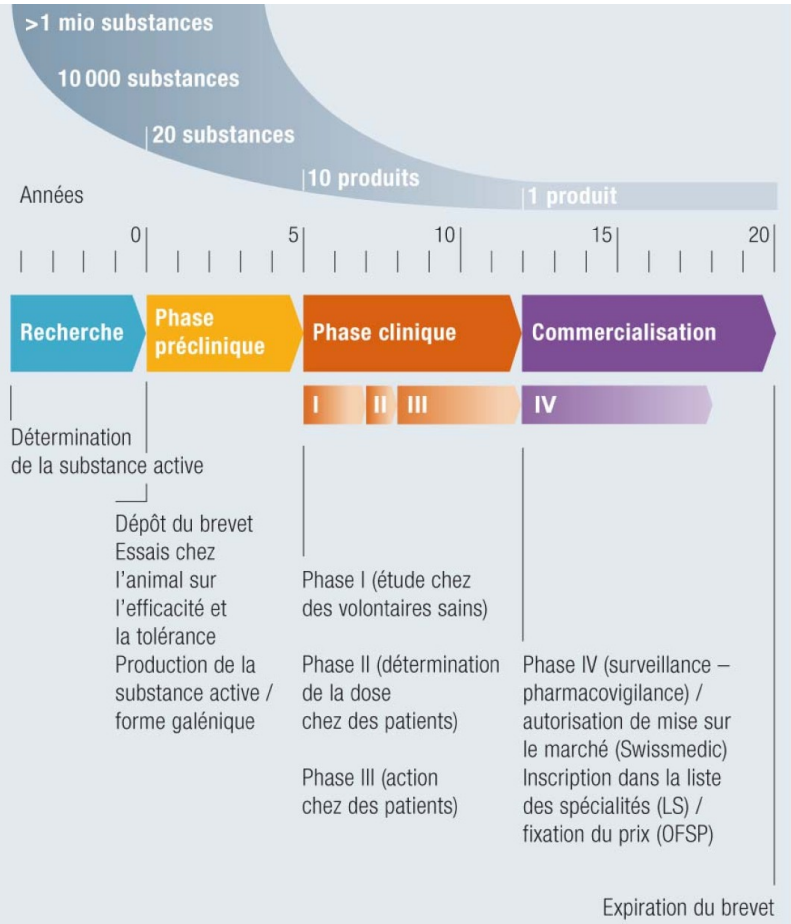


Figure 4 – La genèse d'un médicament. Les différentes étapes de découverte et commercialisation d'un nouveau médicament. Image Interpharma. Source Interpharma. http://www.interpharma.ch/de/index.asp

IV. Le coût de l'innovation thérapeutique

On mesure amplement à la vue de ce parcours de la molécule, de l'hypothèse scientifique à la disponibilité pour le malade, combien le chemin de l'innovation thérapeutique est long, en moyenne une douzaine d'années, et mobilisateur de ressources financières : entre 800 et 1 400 millions de dollars pour un médicament en 2006, d'après le Tufts Center for the Study of Drug Development (Figure 5).* L'innovation thérapeutique présente donc à la fois un coût très élevé et un risque financier majeur. Les grandes entreprises du médicament y consacrent environ 18% de leur chiffre d'affaires. Le temps nécessaire à la genèse d'un médicament mobilise d'importants capitaux sur une longue période pour un résultat éminemment aléatoire. Ce coût avait été estimé à 318 millions de dollars dans les années 80 et à 138 millions dans les années 70 (Figure 5). En dix ans, l'investissement nécessaire à la mise au point d'une nouvelle entité moléculaire a doublé. Ces coûts élevés en recherche et développement peuvent être attribués à une variété de facteurs, tels que l'utilisation croissante dans l'industrie

**:http://csdd.tufts.edu/*

pharmaceutique de technologies de découverte de nouveaux médicaments qui sont coûteux, comme le criblage à haut débit, la chimie combinatoire et pharmaco-génomique. La chimie combinatoire et le criblage à haut débit ont représenté plus de la moitié des dépenses que les entreprises ont consacrées à la recherche « amont » selon une étude du CMR* (Center for Medicines Research International) auprès des dix-sept premières entreprises de ce secteur. Ceci est certainement dû au fait que, aujourd'hui, la masse de données disponibles grâce au séquençage du génome humain, à la résolution des structures tridimensionnelles des protéines par cristallographie à rayon X ou résonance magnétique nucléaire et à l'émergence de larges bases de données de molécules rendent cette approche très coûteuse en temps et en argent.

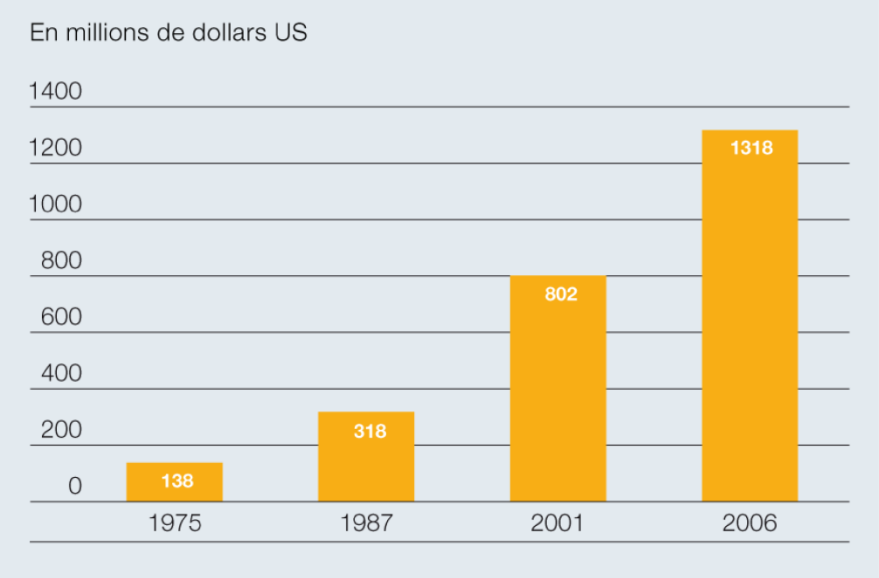


Figure 5 – Les coûts de développement d'un nouveau médicament. Représentation des coûts de découverte d'un nouveau médicament au fil des années. Image Interpharma. Source: Source: Tufts CSDD, Boston, Etats-Unis, 2003 et 2007. http://www.interpharma.ch/de/index.asp

V. Le rôle des méthodes informatiques dans l'innovation thérapeutique

Ainsi, pour des raisons économiques (liées au coût très élevé du criblage à haut débit ou de la chimie combinatoire souvent impossible dans le milieu de la recherche et dans les petites compagnies) et, pour des raisons scientifiques (liées à l'augmentation croissante du nombre de cibles, et notamment les cibles 3D, à la taille des bibliothèques chimique, à l'augmentation de la puissance de calcul des ordinateurs, à une meilleure prédiction des interactions protéine-ligand et aux progrès réalisés en informatique), une nouvelle méthodologie présentant un réel gain en temps et en argent, le CVHD, a été développée.[43, 44] Le CVHD est une technique informatique utilisée en recherche dans le domaine de la conception des médicaments. Il consiste en un parcours test sur de larges librairies de molécules chimiques pour en

*:*http://www.cmr.org/*

sélectionner celles qui ont le plus de chances de se transformer en médicaments. Le CVHD est le fruit des avancées scientifiques dans les domaines de la modélisation moléculaire, la chimie combinatoire et la biologie moléculaire. De nos jours, des millions de molécules doivent être testées en une courte période de temps, d'où la nécessité d'avoir des méthodes *in silico* pour faire un criblage rapide et efficace. L'ordinateur se substitue alors à une partie de l'expérimentation pharmacologique. Ces technologies de modélisation peuvent aussi contribuer à développer de nouvelles approches dans la recherche de molécules efficaces sur une cible donnée. Elles permettent de restreindre l'espace chimique des molécules intéressantes pour une maladie ou une cible thérapeutique et de se focaliser sur les molécules ayant le plus de chances d'avoir de bons résultats aux tests expérimentaux ou d'aboutir à d'éventuels médicaments.[45-48] Le CVHD a pris de plus en plus d'importance et constitue un réel apport permettant d'accélérer le processus de découverte de nouvelles molécules d'intérêt thérapeutique. On part alors de la structure de la cible et l'on tente de mettre au point le modèle de molécule susceptible d'interagir avec elle. Les produits les plus proches sont ensuite synthétisés et testés. C'est ainsi qu'ont été mises au point les anti-protéases actives contre le VIH et d'autres cibles biologiques.[49-54] Toutes les méthodes de CVHD sont généralement utilisées dans les premières phases du développement de molécules têtes de série à fort potentiel thérapeutique, afin de le rendre plus efficace et moins coûteux en temps et en argent dans le processus de découverte de nouveaux médicaments. Les méthodes *in silico* sont aujourd'hui bien insérées dans le processus de processus de découverte de nouveaux traitements dans l'industrie pharmaceutique (Figure 6).[55]

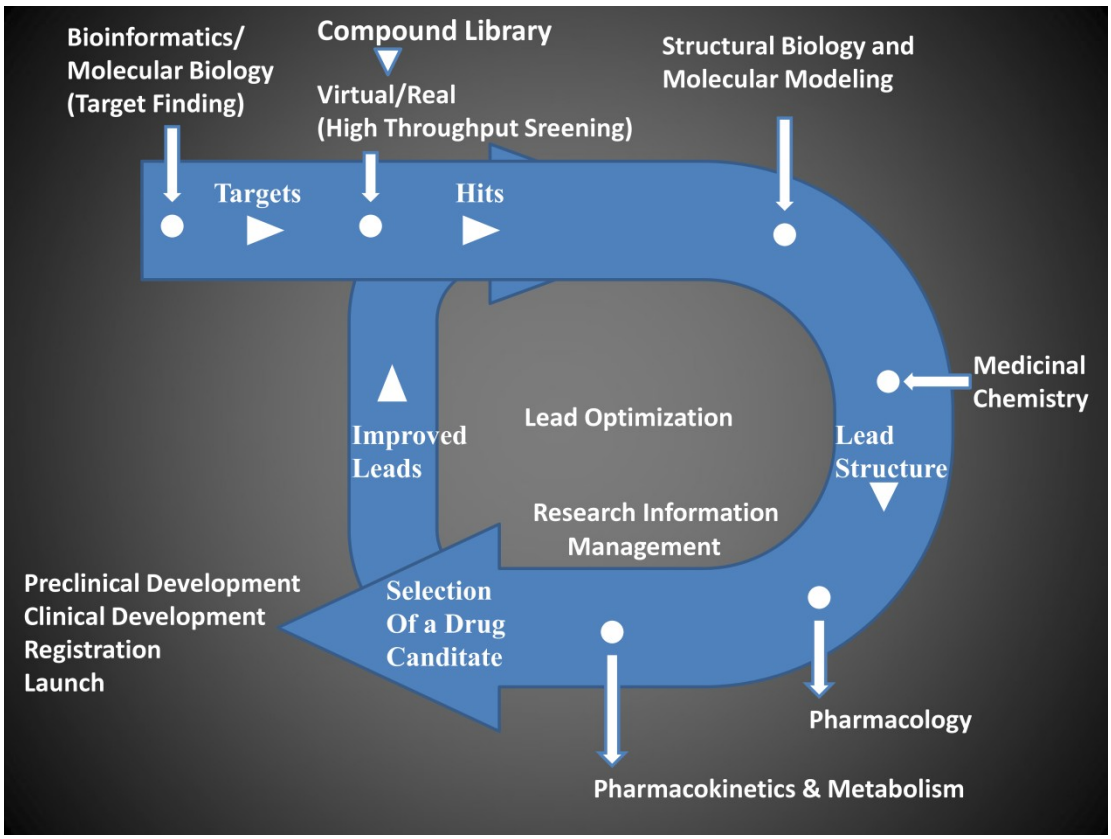


Figure 6 – Pipeline de développement d'un médicament sur lequel sont indiqués les endroits ou les méthodes in silico peuvent intervenir.

Compte tenu de son parcours, un nouveau médicament ne bénéficie en moyenne que d'une dizaine d'années de protection commerciale effective permettant de le rentabiliser. Ainsi, la réduction du temps de mise sur le marché est un challenge perpétuel dans les entreprises pharmaceutiques, étant donné que la durée des phases cliniques et d'obtention des autorisations est difficilement compressible du fait des réglementations en vigueur. Il devient donc indispensable d'optimiser les premières étapes de découverte de nouveaux médicaments. Ceci permettra d'optimiser le coût et les prédictions des méthodes de CVHD.

VI. Présentation des travaux de recherche

Cette thèse présente les principaux résultats issus de la thèse que j'ai effectuée sous la direction de Bernard Maigret, au LORIA (UMR 7503, équipe ORPAILLEUR). Ce travail rassemble des développements méthodologiques et applicatifs. Le point commun de ces deux aspects est la recherche de nouveaux médicaments par le biais de techniques informatiques. Ce travail de thèse m'a permis de développer différents protocoles de CVHD permettant d'optimiser le temps et le coût des différentes approches de CVHD, notamment grâce aux avancées réalisées dans la vitesse des processeurs, la parallélisation des calculs scientifiques,

les grilles de calcul, afin de diminuer le coût en temps de calcul, ainsi que la prise en compte des connaissances extraites des bases de données et des descripteurs physico-chimiques des cibles et des ligands, ceci, afin de mieux prédire l'affinité entre un ligand et une cible.

Le premier chapitre de cette thèse présente un état de l'art des principales méthodes de CVHD qui sont aujourd'hui les plus populaires dans la recherche de nouveaux composés d'intérêt thérapeutique: aussi bien les méthodes basées sur la structure tridimensionnelle des protéines que celles basées uniquement sur la structure des ligands si celle de la protéine cible n'est pas connue. Il résume l'état de l'art des principales méthodes de CVHD et présente aussi les avantages et inconvénients de chaque approche. Il présente un exemple de plateforme logicielle de CVHD (VSM-G) à la mise en place de laquelle j'ai participé. VSM-G repose sur plusieurs concepts novateurs dont la pertinence a été validée sur le plan scientifique en parallèle à son implémentation. Le projet VSM-G s'inscrit dans un contexte de multiples collaborations dont les principales ont impliqué, d'une part l'équipe de Wensheng Cai (Université de Nankai, R.P.Chine), Peter Bladon (Interprobe, Royaume-Uni) et Gilles Moreau pour l'aspect développement, d'autre part Michel Souchet et Sinan Karaboga (Fournier Pharma / Solvay, Daix) pour la campagne d'application/validation.

Le chapitre suivant détaille la méthodologie développée afin d'améliorer les performances en temps de calcul de VSM-G par l'utilisation des grilles de calcul. VSM-G est appliqué au criblage virtuel de millions de molécules. L'émergence des grilles de calcul constitue donc une réelle alternative permettant de distribuer de façon séquentielle ou parallèle l'ensemble des calculs nécessaires au criblage virtuel sur des milliers de processeurs répartis sur des sites géographiques différents de façon à diminuer considérablement le temps nécessaire au CVHD. Celui-ci s'effectue traditionnellement sur un nombre restreint de processeurs d'un ou plusieurs ordinateurs connectés à un réseau local. Ce chapitre a fait l'objet d'une publication. Dans le cadre de ce projet, j'ai collaboré avec Emmanuel Jeannot de l'équipe Algorille du Loria.

Le chapitre trois a fait l'objet de deux papiers avec deux communications orale à deux conférences internationale avec comité de lecture. Il présente la mise en place d'une base de données sur les protéines, les ligands et les interactions protéines-ligands, afin d'en extraire des connaissances permettant de caractériser les ligands actifs sur une cible donnée. Ainsi, deux algorithmes de fouille de données, les règles d'association dans la première publication et les arbres de décision dans la seconde publication, sont utilisés sur les données stockées

dans la base de données afin de déterminer les descripteurs moléculaires et physico-chimiques qui caractérisent les ligands actifs ou inactifs sur une cible biologique donnée. Ce travail a été réalisé avec le concours de Malika Smaïl-Tabbone et Marie-Dominique Devignes de l'équipe Orpailleur du Loria au sein de laquelle j'ai effectué ma thèse.

Le chapitre 4, qui a fait l'objet d'une publication, est une application de l'utilisation de la connaissance obtenue au chapitre 3 sur les descripteurs de ligand dans un protocole de filtrage de larges bases de données de molécules et d'identification de nouvelles molécules thérapeutiques. Ce chapitre décrit aussi la comparaison de cette méthodologie avec d'autres approches plus courantes et déjà bien documentées.

Le chapitre 5 décrit l'adaptation de la base de données décrite au chapitre 3 à la problématique du traitement des résistances du VIH dans le cadre du projet GATES. Ce projet consiste à la mise en place d'une plate forme bioinformatique, couplée à une base de données sur les protéines du VIH, leur ligand, les mutations, les résistances et les données cliniques des patients qui permettent de surmonter les résistances aux antirétroviraux chez les malades du VIH. Ce chapitre a fait l'objet d'une publication. Le projet est financé par la fondation Bill-et-Melinda-Gates et réalisé en partenariat avec le Dr Ouwe Missi Oukem Odile du Centre International de Référence Chantal Biya (CIRCB) pour la Recherche sur la Prévention et la Prise en charge du VIH/SIDA (Cameroun), et le Dr Apolinaire Djikeng de l'institut J. Craig Venter (JCVI, Maryland, USA).

Le chapitre 6 est une conclusion générale sur les différents travaux réalisés au cours de ma thèse et présente les principales perspectives qui découlent de mon travail.

VII. Bibliographie

1. IHGSC. Finishing the euchromatic sequence of the human genome. *Nature.* 2004;431:931-45.
2. Lander ES, Linton LM, Birren Bet al. Initial sequencing and analysis of the human genome. *Nature.* 2001;409:860-921.
3. Levy S, Sutton G, Ng PCet al. The diploid genome sequence of an individual human. *PLoS Biol.* 2007;5:e254.
4. Venter JC, Adams MD, Myers EWet al. The sequence of the human genome. *Science.* 2001;291:1304-51.
5. Hopkins AL, Groom CR. The druggable genome. *Nat Rev Drug Discov.* 2002;1:727-30.
6. Russ AP, Lampel S. The druggable genome: an update. *Drug Discov Today.* 2005;10:1607-10.
7. Paolini GV, Shapland RH, van Hoorn WP, Mason JS, Hopkins AL. Global mapping of pharmacological space. *Nat Biotechnol.* 2006;24:805-15.
8. Overington JP, Al-Lazikani B, Hopkins AL. How many drug targets are there? *Nat Rev Drug Discov.* 2006;5:993-6.
9. Weinshilboum RM, Wang L. Pharmacogenetics and pharmacogenomics: development, science, and translation. *Annu Rev Genomics Hum Genet.* 2006;7:223-45.
10. Loriot MA, Beaune P. [Pharmacogenomics: the link between genes and response to drugs]. *Med Sci (Paris).* 2004;20:634-6.
11. Verma IM, Weitzman MD. Gene therapy: twenty-first century medicine. *Annu Rev Biochem.* 2005;74:711-38.
12. Sugaya N, Ikeda K, Tashiro Tet al. An integrative in silico approach for discovering candidates for drug-targetable protein-protein interactions in interactome data. *BMC Pharmacol.* 2007;7:10.
13. Yanofsky C. Gene structure and protein structure. *Harvey Lect.* 1967;61:145-68.
14. Watson JD, Crick FH. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature.* 1953;171:737-8.
15. Watson JD, Crick FH. Genetical implications of the structure of deoxyribonucleic acid. *Nature.* 1953;171:964-7.
16. Crick FH, Barnett L, Brenner S, Watts-Tobin RJ. General nature of the genetic code for proteins. *Nature.* 1961;192:1227-32.
17. Anfinsen CB. Principles that govern the folding of protein chains. *Science.* 1973;181:223-30.
18. Onuchic JN, Wolynes PG. Theory of protein folding. *Curr Opin Struct Biol.* 2004;14:70-5.
19. Dill KA, Ozkan SB, Weikl TR, Chodera JD, Voelz VA. The protein folding problem: when will it be solved? *Curr Opin Struct Biol.* 2007;17:342-6.
20. Moult J. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr Opin Struct Biol.* 2005;15:285-9.
21. Parrish JR, Gulyas KD, Finley RL, Jr. Yeast two-hybrid contributions to interactome mapping. *Curr Opin Biotechnol.* 2006;17:387-93.
22. Uetz P, Finley RL, Jr. From protein networks to biological systems. *FEBS Lett.* 2005;579:1821-7.
23. Devos D, Russell RB. A more complete, complexed and structured interactome. *Curr Opin Struct Biol.* 2007;17:370-7.
24. Gavin AC, Bosche M, Krause Ret al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature.* 2002;415:141-7.

25. Ho Y, Gruhler A, Heilbut Aet al. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature.* 2002;415:180-3.
26. Krogan NJ, Cagney G, Yu Het al. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature.* 2006;440:637-43.
27. Gandhi TK, Zhong J, Mathivanan Set al. Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nat Genet.* 2006;38:285-93.
28. Kiemer L, Cesareni G. Comparative interactomics: comparing apples and pears? *Trends Biotechnol.* 2007;25:448-54.
29. Stelzl U, Worm U, Lalowski Met al. A human protein-protein interaction network: a resource for annotating the proteome. *Cell.* 2005;122:957-68.
30. Jeong H, Mason SP, Barabasi AL, Oltvai ZN. Lethality and centrality in protein networks. *Nature.* 2001;411:41-2.
31. Berman HM, Westbrook J, Feng Zet al. The Protein Data Bank. *Nucleic Acids Res.* 2000;28:235-42.
32. Liu HL, Hsu JP. Recent developments in structural proteomics for protein structure determination. *Proteomics.* 2005;5:2056-68.
33. Stewart L, Clark R, Behnke C. High-throughput crystallization and structure determination in drug discovery. *Drug Discov Today.* 2002;7:187-96.
34. Lindsay MA. Target discovery. *Nat Rev Drug Discov.* 2003;2:831-8.
35. Kunkel EJ. Systems biology in drug discovery. *Conf Proc IEEE Eng Med Biol Soc.* 2006;1:37.
36. Lindsay MA. Finding new drug targets in the 21st century. *Drug Discov Today.* 2005;10:1683-7.
37. Bajorath J. Integration of virtual and high-throughput screening. *Nat Rev Drug Discov.* 2002;1:882-94.
38. Blum LC, Reymond JL. 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. *J Am Chem Soc.* 2009;131:8732-3.
39. Cheeseright TJ, Mackey MD, Melville JL, Vinter JG. FieldScreen: virtual screening using molecular fields. Application to the DUD data set. *J Chem Inf Model.* 2008;48:2108-17.
40. Irwin JJ, Shoichet BK. ZINC--a free database of commercially available compounds for virtual screening. *J Chem Inf Model.* 2005;45:177-82.
41. Wishart DS, Knox C, Guo ACet al. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* 2008;36:D901-6.
42. Bleicher KH, Bohm HJ, Muller K, Alanine AI. Hit and lead generation: beyond high-throughput screening. *Nat Rev Drug Discov.* 2003;2:369-78.
43. Hou T, Xu X. Recent development and application of virtual screening in drug discovery: an overview. *Curr Pharm Des.* 2004;10:1011-33.
44. Shoichet BK. Virtual screening of chemical libraries. *Nature.* 2004;432:862-5.
45. Dobson CM. Chemical space and biology. *Nature.* 2004;432:824-8.
46. Lipinski C, Hopkins A. Navigating chemical space for biology and medicine. *Nature.* 2004;432:855-61.
47. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev.* 2001;46:3-26.
48. Rigby AC. Exploring novel chemical space through the use of computational and structural biology. *Comb Chem High Throughput Screen.* 2009;12:927-8.
49. 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.
50. Alvarez JC. High-throughput docking as a source of novel drug leads. *Curr Opin Chem Biol.* 2004;8:365-70.

51. Daneshtalab M. Discovery of chlorogenic acid-based peptidomimetics as a novel class of antifungals. A success story in rational drug design. *J Pharm Pharm Sci.* 2008;11:44s-55s.
52. Ghosh S, Nie A, An J, Huang Z. Structure-based virtual screening of chemical libraries for drug discovery. *Curr Opin Chem Biol.* 2006;10:194-202.
53. Gruneberg S, Stubbs MT, Klebe G. Successful virtual screening for novel inhibitors of human carbonic anhydrase: strategy and experimental confirmation. *J Med Chem.* 2002;45:3588-602.
54. 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;46:2656-62.
55. Jorgensen WL. The many roles of computation in drug discovery. *Science.* 2004;303:1813-8.

« […] [I]nformatics is a real aid to discovery when analyzing biological functions […]. […] I was convinced of the potential of the computational approach, which I called in silico, to underline its importance as a complement to in vivo and in vitro experimentation. »

Antoine Danchin

CHAPITRE 1 - Le criblage virtuel à haut débit (CVHD) : But, champ d'application et état de l'art

Dans ce chapitre, nous allons présenter les concepts et l'historique des techniques de criblage virtuel et de haut débit. Nous montrerons leurs applications à la recherche dans des banques de données moléculaires permettant la sélection de molécules, à l'analyse de la similarité et de la diversité des molécules. Nous finirons par la description d'un entonnoir de CVHD multi étapes VSM, qui permet d'optimiser le coût du CVHD en combinant de façon séquentielle différentes méthodes de CVHD.

Sommaire

I. Introduction

Les progrès incessants de la chimie, de la génomique, de la robotique, de l'informatique et de bien d'autres disciplines scientifiques ont permis depuis les années 1990 d'établir le CEHD comme la principale méthode en chimie médicinale de sélection de nouvelles molécules têtes de série ou candidats médicaments. Mais, en raison du coût élevé de la synthèse et du criblage d'une très grande quantité de composés, il apparaît un besoin croissant en outils efficaces pour concevoir et classifier de vastes librairies chimiques, afin d'accroître le contenu de leurs informations. Le CVHD se définit comme un procédé *in silico* permettant de filtrer et de donner un score à des molécules d'une bibliothèque en fonction de leur affinité prédite avec une cible biologique. Ainsi, il permet de prédire l'activité des petites molécules organiques présentes dans des grandes banques de données publiques ou privées et de focaliser l'approche expérimentale sur les molécules candidates les plus prometteuses, alors que le CEHD nécessite une plate forme technologique performante et que les informations sur la structure 3D de la cible ne sont pas indispensables. Le CVHD, en plus des informations sur les structures 3D de la cible, nécessite d'autres informations sur les propriétés physicochimiques et géométriques des composés actifs. En fonction de la disponibilité ou non de ces informations, on a différentes méthodes de CVHD. Les méthodes basées sur la structure (structure-based) sont utilisées lorsque les données sur les structures 3D de la cible biologique sont disponibles. La forte progression des structures 3D des macromolécules disponibles a provoqué un fort développement des méthodes basées sur la structure. Les méthodes basées sur les ligands (ligand-based) sont utilisées lorsque seules les informations émanant des composés actifs sur la cible biologique sont disponibles et non les informations sur la structure 3D de la cible biologique. Lorsque les informations aussi bien sur les composés actifs sur la cible biologique que celles sur les structures 3D de la cible biologique sont disponibles, il est possible d'utiliser les deux méthodes séparément afin de comparer leur efficacité ou bien de les utiliser simultanément afin de tirer le meilleur des deux méthodes. De nouvelles méthodes de CVHD commencent à émerger. Elles intègrent le criblage anti-cible associé à la pharmacocinétique (absorption, distribution, métabolisme, excrétion) et la toxicité (ADME/Tox) afin de réduire les complications possibles dans les étapes suivantes du processus de développement de nouveaux médicaments.

II. L'univers chimique

L'univers de l'espace chimique « possible » est très large : l'espace virtuel de petites molécules (Poids moléculaire < 500) qui peuvent en principe être créées est estimé à 10^{60}, très loin des capacités des synthèses par chimie combinatoire les plus ambitieuses.[1] Par contre, l'espace chimique « actuel » tel que décrit dans la littérature est bien plus limité : par exemple, le registre CAS (Chemical Abstract Service) qui est une des plus larges ressources de composés contient un peu plus de 52 millions de petites molécules organiques et inorganiques, ce qui est très loin des quantités théoriquement disponibles. Seule une extrêmement petite fraction de cet espace chimique « possible » a jusqu'à présent été explorée dans le cadre de recherche de nouveaux médicaments (il est certain que des contraintes de synthèse y ont fortement contribué). Ceci peut expliquer en partie la faible quantité de nouvelles molécules introduites sur le marché ces dernières années (17 nouvelles molécules en 2007 avec des coûts qui continuent à croître de façon non proportionnelle).[2, 3] Cependant, plusieurs approches sont actuellement explorées afin d'explorer un espace chimique plus large permettant de découvrir de nouvelles molécules à visées thérapeutiques. Ces approches incluent l'utilisation d'entités thérapeutiques déjà validées ou de produits naturels récemment isolés dont l'utilisation peut fournir des fragments moléculaires qui seront ensuite combinés pour explorer un espace chimique plus large et augmenter les chances de produire une molécule active. Ainsi, lors de la recherche de nouvelles entités thérapeutiques, il est indispensable d'augmenter la diversité moléculaire des composés disponibles en essayant de couvrir l'espace chimique le plus large possible.

II.1. Exploration de l'espace chimique afin d'identifier de nouvelles entités thérapeutiques

II.1.1. La chimie combinatoire

La chimie combinatoire (réelle ou virtuelle) est apparue naturellement comme une option viable au problème de la diversité moléculaire. Aujourd'hui, c'est un moyen pratique pour prédire et synthétiser une grande quantité de molécules en chimie pharmaceutique et agrochimique[4-6]. Comme moteur de diversité, cet outil est devenu indispensable et a joué un rôle important dans le progrès de la synthèse automatique et parallèle survenu ces vingt dernières années. Cette méthode repose sur l'idée d'obtenir le plus grand nombre de produits possible d'une réaction particulière et ceci sous certaines conditions[7, 8] (la Figure I.1 pour plus

d'exemples). Comme le mot l'indique, ces possibilités dites «combinatoires» ne sont pas infinies mais très nombreuses, d'où le problème du traitement (réel ou virtuel) de ces molécules. Aux données combinatoires s'ajoutent de nouvelles molécules, issues des synthèses, des extractions et d'autres procédés chimiques, dans les bases de données chimiques à caractère académique ou industriel. Ainsi, chaque année, le CAS voit sa base de molécules chimiques augmenter de millions de nouveaux composants. Les structures, les propriétés physicochimiques et biologiques de ces molécules sont ensuite codées et enregistrées, générant plus d'informations. L'organisation, l'analyse, la recherche et la gestion de cette grande quantité d'informations ouvrent de nouvelles possibilités aux techniques novatrices de chimie informatique, parmi lesquelles on compte le CEHD ou CVHD, la fouille de données (data-mining), etc.

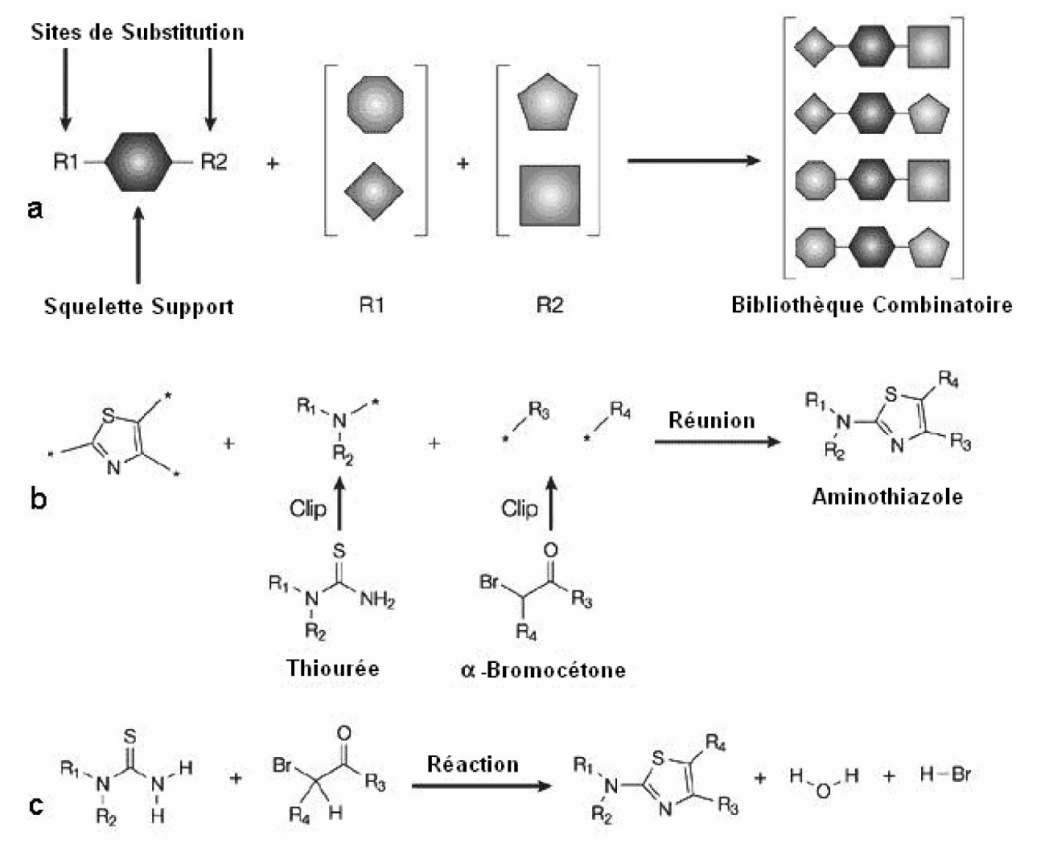


Figure I.1 - Génération d'une bibliothèque virtuelle, où deux approches sont couramment utilisées: (a) La première est basée sur les structures de Markush. (b) La deuxième consiste à attacher systématiquement les réactifs aux sites actifs. (c) Dans une variation de la deuxième approche, des parties spécifiques des réactants sont spécifiées ainsi que la nature des réactions possibles.

II.1.2. Les Chemins aléatoires (Random walks)

Cette méthode permet d'explorer la diversité moléculaire en utilisant la génération aléatoire des structures moléculaires à partir d'un sous-ensemble constitué de molécules ayant un intérêt thérapeutique précis. Étant donné les dimensions de l'espace chimique des petites molécules, la probabilité de recherche fructueuse par les chemins purement aléatoires est faible, ce qui explique l'échec des premières mises en œuvre de la chimie combinatoire qui n'incluaient pas les considérations sur « la qualité » des médicaments.[9] De plus en plus, l'attention est portée sur les bibliothèques focalisées, basées soit sur des échafaudages moléculaires productifs (médicaments existant et structures privilégiées), soit sur l'imitation de la chimie combinatoire de la nature par des méthodes basées sur la diversité.[10]

II.1.3. Constructions à partir des structures existantes

La base la plus fructueuse pour la découverte d'un nouveau médicament est de partir d'un médicament déjà connu. Clairement, il existe beaucoup d'exemples de médicaments qui sont des variations d'anciens médicaments : par exemple les β-bloquants, les antidépresseurs tricycliques, le 1,4-dihydropyridine bloquant des canaux de calcium, etc..[11] Parfois des médicaments sont produits en réponse à la concurrence ou pour des besoins économiques. Cette approche produit fréquemment des molécules très similaires mais avec des propriétés pharmacocinétiques et pharmacodynamiques différentes. De plus, puisque ces molécules sont déjà connues dans l'occupation d'un espace pharmacologique validé, elles peuvent être employées comme « têtes de série » pour de nouvelles indications thérapeutiques. L'approche basée sur l'optimisation sélective des effets secondaires (SOSA) permet de cribler une bibliothèque de médicaments existants (avec des valeurs de toxicité et biodisponibilité connue chez l'homme) sur de nouvelles cibles. L'optimisation structurale subséquente permet de convertir l'effet secondaire pour un médicament existant en une activité principale d'une nouvelle entité thérapeutique.[12]

II.1.4. L'espace chimique des produits naturels

Historiquement, les produits naturels et leurs dérivés ont été une source majeure d'agents thérapeutiques. Presque 50 % des nouvelles entités chimiques présentées dans les années 1980 et les années 1990 ont été tirées directement ou indirectement des structures de produits naturels.[13] Il y a maintenant un regain d'intérêt de la biosynthèse des squelettes de produit naturels et de l'amélioration des méthodes synthétiques qui peuvent plus aisément produire

des structures complexes issues des produits naturels.[14, 15] Traditionnellement, les structures complexes des substances naturelles sont traitées sur la base "une molécule à la fois". Des méthodes synthétiques plus récentes s'efforcent, grâce à la synthèse basée sur la diversité, de permettre, à partir d'un bloc chimique simple au départ, la génération d'une grande collection d'entités diverses et complexes. Pour faciliter l'exploration d'un espace chimique de composés basés sur les produits naturels, Waldmann et ses collaborateurs ont fourni une classification structurale des produits naturels (SCONPs) basée sur les squelettes sous-jacents présents dans ces produits naturels.[16] Cette organisation hiérarchique basée sur le squelette structural des composés naturels peut fournir des conseils sur la sélection de motifs moléculaires spécifiques utilisés dans le développement de bibliothèques chimiques basées sur les composés naturels.

II.1.5. Assemblage des fragments

La découverte de nouveaux médicaments basée sur les fragments repose sur le concept de base que la complémentarité moléculaire est plus facilement et efficacement explorée avec les petits fragments moléculaires de taille allant jusqu'à 12 atomes, puisque le nombre de tels fragments, ~10^7, est beaucoup plus petit que le grand nombre de molécules semblables au médicament qui est supérieur à 10^{60}. Des exemples récents d'application de cette technique ont été détaillés par Congreve et al.[17] La synthèse guidée par un modèle peut être considérée comme une extension de l'assemblage par fragments : par exemple, la biosynthèse peut être utilisée comme modèle d'assemblage de fragments moléculaires afin de conduire à la formation des liaisons covalentes entre fragments et former une molécule active.[18] Un excellent exemple de succès de cette approche est la synthèse de l'inhibiteur de l'acétylcholine estérase avec une affinité femto molaire à partir des motifs de la tacrine et du phenanthridinium.[19]

II.2. Les collections de ligands et de cibles pour le criblage

II.2.1. Les collections de ligands

Aujourd'hui, il existe une pléthore de bases de données et de bibliothèques de molécules chimiques provenant, soit de la chimie combinatoire, ou des substances naturelles, soit des molécules, déjà connues comme intéressantes sur certaines cibles biologiques, et qui peuvent être potentiellement des médicaments « drug-like ». Tout cela contribue à élargir l'espace chimique virtuel. Une exploration systématique d'une petite partie de cet espace avec des

molécules ayant jusqu'à onze atomes lourds a été récemment réalisée.[20] Après l'exclusion des molécules peu convenables, plus de 13 millions de composés différents sont sélectionnés. Une molécule de médicament typique peut être jusqu'à deux fois aussi grande que les composés examinés dans cette étude (masse moyenne de 340 Daltons, environ 24 atomes lourds).[21] Le nombre de molécules « drug-like » accessibles aux procédures de synthèse actuelles est de l'ordre de 10^{60} à 10^{100}.[22] Ces nombres indiquent que nous couvrons une fraction presque négligeable de l'espace chimique virtuel. Les composés pour le criblage peuvent être obtenus à partir des bases de données de structures connues, de bibliothèques combinatoires ou des programmes de conception de novo. En raison des problèmes de synthèse, on considère souvent seulement les structures connues. Des bases de données typiques avec des composés organiques de laboratoire par exemple MDL* ou SPRESI** ne sont pas des sources appropriées de composés de criblage en raison des propriétés non « drug-like » de la plupart des entrées. (En fait, ces bases de données sont utilisées comme des références pour des non médicaments. Cf ci-dessous). De meilleures sources sont des collections disponibles en interne dans les laboratoires pharmaceutiques ou offertes par les vendeurs de composés chimiques, contenant des composés historiques et des bibliothèques combinatoires. Dans la base de données MDL de composés pour le criblage, plus de 3 millions de composés pour le criblage sont disponibles avec des informations de fournisseur. Malheureusement, toutes ces bases de données ont besoin d'un vaste nettoyage pour être appropriées pour le criblage de nouveaux médicaments. Très récemment, la ZINC, une grande bibliothèque nettoyée commerciale de composés pour le criblage est devenue disponible.[23] Les données de référence de composés pharmaceutiques aux différentes étapes de développement peuvent être prises du MDL Drug Data Report+, du World Drug Index++, ou de la base de données MDL Comprehensive Medicinal Chemistry+++. Ces bases de données, qui sont de grande taille, sont aujourd'hui d'un réel apport en diversité de molécules pour les méthodes in silico de découverte de nouvelles molécules thérapeutiques mais peuvent constituer une réelle limitation à cause de la quantité des données à traiter.

II.2.2. Les ressources pour les cibles

UniProt est le catalogue le plus complet de l'information sur les protéines. Il s'agit d'une collection de séquences de protéines et de fonctions créée en joignant les informations contenues dans Swiss-Prot, TrEMBL et PIR. UniProt a trois composantes, chacune optimisée pour différents usages. L'UniProt Knowledgebase (UniProtKB) est le point central d'accès

*http://www.akosgmbh.de/Symyx/software/databases/index.htm
**http://www.spresi.com/
+http://www.cwmglobalsearch.com/Symyx/software/databases/mddr.htm
++http://thomsonreuters.com/products_services/science/science_products/a-z/world_drug_index
+++http://www.cwmglobalsearch.com/Symyx/software/databases/cmc-3d.htm

aux protéines ; il contient des informations très diverses, y compris la fonction, la classification et les références croisées. La base de données PDB est le dépôt mondial d'informations sur les structures tridimensionnelles de grandes molécules biologiques en complexe ou non, y compris des protéines et des acides nucléiques. Au mardi 23 mars 2010 la PDB contient 64 229 structures.

III. Les descripteurs moléculaires et indices de similarité

La structure moléculaire d'un composé organique ou inorganique détermine ses propriétés. Les descripteurs moléculaires représentent les informations topologique, structurale, géométrique et physico chimique permettant de caractériser les molécules et macromolécules (voir Figure I.2). Ils peuvent être calculés à partir de la structure (constitution, configuration et conformation moléculaires) ou des propriétés (physiques, chimiques, biologiques) appartenant aux molécules.[24, 25] Ils sont généralement utilisés pour la recherche de similarité, l'analyse de la diversité/similarité des bibliothèques de composés et de cibles. Les descripteurs moléculaires sont aussi fréquemment utilisés pour développer des modèles statistiques, pour la prédiction informatique de l'activité, la fixation du ligand au récepteur ou les propriétés toxicologiques de composés à partir de leurs structures.

III.1. Descripteurs moléculaires des ligands

III.1.1. Les descripteurs constitutionnels (0D)

Les descripteurs constitutionnels incluent l'information d'ordre des atomes et des liaisons. Ils sont importants pour décrire les propriétés et définir les règles de prédiction des molécules drug-like, telles que la règle des 5 de Lipinski[26], celle publiée par *Veber*[27] *et al* et les modèles statistiques des propriétés pharmacodynamiques, pharmacocinétiques et toxicologiques des molécules.

III.1.2. Les descripteurs physico chimiques

Les descripteurs physico chimiques sont généralement utilisés en combinaison avec d'autres descripteurs pour représenter les éléments contribuant à des propriétés pharmacodynamiques, pharmacocinétiques, toxicologiques spécifiques et les composés qui se fixent à des sites catalytiques de certains récepteurs. Ces descripteurs sont aussi utilisés pour évaluer les modes

de liaisons chimiques, pour estimer l'énergie libre et la constance de proportion des réactions chimiques ainsi que la présence ou l'absence de fragments.

III.1.3. Les descripteurs topologiques (2D)

Les indices topologiques sont utilisés pour représenter le squelette structural et les caractéristiques responsables d'une activité particulière, de la fixation au récepteur et du mode de liaison chimique de la molécule.

III.1.4. Les descripteurs géométriques (3D/4D)

Les descripteurs géométriques concernent l'arrangement en 3D des atomes. Les descripteurs conformationnels (4D) représentent l'arrangement spatial thermodynamique stable des atomes dans une molécule. Les descripteurs géométriques sont utilisés pour prédire l'affinité de liaison et assigner la similarité moléculaire ainsi que pour prédire les propriétés pharmacodynamiques, pharmaco cinétiques et toxicologiques particulières.

III.1.5. Les empreintes moléculaires (1D/2D/3D)

L'empreinte chimique d'une molécule (fingerprint) est une chaîne de bits (séquence de 0 et de 1 chiffre, coefficient de Tanimoto), dans laquelle sont codées les informations sur la structure de la molécule. Elle permet de représenter de façon compacte la structure chimique d'une molécule.

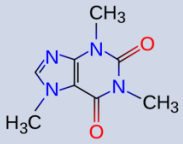
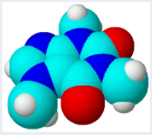
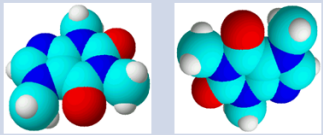
Représentation moléculaire	Descripteurs	Exemples
0D $C_8H_{10}N_4O_2$	Nombre d'atomes, Nombre liaison, poids moléculaire	**Poids moléculaire, nombre atomes hydrogène, nombres atomes carbone, nombre hétéroatomes**
1D $C_8H_{10}N_4O_2$	Nombre de fragments	**Nombre de carbone primaire sp^3, nombre de groupe ammonium, surface polaire basé sur les fragments**
2D 	Descripteurs topologiques	**Index zagreb, index wiener, vecteur autocorrelation 2D**
3D 	Descripteurs géométriques	**3D balaban index, fonction de distribution radiale, surface polaire, volume de la poche, analyse comparative du champs moléculaire (CoMFA), Coordonnées 3D**
4D 	Descripteurs de conformations	**Coordonnées 3D + échantillonnage des conformations**

Figure I.2 - Quelques exemples de descripteurs et leur classification en 1D, 2D, et 3D de la Cafeïne.

III.1.6. Sélection des descripteurs moléculaires

Le calcul et la sélection des descripteurs sont des facteurs déterminants de la réussite du criblage virtuel de molécules. Beaucoup de questions doivent donc être posées. Si des propriétés physicochimiques sont utilisées, il faut fixer à l'avance lesquelles seront retenues et comment elles devront être calculées. Dans le cas de descripteurs structuraux, il faut choisir le niveau de représentation (1D, 2D, 3D ou 4D) en sachant que l'approche 1D présente de nombreux avantages mais est d'un niveau descriptif incomplet. Les descripteurs 2D reflètent bien les propriétés physiques et la réactivité dans la plupart des cas, mais l'activité biologique est étroitement liée à la représentation 3D. Cependant, l'utilisation de structures 3D dans la caractérisation des molécules présente des problèmes de conformation, d'énergie et aussi de disponibilité des bases de données 3D. D'autre part, les tautomères et les ions présentent de nouvelles contraintes. Des approches dites « mixtes » sont très utilisées actuellement mais il faut choisir un groupe de descripteurs en veillant à leur indépendance et à leur utilité. Dans ce choix, le problème à traiter est souvent NP complet, c'est-à-dire un problème pour lequel le temps de résolution peut s'avérer exponentiel. Ainsi, l'usage de techniques d'apprentissage automatique (arbres de décisions, règles d'associations, etc..) semble nécessaire. En raison de

l'existence de bases de molécules de plus en plus grandes, le facteur de vitesse de traitement ne pourra pas être négligé au moment de choisir la représentation optimale. Il est important de noter qu'il n'existe pas de « bon » ou de « mauvais » descripteur : l'utilité et l'efficacité sont étroitement liées aux types de molécules à traiter ainsi qu'aux calculs à effectuer. Par conséquent, la plupart des descripteurs connus aujourd'hui sont employés de préférence dans le contexte pour lesquels ils ont été créés.

III.1.7. Les indices de similarité des ligands

Pour mesurer la (dis) similarité moléculaire, on utilise des fonctions qui transforment les différences entre deux molécules en nombres réels, généralement dans l'intervalle unité [0-1]. Cette quantité fournit une mesure quantitative du niveau de ressemblance chimique pour un jeu de descripteurs donnés.[28] Les mesures de similarité sont généralement constituées de deux éléments : une représentation mathématique de l'information chimique pertinente (en forme de groupes, graphes, vecteurs ou fonctions) et un index compatible avec la représentation. Par exemple, si nous représentons une molécule M_i sous la forme d'un vecteur où chaque composante i correspond à un descripteur moléculaire individuel d_i. D'un point de vue formel, ce vecteur positionne la molécule M dans un point de l'espace vectoriel V, dans lequel chacun des axes correspond à un descripteur (figure.I.3). Cet espace vectoriel s'appelle « l'espace structural ».[29] La (dis) similarité moléculaire entre deux molécules (M1, M2) sera intuitivement reliée à la distance entre les deux points dans cet espace particulier. La règle de calcul de cette distance est appelée «métrique».

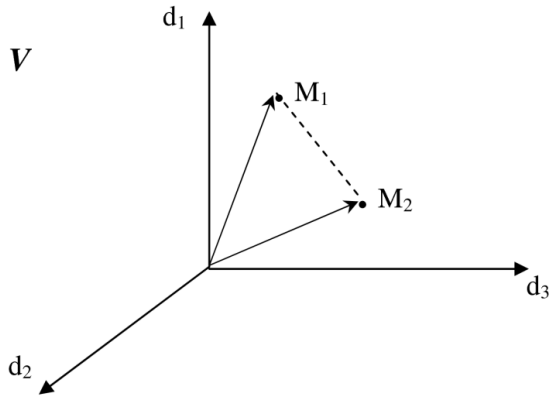


Figure I.3 - L'espace structural de deux molécules représentées par des descripteurs d1, d2 et d3.

Ainsi, toute mesure adéquate de la similarité doit être cohérente avec les propriétés d'une distance mathématique.[30] L'évaluation de similarité peut être abordée par des corrélations, des mesures de distance ou des approches probabilistes ou associatives. La performance de différentes mesures de similarité est le sujet de nombreux travaux.[31] Remarquons que l'évaluation de similarité se fait dans l'espace structural défini par les descripteurs choisis au moyen d'une métrique fixée et non par rapport aux distances interatomiques dans l'espace 3D.

III.2. Descripteurs moléculaires des cibles

III.2.1. La structure primaire

La structure primaire, ou séquence, d'une protéine correspond à la succession linéaire des acides aminés (ou résidus) la constituant. Les protéines sont donc des polymères d'acides aminés, reliés entre eux par des liaisons peptidiques.

III.2.2. Structure secondaire

La structure secondaire décrit le repliement local de la chaîne principale* d'une protéine. 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. Ainsi, une protéine peut être décrite par une séquence d'acides aminés mais aussi par un enchaînement d'éléments de structure secondaire. De plus, certaines conformations se trouvent nettement favorisées car stabilisées par des liaisons hydrogène(s) entre les groupements amide (-NH) et carbonyle (-CO) du squelette peptidique. Il existe trois catégories principales de structures secondaires selon l'échafaudage de liaisons hydrogène(s) et donc selon le repliement des liaisons peptidiques : les hélices, les feuillets et les coudes.

III.2.3. Structure tertiaire

La structure tertiaire d'une protéine correspond au repliement de la chaîne polypeptidique dans l'espace. On parle plus couramment de structure tridimensionnelle ou structure 3D. La structure 3D d'une protéine est intimement liée à sa fonction: lorsque cette structure est cassée par l'emploi d'agents dénaturants, la protéine perd sa fonction: elle est dénaturée.

**la chaîne principale d'une protéine correspond aux atomes impliqués dans la structure de base du polypeptide (-NH-CαH-CO-). Les chaînes latérales des acides aminés (souvent notées -R) n'appartiennent donc pas au squelette carboné.*

III.2.4. Structure quaternaire

La structure quaternaire des protéines regroupe l'association d'au moins deux chaînes polypeptidiques identiques ou différentes par des liaisons non-covalentes (liaison H, liaison ionique, interactions hydrophobes), mais rarement des ponts disulfures. L'effet hydrophobe est un facteur prépondérant dans l'assemblage des éléments structuraux, y compris dans l'association des sous-unités. Chacune de ces chaînes est appelée monomère (ou sous-unité) et l'ensemble est désigné sous le nom d'oligomère ou protéine multimérique.

III.2.5. Interactions responsables de la stabilité conformationnelle

Il est généralement admis que la structure d'une protéine "native" est thermodynamiquement la structure la plus stable. À l'exception des ponts disulfures qui n'existent que dans certaines protéines, principalement les protéines exocellulaires, les interactions qui stabilisent la conformation de ces molécules sont des interactions non covalentes. Toutes les interactions de ce type, qui interviennent dans les petites molécules, existent également dans les protéines. D'autre part, les interactions non covalentes ont lieu entre les divers groupes d'une protéine, mais aussi entre ces groupes et les molécules de solvant. Ainsi, l'énergie conformationnelle d'une molécule protéique est la somme de plusieurs contributions. Certaines de ces contributions résultent de facteurs intrinsèques à la protéine : ce sont les interactions de Van der Waals (non-bonded interactions) qui comportent un terme d'attraction et un terme de répulsion, les potentiels de torsion, les énergies de contrainte dans les angles ou les longueurs de liaison. D'autres proviennent d'interactions intramoléculaires influencées par le solvant, comme les liaisons hydrogène(s) et les interactions électrostatiques. D'autres enfin sont principalement déterminées par le solvant, ce sont les interactions hydrophobes. Les liaisons hydrogène(s) et les interactions hydrophobes présentent une dépendance de signe opposé par rapport à la température. Les liaisons hydrogène(s) sont plus stables à basse température, à l'inverse des interactions hydrophobes; par suite, la température correspondant au maximum de stabilité dépend de la proportion de ces interactions et, par conséquent, varie d'une protéine à l'autre. La structure native d'une protéine résulte d'un équilibre subtil entre différentes interactions stabilisantes et l'entropie conformationnelle qui tend à déstabiliser l'ensemble.

III.2.6. Effet hydrophobe

La séquence d'une protéine comporte une certaine proportion d'acides aminés polaires (hydrophiles) et non polaires (hydrophobes). Leurs interactions avec les molécules d'eau conditionnent la manière dont la chaîne polypeptidique se replie. Les acides aminés non polaires auront tendance à éviter l'eau. Inversement, les résidus polaires vont chercher à rester à proximité du solvant aqueux. Ainsi, dans le cas des protéines solubles, il se forme un cœur hydrophobe au centre de la structure tertiaire, tandis que les groupes polaires restent plutôt en surface. Dans le cas des protéines transmembranaires, le problème est inverse. L'environnement membranaire est globalement hydrophobe. Ainsi, les acides aminés hydrophiles vont se retrouver au cœur de la protéine tandis que les acides aminés hydrophobes vont se retrouver en surface. Des résidus hydrophiles peuvent se retrouver à la surface des protéines membranaires, en contact avec le milieu hydrophobe. Dans ce cas, il y a de fortes chances que ces résidus soient impliqués dans des interactions avec d'autres résidus hydrophiles de la même ou d'une autre protéine.

III.2.7. Les descripteurs moléculaires de la cavité

Un certain nombre de propriétés sont utilisées pour décrire la cavité des protéines présentant un intérêt dans le processus de fixation de la protéine au ligand. Ainsi la cavité peut être représentée par sa forme, sa surface ou son volume. Les groupements chimiques et les propriétés de la cavité nécessaires à l'activité de la protéine peuvent être codés sous forme de bit (0,1) en une empreinte (fingerprint) permettant de caractériser le site de fixation de la protéine considérée.

IV. La dynamique moléculaire

IV.1. Principe

Une simulation de dynamique moléculaire consiste à simuler par le calcul informatique l'évolution d'un système de particules au cours du temps. Ces simulations servent de modèles structuraux et dynamiques pour la compréhension ou la prédiction de résultats expérimentaux. Dans la pratique, cela revient concrètement à simuler le mouvement d'un groupe d'atomes dans le temps. Les simulations par dynamique moléculaire sont majoritairement utilisées pour étudier l'espace conformationnel des macromolécules biologiques ou des petites molécules

chimiques.[32-34] Elles sont valables pour mieux appréhender le comportement des protéines et leurs ligands dans une période de temps donnée. Il est également possible d'étudier l'effet de molécules de solvant explicites sur la structure protéique ou du complexe protéine-ligand. Ces méthodes ne sont pas uniquement utilisées pour rationaliser l'utilisation de structures issues de mesures expérimentales, mais sont également appliquées pour raffiner la plupart des structures issues de la cristallographie par rayons X et de la RMN. L'augmentation de la puissance informatique a permis d'allonger le temps de simulation qui est passé, en une vingtaine d'années, de la pico seconde à la nano seconde et parfois même à la micro seconde. Des progrès également dans les interfaces (membrane simulée, solvant explicite) permettent de mimer l'évolution du système dans un environnement plus complexe que le vide. De meilleurs champs de forces ont par ailleurs vu le jour, impliquant un meilleur traitement des interactions électrostatiques à longue distance. Toutefois, certains problèmes peuvent gêner l'utilisation de ces méthodes, en particulier le piégeage du complexe dans un minimum local qui ne serait pas représentatif de l'espace conformationnel et dans lequel évolue l'ensemble. Les programmes couramment utilisés dans la simulation des biomolécules par dynamique moléculaire sont AMBER[35], CHARMM[36], GROMOS[37] et NAMD[38].

IV.2. Intérêt de la dynamique moléculaire en amont du CVHD

La faiblesse majeure des algorithmes de CVHD est l'absence ou le peu de prise en compte de la flexibilité de la protéine et du ligand lors de l'opération d'amarrage entre le ligand et la protéine.[39] Ceci ne permet donc pas une co-adaptation optimale du ligand et du récepteur. La dynamique moléculaire est capable de traiter la flexibilité de la protéine et du ligand. Par conséquent, le couplage des deux puissantes techniques que sont le CVHD et la dynamique moléculaire doit théoriquement augmenter le pouvoir prédictif de notre modèle.

V. Méthodes CVHD basées sur la structure du ligand (ligand-based)

V.1. Méthodes basées sur la topologie du ligand (2D)

Ce sont des méthodes qui sont basées sur les informations topologiques et la table de connections des molécules chimiques.

V.1.1. Empreinte (fingerprint) topologique

Cette méthode permet de calculer de façon rapide et efficace l'empreinte topologique de toutes les molécules présentes dans une base de données, en ignorant les coordonnées des atomes.[40] Ainsi, en utilisant des indicateurs de similarité tels que l'index de Tanimoto[41], ou le Daylight fingerprint[42], on peut cribler les empreintes topologiques pré calculées pour une large base de données et en extraire les molécules les plus semblables au composé connu comme actif sur une cible biologique. Un aperçu des méthodes basé sur la similarité 2D des empreintes topologiques et leurs performances durant le CVHD a été déjà décrit dans plusieurs travaux.[43, 44] Cependant, les études ont démontré que les touches obtenues en utilisant cette méthode sont moins diverses que celles obtenues par les méthodes qui nécessitent des informations 3D du ligand actif.[45] La diversité étant un facteur important, ceci justifie l'utilisation ultérieure de méthodes un peu plus complexes et qui nécessitent un long temps de calcul.

V.1.2. Arbre de propriétés

C'est une méthode basée sur les descripteurs moléculaires qui utilisent les arbres de propriétés pour une analyse par similarité de larges bases de données de molécules.[46, 47] Toutes les molécules sont décrites par un arbre permettant de caractériser la composition et la hiérarchie de leur descripteur moléculaire. L'arbre décrit les groupements chimiques les plus importants dans la molécule en tenant compte de l'ensemble moléculaire. Ensuite, on utilise un algorithme d'alignement afin d'effectuer un alignement moléculaire basé sur la correspondance entre les différents groupements chimiques fonctionnels des molécules actives. Plusieurs arbres de descripteurs moléculaires peuvent être combinés pour former un multi arbre de descripteurs moléculaires qui présente de réelles performances dans l'identification de nouvelles molécules d'intérêt thérapeutique.[46]

V.2. Méthodes basées sur les descripteurs de la distribution des paires d'atomes centrés (transition de l'espace 2D au 3D)

C'est une méthode basée sur l'utilisation des propriétés des paires d'atomes centrés afin de filtrer de façon rapide de larges bases de données en se basant sur la topologie 2D des molécules telle que la recherche avancée des motifs chimiques (CATS (Chemically Advanced Template Search)).[48] Elle utilise les empreintes 2D topologiques pour une comparaison par paires de molécules. CATS3D est tridimensionnelle et plus gourmande en temps de calcul.

Elle utilise un vecteur de corrélation qui représente la conformation 3D des molécules, permettant ainsi d'augmenter l'exactitude des prédictions.[49] Actuellement, il existe plusieurs autres méthodologies qui sont des extensions utilisant des descripteurs des paires d'atomes centrés, telles que SURFCATS[50], une extension de CATS3D, Similog[51], ou les descripteurs d'entropie de Shannon (SHED)[52]. Plusieurs études prospectives utilisant ces méthodes ont clairement démontré leur potentiel et l'identification de molécules actives.[49-53]

V.3. Méthodes basées sur la représentation géométrique des structures moléculaires (approche 3D)

V.3.1. Superposition des molécules en une conformation ou en ensemble de conformations

L'alignement 3D des molécules en une ou en un ensemble de conformations constitue une autre méthode très répandue du CVHD.[54, 55] La molécule bioactive est utilisée comme structure de référence de la superposition et comparée aux autres molécules de la base de données criblées. Des points pharmacophoriques peuvent être générés pour chacune des molécules de la base de données et utilisés pour l'alignement.[56] Ces groupements chimiques sont ceux qui peuvent être importants dans l'activité biologique de la molécule. L'utilisation des points pharmacophoriques permet la réduction du temps de l'alignement et du criblage de la base de données.

V.3.2. Modélisation de pharmacophore basée sur le ligand

Cette méthode est l'une des plus utilisées actuellement dans le CVHD. Elle permet de définir pour chaque molécule active un pharmacophore qui décrit l'arrangement 3D des propriétés stériques et électroniques clés de la molécule nécessaire à l'activation ou à l'inactivation du processus biologique.[57] La perception du pharmacophore basée sur la similarité géométrique est souvent décrite comme une approche de recherche des analogues des molécules actives, où le pharmacophore est déduit des interactions clés et groupements chimiques importants pour l'activité du ligand en tenant compte de sa flexibilité conformationnelle. Ces dernières années ont vu émerger des applications et des études de recherche de nouveaux ligands basées sur des pharmacophores de ligands.[58-61]

V.3.3. Criblage virtuel basé sur la forme géométrique

La comparaison des formes moléculaires comme la complémentarité stérique joue un rôle important dans la fixation du ligand et peut être utilisée dans le CVHD comme une mesure permettant de quantifier la similarité moléculaire.[62] Ainsi, différents algorithmes d'amarrage moléculaire (FRED[63], ROCS[64, 65], SHEF[66]) utilisés pour le CVHD utilisent la complémentarité de forme pour situer de façon rapide le ligand dans le site actif de la protéine. L'idée de base de cette méthodologie de CVHD est de générer une image complémentaire du site actif de la cible biologique en étudiant la forme du ligand actif et de l'utiliser pour cribler les molécules d'une base de données afin d'extraire les molécules les plus similaires. Ceci suppose qu'on dispose de la structure de référence du ligand actif dans sa cible biologique. De récentes études ont montré l'efficacité de cette méthode pour l'identification de nouvelles molécules actives.[67, 68]

VI. Méthodes de CVHD basées sur la structure de la cible (structure-based)

VI.1. Amarrages moléculaires

L'amarrage moléculaire des protéines et des ligands est considéré aujourd'hui comme une des méthodes principales de CVHD basées sur la structure. Le processus d'amarrage moléculaire entre protéine et ligand peut être divisé en deux étapes. En premier, il faut placer correctement le ligand dans le site de fixation de la protéine. En second, il faut calculer l'affinité du ligand au site de fixation de la protéine par une fonction de score. Au fil des années, on note l'apparition d'un grand nombre de programmes d'amarrage moléculaire avec une grande diversité des fonctions de score basées sur différents algorithmes et champs d'application. Ainsi, on distingue des algorithmes utilisant une construction incrémentale (FlexX[69]), les algorithmes basés sur la forme (DOCK[70]), les algorithmes génétiques (GOLD[71]), la recherche systématique (Glide[72, 73]), les simulations de Monte Carlo (LigandFit[74]) et les algorithmes basés sur la similarité des surfaces moléculaires (Surflex[75]). Les différents programmes d'amarrage moléculaire sont en général capables de générer des poses de ligand dans le site actif de la protéine similaires à celles déterminées expérimentalement.[76] Cependant, contrairement aux autres méthodes rapides de CVHD telles que celles basées sur les pharmacophores, l'amarrage moléculaire nécessite un perpétuel compromis entre les temps de calcul et la précision des résultats.

VI.2. Méthodes basées sur les pharmacophores

VI.2.1. Empreinte des Pharmacophores d'interaction

Les empreintes des pharmacophores d'interaction sont parmi les méthodes les plus rapides du CVHD utilisant l'analyse statistique basée sur des données structurales (FLIP[77], SIFt[78], p-SIFts[79]). Parce qu'elle utilise des données compactes, binaires, représentant les interactions présentes dans le complexe protéine-ligand, cette méthode permet un clustering et une analyse rapide des larges collections de composés. L'efficacité de cette approche a été démontrée dans la recherche de similarité entre molécules actives et celle d'une large base de données, ainsi que dans la génération d'un profil sélectif de petites molécules organiques.[79]

VI.2.2. Modèle de pharmacophore 3D

Cette méthodologie utilise un modèle de pharmacophore 3D plutôt que les coordonnées 3D des atomes de la protéine afin de cribler une base de données de composés en fonction de la superposition géométrique des pharmacophores 3D de ceux-ci à ceux du modèle. Ce qui permet aux méthodes de criblage par les pharmacophores d'être beaucoup moins exigeantes en temps de calcul que les algorithmes d'amarrage moléculaire.[56] Le modèle de pharmacophore peut être généré par différents programmes à partir de la structure 3D d'un complexe protéine-ligand.[56, 80] Ce modèle sera utilisé plus tard pour un CVHD par des plateformes externes (CATALYST, MOE, and PHASE)[81]. Des développements et des recherches récentes sur l'utilisation des pharmacophores 3D basés sur la structure ont montré un grand succès dans l'identification de composés bioactifs, ainsi qu'un grand potentiel de la méthode dans le CVHD.[82-87]

VII. Profilage de l'activité et criblage parallèle

Dans la recherche des composés tête de série, en plus de l'activité sur la cible biologique, il est aussi intéressant de connaître l'effet sur d'autres cibles des molécules à envoyer en test expérimental afin de minimiser les risques dans les phases suivantes du développement du médicament.[88] Un certain nombre d'enzymes, de récepteurs et de canaux ioniques, aussi appelés « anti cible », ont été identifiés comme les principaux responsables des effets secondaires et ADME/Tox. Ce qui fait qu'un composé ayant peu ou pas d'effet sur ces cibles avec des conséquences cardiovasculaires, toxiques ou métaboliques a plus de chance de

passer les étapes suivantes du processus de conception de nouveaux médicaments. Ainsi, le profilage de l'activité des composés intéressants, au début du processus de découverte de nouveaux médicaments, peut réduire significativement le coût des méthodes expérimentales, ainsi que le risque d'échecs. De plus, la technique de criblage parallèle a une grande valeur ajoutée dans la mise en évidence des modes de liaison inconnus et il permet aussi de cribler les médicaments approuvés sur des centaines de cibles afin d'identifier des interactions inconnues. Ces médicaments pourront ainsi être approuvés pour de nouveaux traitements avec un coût financier et expérimental considérablement faible.

VII.1. Criblage parallèle en utilisant les méthodes basées sur les ligands

Cette méthode de profilage de l'activité *in silico* basée sur les ligands est liée à la disponibilité des données pharmacologiques. Plusieurs bases de données publiques (AffinDB[89], PDBbind[90]), ainsi que plusieurs librairies commerciales (WOMBAT[91], MDL Drug Report[92]) sont utilisées par les scientifiques pour cribler des petites molécules organiques sur des cibles protéiques. Cette méthodologie nous permet de reclasser efficacement les composés tête de série par l'identification de ligands analogues.[93, 94]

VII.2. Criblage parallèle en utilisant les méthodes basées sur les cibles

Cette méthode de profilage de l'activité *in silico* basée sur les cibles est liée à la disponibilité des données structurales sur la cible moléculaire. Car, malgré le nombre croissant de structures disponibles dans la PDB, celles-ci ne sont pas réparties de façon égale entre les différentes familles de protéines.

VII.2.1. L'amarrage moléculaire inverse

Cette méthode permet de tester une large base de données de molécules sur plusieurs cibles biologiques afin de mettre en évidence les interactions possibles avec ces différentes cibles. Ensuite, on utilise une fonction de score afin de classer les molécules en fonction de leur affinité prédite.[93-95] Cependant, c'est une méthodologie qui consomme un grand temps de calcul et son automatisation n'est pas évidente.

VII.2.2. Le criblage parallèle basé sur les pharmacophores

Cette méthode permet de cribler une ou plusieurs conformations de molécules ou une large base de données de molécules sur un ou plusieurs modèles pharmacophoriques de cible.[96-98]

La valeur de correspondance est calculée et correspond à la mesure de la similarité des groupements chimiques de la molécule avec celle du pharmacophore.

VIII. Méthodes de modélisation des données

Les méthodes par modélisation des données sont des outils statistiques qui permettent d'explorer les relations pouvant exister entre les descripteurs de molécules calculés à partir des structures chimiques, et des propriétés pharmacocinétiques ou biologiques déterminées expérimentalement.[99] Ces méthodes peuvent être utilisées aussi bien pour l'identification que pour l'optimisation de molécules tête de série.[100]

VIII.1. Méthode basée sur les descripteurs

Les méthodes basées sur les descripteurs utilisent une représentation numérique de la structure chimique pour en déduire un modèle, avec le postulat qu'il y a toujours une fonction qui met en corrélation les propriétés biologiques d'une molécule avec sa structure. Une autre application importante de cette méthodologie est de pouvoir calculer la capacité d'une molécule d'être un candidat médicament (drug-like) ou une tête de série (lead-like), ainsi que de prédire les propriétés ADME/Tox durant les premières étapes du processus d'identification de nouveaux médicaments afin de réduire les risques d'échecs aux étapes expérimentales.[101, 102]

VIII.1.1. Approches linéaires

Historiquement, la méthode la plus répandue est la Relation Quantitative Structure-Activité (QSAR).[103] Elle permet d'établir la corrélation entre certains descripteurs de la structure moléculaire et les propriétés biologiques mesurées de la molécule afin de pouvoir générer un modèle prédictif. Ceci peut se faire par l'utilisation des propriétés 2D, des fragments (Hologramme QSAR)[104] et des propriétés 3D (3D QSAR).[105, 106] Aujourd'hui, le 3D QSAR est l'une des approches linéaires la plus utilisée et la plus performante. Elle présente un net avantage dans la visualisation intuitive du modèle et fournit des informations nécessaires à l'optimisation future des molécules ainsi qu'il est décrit dans des études récentes. Le 3D QSAR permet seulement l'utilisation d'une conformation lors du développement du modèle, ce qui rend le modèle fortement dépendant de l'alignement et permet le traitement d'un petit ensemble de données. Alors, l'utilisation des informations 4D (4D QSAR) permet de

surmonter cette limitation en incluant plusieurs conformations des molécules et une liberté dans l'alignement.[107, 108] Une autre limitation des méthodes linéaires est l'utilisation des méthodes statistiques linéaires telles que le MLR (Multiple Linear Regression), PCA (Principal Component Analysis) et le PLS (Partial Least square) pour la génération du modèle. Ces méthodes statistiques ont l'avantage de la rapidité et de la simplicité.[109] Cependant, dans plusieurs cas, elles ne peuvent pas être utilisées pour corréler avec succès les molécules qui ont une grande diversité, les données avec du bruit et les dépendances non linéaires telles que la prédiction des propriétés ADME/Tox des larges bases de données. Les approches non linéaires peuvent permettre de surmonter ces limitations.

VIII.1.2. Approches non linéaires

L'utilisation de la chimie combinatoire et du criblage à haut débit génère une quantité considérable de données qui peuvent être analysées, non pas par les méthodes linéaires, mais par les méthodes automatisées d'apprentissage et d'analyse statistique. La recherche pharmaceutique tire un grand bénéfice de l'application des méthodes non linéaires telles que : les réseaux de neurones (NN[110]), les machines à vecteurs de support (SVM[111]), les K *plus proches voisins* (k-NN[112]), le partitionnement récursif (RP[113]) et les méthodes de clustering[114]. Ces dernières années, la méthode NN a gagné en popularité et est devenue l'une des principales méthodes d'analyse statistique en industrie pharmaceutique.

VIII.2. System expert et base de données

Les systèmes experts sont des programmes basés sur une analyse experte, à base de connaissances des données qui, après entraînement, sont capables de déduire de nouvelles informations à partir de précédentes informations. Ces programmes sont généralement basés sur des collections de données développées par des experts du domaine de recherche. Ces connaissances hautement structurées sont alors utilisées pour prédire des propriétés basées sur des règles (approche basée sur des règles).[93] Les informations nécessaires sont collectées à partir de différentes études, stockées dans des bases de données et seront utilisées plus tard pour déduire un modèle qui permettra de prédire a priori les cibles biologiques potentielles, les voies métaboliques ou l'intérêt éventuel d'un composé à partir de sa structure chimique seulement. La principale limite de cette approche est la mise à jour incessante des données collectées ainsi que la complexité des descripteurs souvent utilisés.[115]

IX. Exemple de plate forme de criblage virtuel à haut débit : VSM

IX.1. But

De nombreuses méthodes de CVHD sont disponibles comme complément au CEHD dans le processus de découverte de nouvelles molécules thérapeutiques. Mais aucune de ces méthodes ne garantit un niveau de précision et de taille de la base de données criblée comparable à celui du CEHD. Il s'avère donc nécessaire de mettre sur pied des plateformes de CVHD multi étapes, permettant de combiner, dans un entonnoir de criblage de façon séquentielle, différentes méthodes de CVHD. Cette approche hiérarchique permet de combiner en une même expérience de CVHD différentes méthodes afin d'optimiser les paramètres du CVHD telles que la taille des bases de données criblées, la précision des résultats et le gain en temps de calcul de façon à réduire résolument les coûts des premières étapes de découverte de médicament. VSM (Virtual Screening Manager) est une plate-forme de CVHD qui combine de façon séquentielle, dans un entonnoir, plusieurs méthodes de CVHD.[116] Elle permet de combiner l'utilisation séquentielle des méthodes les plus rapides basées sur les ligands en amont de l'entonnoir d'amarrage pour pré filtrer les larges bases de données à cribler aux méthodes nécessitant un grand temps de calcul, basées sur la structure des protéines. Ceci permet d'augmenter la précision et de diminuer les coûts en temps et en argent du CVHD en combinant les avantages des différentes méthodes.

IX.2. Description du prototype et validation

VSM est constituée par une série de différentes méthodes de CVHD basées sur les ligands et sur la structure des protéines, organisées séquentiellement dans une stratégie d'entonnoir. Les techniques s'étendent de méthodes simples aux plus sophistiquées, combinant la rapidité des premiers filtres et la précision des derniers. À chaque étape du processus, le filtre renonce aux composés inopportuns. Les filtres les plus simples et les plus rapides sont utilisés au début dans le processus de filtrage, réservant les méthodes consommant plus de temps pour les dernières étapes. L'entonnoir de criblage de VSM est constitué par un pré filtrage des bases de données à cribler par des méthodes basées sur les ligands (règle des cinq, ADME/Tox), suivi par un pré filtrage avec des méthodes basées sur la structure. Tout d'abord, on a un amarrage moléculaire rigide (SHEF) basé sur la forme géométrique des molécules, rapide, suivi par un amarrage moléculaire flexible (GOLD) gourmand en temps de calcul et beaucoup

plus précis; enfin, l'entonnoir de criblage se termine par de la dynamique moléculaire (NAMD) beaucoup plus gourmande en temps de calcul afin d'affiner les résultats. Ce principe d'entonnoir permet à chaque étape de restreindre le nombre de molécules qui vont passer à l'étape suivante et de diminuer ainsi le coût en temps de calcul. VSM a déjà été implémenté au laboratoire, validé sur la cible biologique LXR β afin d'améliorer l'enrichissement en molécules actives comparé à des méthodes classiques d'amarrage moléculaire.[116] J'ai personnellement contribué à ce travail en analysant les résultats de l'étude pilote sur les composés fournis par l'entreprise pharmaceutique Solvay.

IX.3. Avantages, limitations et développements prévus

VSM permet la sélection dans de larges bases de données en un temps relativement court et avec un taux d'enrichissement en molécules actives relativement élevé comparé aux méthodes classiques de CVHD. Elle permet de combiner différentes méthodes de CVHD et de profiter des avantages de chacune des méthodes. Cependant, elle est limitée pas son taux élevé de faux positifs et par le temps de calcul pour les bases de données de grande taille. Ainsi, les principaux développements futurs seraient, d'une part, son implémentation sur grille de calcul afin de diminuer le temps de calcul et, d'autre part, l'intégration de certaines connaissances du domaine sous forme de contraintes dans l'entonnoir de criblage afin de diminuer le taux de faux positifs.

X. Conclusion

Chaque méthode de CVHD décrite plus haut possède ses propres avantages et inconvénients et doit être utilisée après une analyse approfondie des besoins, des données et de la plate forme technologique disponible. Mais le coût en temps de calcul, la précision des résultats, la précision des fonctions de score et la taille des données disponibles constituent aujourd'hui les principaux handicaps du CVHD.

Aujourd'hui, le filtrage des bibliothèques de molécules conduisant à des chimiothèques focalisées, l'implémentation des méthodes CVHD sur des grilles de calcul et l'introduction de connaissances du domaine constituent les principales voies d'amélioration du CVHD.

Quand il est appliqué à haut débit, le criblage virtuel, basé sur la structure de la cible, nécessite une grande capacité de calcul afin d'effectuer des millions d'amarrages moléculaires

de ligands potentiels avec des centaines de conformations de la cible biologique. Malgré le faible coût et l'augmentation de la vitesse des ordinateurs, un amarrage moléculaire (un ligand sur une cible) prend en moyenne 2 secondes. Ainsi, le temps total pour effectuer 100 000 000 calculs serait de 381 ans ! En plus, le criblage de millions de molécules sur différents sites géographiques nécessite une grande capacité de communication réseaux pour les échanges de données. Ainsi, l'accès à des ressources avec de grandes capacités de calcul telles que les grilles de calcul est donc indispensable afin d'augmenter la rapidité des calculs.

Il existe aujourd'hui une grande panoplie de données biologiques, structurales et physico chimiques sur les protéines et les ligands. Plusieurs études ont montré que l'intégration de ces données et leur analyse par des méthodes de fouille de données, statistiques ou symboliques, sont capables de fournir des unités connaissance sous forme de contraintes pouvant servir de filtre pour les larges bases de molécules. C'est ce qu'on appelle l'extraction de connaissance dans les grandes bases de données. Les unités de connaissance découvertes peuvent concerner par exemple la façon dont un composé se lie à une cible donnée. Les données dont on peut extraire cette connaissance proviennent de sources diverses et hétérogènes, d'où la nécessité d'une bonne intégration de toutes ces données. Les connaissances extraites pourront servir ensuite de contrainte ou de filtre dans l'entonnoir de criblage virtuel, contribuant à la sélection des composés en fonction de leurs propriétés physico chimiques et de leurs données d'interactions avec la cible biologique.

XI. Bibliographie

1. Dobson CM. Chemical space and biology. *Nature.* 2004;432:824-8.
2. Insights. B. Drug approval trends at the FDA and EMEA; [cited March 10, 2010, Available from: www.globalbusinessinsights.com.
3. Owens J. Big pharma slims down to bolster productivity. *Nat Rev Drug Discov.* 2007;6:173-4.
4. Kappel JC, Fan YC, Lam KS. Application of the "libraries from libraries" concept to "one-bead one-compound" combinatorial chemistry. *Adv Exp Med Biol.* 2009;611:21-2.
5. Moos WH, Hurt CR, Morales GA. Combinatorial chemistry: oh what a decade or two can do. *Mol Divers.* 2009;13:241-5.
6. Weber L. High-diversity combinatorial libraries. *Curr Opin Chem Biol.* 2000;4:295-302.
7. Bures MG, Martin YC. Computational methods in molecular diversity and combinatorial chemistry. *Curr Opin Chem Biol.* 1998;2:376-80.
8. Van Hijfte L, Marciniak G, Froloff N. Combinatorial chemistry, automation and molecular diversity: new trends in the pharmaceutical industry. *J Chromatogr B Biomed Sci Appl.* 1999;725:3-15.

9. Leeson PD, Springthorpe B. The influence of drug-like concepts on decision-making in medicinal chemistry. *Nat Rev Drug Discov.* 2007;6:881-90.
10. Verdine GL. The combinatorial chemistry of nature. *Nature.* 1996;384:11-3.
11. Wermuth CG. Similarity in drugs: reflections on analogue design. *Drug Discov Today.* 2006;11:348-54.
12. Wermuth CG. Selective optimization of side activities: another way for drug discovery. *J Med Chem.* 2004;47:1303-14.
13. Newman DJ, Cragg GM. Natural Products as Sources of New Drugs over the Last 25 YearsâŠ¥. *Journal of Natural Products.* 2007;70:461-477.
14. Clardy J, Walsh C. Lessons from natural molecules. *Nature.* 2004;432:829-37.
15. Schreiber SL. Target-oriented and diversity-oriented organic synthesis in drug discovery. *Science.* 2000;287:1964-9.
16. Koch MA, Schuffenhauer A, Scheck Met al. Charting biologically relevant chemical space: a structural classification of natural products (SCONP). *Proc Natl Acad Sci U S A.* 2005;102:17272-7.
17. Congreve M, Chessari G, Tisi D, Woodhead AJ. Recent developments in fragment-based drug discovery. *J Med Chem.* 2008;51:3661-80.
18. Kolb HC, Sharpless KB. The growing impact of click chemistry on drug discovery. *Drug Discov Today.* 2003;8:1128-37.
19. Lewis WG, Green LG, Grynszpan Fet al. Click chemistry in situ: acetylcholinesterase as a reaction vessel for the selective assembly of a femtomolar inhibitor from an array of building blocks. *Angew Chem Int Ed Engl.* 2002;41:1053-7.
20. Fink T, Bruggesser H, Reymond JL. Virtual exploration of the small-molecule chemical universe below 160 Daltons. *Angew Chem Int Ed Engl.* 2005;44:1504-8.
21. Feher M, Schmidt JM. Property distributions: differences between drugs, natural products, and molecules from combinatorial chemistry. *J Chem Inf Comput Sci.* 2003;43:218-27.
22. Bohacek R, McMartin C, Guida W. The art and practice of structure-based drug design: A molecular modeling perspective. *Medicinal Research Reviews.* 1996;16:3-50.
23. Irwin JJ. Using ZINC to acquire a virtual screening library. *Curr Protoc Bioinformatics.* 2008;Chapter 14:Unit 14 6.
24. Brown RD. Descriptors for diversity analysis. *Perspectives in Drug Discovery and Design.* 1997;7/8:31-49.
25. Gasteiger J, Engel T. *Chemoinformatics: A Textbook*. Wiley-VCH: 2003.
26. Lipinski CA. Chris Lipinski discusses life and chemistry after the Rule of Five. *Drug Discov Today.* 2003;8:12-6.
27. Veber DF, Johnson SR, Cheng HY, Smith BR, Ward KW, Kopple KD. Molecular properties that influence the oral bioavailability of drug candidates. *J Med Chem.* 2002;45:2615-23.
28. Willett P, Barnard JM, Downs GM. Chemical Similarity Searching. *Journal of Chemical Information and Computer Sciences.* 1998;38:983-996.
29. Maggiora GM, Shanmugasundaram V. Molecular similarity measures. *Methods Mol Biol.* 2004;275:1-50.
30. Petitjean M. Three-Dimensional Pattern Recognition from Molecular Distance Minimizationâ€ *Journal of Chemical Information and Computer Sciences.* 1996;36:1038-1049.
31. Holliday JD, Hu CY, Willett P. Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2D fragment bit-strings. *Comb Chem High Throughput Screen.* 2002;5:155-66.

32. Hassan SA, Gracia L, Vasudevan G, Steinbach PJ. Computer simulation of protein-ligand interactions: challenges and applications. *Methods Mol Biol.* 2005;305:451-92.
33. Ponder JW, Case DA. Force fields for protein simulations. *Adv Protein Chem.* 2003;66:27-85.
34. Garner J, Deadman J, Rhodes D, Griffith R, Keller PA. A new methodology for the simulation of flexible protein-ligand interactions. *J Mol Graph Model.* 2007;26:187-97.
35. Case DA, Cheatham TE, 3rd, Darden Tet al. The Amber biomolecular simulation programs. *J Comput Chem.* 2005;26:1668-88.
36. Brooks BR, Brooks CL, 3rd, Mackerell AD, Jr.et al. CHARMM: the biomolecular simulation program. *J Comput Chem.* 2009;30:1545-614.
37. Christen M, Hunenberger PH, Bakowies Det al. The GROMOS software for biomolecular simulation: GROMOS05. *J Comput Chem.* 2005;26:1719-51.
38. Phillips JC, Braun R, Wang Wet al. Scalable molecular dynamics with NAMD. *J Comput Chem.* 2005;26:1781-802.
39. Hartmann C, Antes I, Lengauer T. Docking and scoring with alternative side-chain conformations. *Proteins.* 2009;74:712-26.
40. McGaughey GB, Sheridan RP, Bayly CIet al. Comparison of topological, shape, and docking methods in virtual screening. *J Chem Inf Model.* 2007;47:1504-19.
41. Godden JW, Xue L, Bajorath J. Combinatorial preferences affect molecular similarity/diversity calculations using binary fingerprints and Tanimoto coefficients. *J Chem Inf Comput Sci.* 2000;40:163-6.
42. *Daylight Fingerprints*, Daylight Chemical Informations Systems: Santa Fe, NM, 2008.
43. Hert J, Willett P, Wilton DJet al. Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. *Org Biomol Chem.* 2004;2:3256-66.
44. Hert J, Willett P, Wilton DJet al. Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J Chem Inf Comput Sci.* 2004;44:1177-85.
45. Brown RD, Martin YC. Designing combinatorial library mixtures using a genetic algorithm. *J Med Chem.* 1997;40:2304-13.
46. Hessler G, Zimmermann M, Matter Het al. Multiple-ligand-based virtual screening: methods and applications of the MTree approach. *J Med Chem.* 2005;48:6575-84.
47. Rarey M, Dixon JS. Feature trees: a new molecular similarity measure based on tree matching. *J Comput Aided Mol Des.* 1998;12:471-90.
48. Schneider G, Neidhart W, Giller T, Schmid G. "Scaffold-Hopping" by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angew Chem Int Ed Engl.* 1999;38:2894-2896.
49. Renner S, Noeske T, Parsons CG, Schneider P, Weil T, Schneider G. New allosteric modulators of metabotropic glutamate receptor 5 (mGluR5) found by ligand-based virtual screening. *Chembiochem.* 2005;6:620-5.
50. Renner S, Schneider G. Scaffold-hopping potential of ligand-based similarity concepts. *ChemMedChem.* 2006;1:181-5.
51. Schuffenhauer A, Floersheim P, Acklin P, Jacoby E. Similarity metrics for ligands reflecting the similarity of the target proteins. *J Chem Inf Comput Sci.* 2003;43:391-405.
52. Gregori-Puigjane E, Mestres J. SHED: Shannon entropy descriptors from topological feature distributions. *J Chem Inf Model.* 2006;46:1615-22.
53. Renner S, Schwab CH, Gasteiger J, Schneider G. Impact of conformational flexibility on three-dimensional similarity searching using correlation vectors. *J Chem Inf Model.* 2006;46:2324-32.

54. Mestres J, Veeneman GH. Identification of "Latent Hits" in Compound Screening Collections. *Journal of Medicinal Chemistry.* 2003;46:3441-3444.
55. Lemmen C, Lengauer T. Computational methods for the structural alignment of molecules. *J Comput Aided Mol Des.* 2000;14:215-32.
56. Wolber G, Dornhofer AA, Langer T. Efficient overlay of small organic molecules using 3D pharmacophores. *J Comput Aided Mol Des.* 2006;20:773-88.
57. Wermuth, C G, Ganellinet al. *GLOSSARY OF TERMS USED IN MEDICINAL CHEMISTRY (IUPAC RECOMMENDATIONS 1997)*. Academic Press: San Diego, CA, ETATS-UNIS, 1998; Vol. 33, p 11.
58. Evans DA, Doman TN, Thorner DA, Bodkin MJ. 3D QSAR methods: Phase and Catalyst compared. *J Chem Inf Model.* 2007;47:1248-57.
59. Ortuso F, Langer T, Alcaro S. GBPM: GRID-based pharmacophore model: concept and application studies to protein-protein recognition. *Bioinformatics.* 2006;22:1449-55.
60. Chimenti F, Bolasco A, Manna Fet al. Synthesis, biological evaluation and 3D-QSAR of 1,3,5-trisubstituted-4,5-dihydro-(1H)-pyrazole derivatives as potent and highly selective monoamine oxidase A inhibitors. *Curr Med Chem.* 2006;13:1411-28.
61. Patel Y, Gillet VJ, Bravi G, Leach AR. A comparison of the pharmacophore identification programs: Catalyst, DISCO and GASP. *J Comput Aided Mol Des.* 2002;16:653-81.
62. Haigh JA, Pickup BT, Grant JA, Nicholls A. Small molecule shape-fingerprints. *J Chem Inf Model.* 2005;45:673-84.
63. McGann MR, Almond HR, Nicholls A, Grant JA, Brown FK. Gaussian docking functions. *Biopolymers.* 2003;68:76-90.
64. Nicholls A, Grant JA. Molecular shape and electrostatics in the encoding of relevant chemical information. *J Comput Aided Mol Des.* 2005;19:661-86.
65. Rush TS, 3rd, Grant JA, 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;48:1489-95.
66. Cai W, Xu J, Shao X, Leroux V, Beautrait A, Maigret B. SHEF: a vHTS geometrical filter using coefficients of spherical harmonic molecular surfaces. *J Mol Model.* 2008;14:393-401.
67. Bostrom J, Berggren K, Elebring T, Greasley PJ, Wilstermann M. Scaffold hopping, synthesis and structure-activity relationships of 5,6-diaryl-pyrazine-2-amide derivatives: a novel series of CB1 receptor antagonists. *Bioorg Med Chem.* 2007;15:4077-84.
68. Sykes MJ, Sorich MJ, Miners JO. Molecular modeling approaches for the prediction of the nonspecific binding of drugs to hepatic microsomes. *J Chem Inf Model.* 2006;46:2661-73.
69. Rarey M, Kramer B, Lengauer T, Klebe G. A fast flexible docking method using an incremental construction algorithm. *J Mol Biol.* 1996;261:470-89.
70. Ewing TJ, Makino S, Skillman AG, Kuntz ID. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *J Comput Aided Mol Des.* 2001;15:411-28.
71. Jones G, Willett P, Glen RC, Leach AR, Taylor R. Development and validation of a genetic algorithm for flexible docking. *J Mol Biol.* 1997;267:727-48.
72. Halgren TA, Murphy RB, Friesner RAet al. Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J Med Chem.* 2004;47:1750-9.

73. Friesner RA, Banks JL, Murphy RBet al. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J Med Chem.* 2004;47:1739-49.
74. Venkatachalam CM, 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;21:289-307.
75. Jain AN. Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *J Med Chem.* 2003;46:499-511.
76. Warren GL, Andrews CW, Capelli AMet al. A critical assessment of docking programs and scoring functions. *J Med Chem.* 2006;49:5912-31.
77. Prabha K, Amit K. Application of Pharmacophore Fingerprints to Structure-Based Design and Data Mining. In *Pharmacophores and Pharmacophore Searches*, Prof. Dr. Thierry Langer, D. R. D. H., Ed.; 2006; pp 193-206.
78. 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;47:337-44.
79. Chuaqui C, Deng Z, Singh J. Interaction profiles of protein kinase-inhibitor complexes and their application to virtual screening. *J Med Chem.* 2005;48:121-33.
80. Wolber G, Langer T. LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *J Chem Inf Model.* 2005;45:160-9.
81. Wolber G, Seidel T, Bendix F, Langer T. Molecule-pharmacophore superpositioning and pattern matching in computational drug design. *Drug Discov Today.* 2008;13:23-9.
82. Adane L, Patel DS, Bharatam PV. Shape- and chemical feature-based 3D-pharmacophore model generation and virtual screening: identification of potential leads for P. falciparum DHFR enzyme inhibition. *Chem Biol Drug Des.* 75:115-26.
83. Aparoy P, Kumar Reddy K, Kalangi SK, Chandramohan Reddy T, Reddanna P. Pharmacophore modeling and virtual screening for designing potential 5-Lipoxygenase inhibitors. *Bioorg Med Chem Lett.* 20:1013-8.
84. Kansal N, Silakari O, Ravikumar M. Three dimensional pharmacophore modelling for c-Kit receptor tyrosine kinase inhibitors. *Eur J Med Chem.* 45:393-404.
85. Krovat EM, Fruhwirth KH, Langer T. Pharmacophore identification, in silico screening, and virtual library design for inhibitors of the human factor Xa. *J Chem Inf Model.* 2005;45:146-59.
86. Leach AR, Gillet VJ, Lewis RA, Taylor R. Three-dimensional pharmacophore methods in drug discovery. *J Med Chem.* 53:539-58.
87. Schuster D, Maurer EM, Laggner Cet al. The discovery of new 11beta-hydroxysteroid dehydrogenase type 1 inhibitors by common feature pharmacophore modeling and virtual screening. *J Med Chem.* 2006;49:3454-66.
88. Choong NW, Cohen EE. Forthcoming receptor tyrosine kinase inhibitors. *Expert Opin Ther Targets.* 2006;10:793-7.
89. Block P, Sotriffer CA, Dramburg I, Klebe G. AffinDB: a freely accessible database of affinities for protein-ligand complexes from the PDB. *Nucleic Acids Res.* 2006;34:D522-6.
90. Wang R, Fang X, Lu Y, Yang CY, Wang S. The PDBbind database: methodologies and updates. *J Med Chem.* 2005;48:4111-9.
91. Marius O, Maria M, Liliana Oet al. WOMBAT: World of Molecular Bioactivity. In *Chemoinformatics in Drug Discovery*, Prof. Dr. Tudor, I. O., Ed.; 2005; pp 221-239.
92. Sheridan RP, Shpungin J. Calculating similarities between biological activities in the MDL Drug Data Report database. *J Chem Inf Comput Sci.* 2004;44:727-40.

93. Ekins S, Mestres J, Testa B. In silico pharmacology for drug discovery: methods for virtual ligand screening and profiling. *Br J Pharmacol.* 2007;152:9-20.
94. Poulain R, Horvath D, Bonnet Bet al. From hit to lead. Analyzing structure-profile relationships. *J Med Chem.* 2001;44:3391-401.
95. Toledo-Sherman LM, Chen D. High-throughput virtual screening for drug discovery in parallel. *Curr Opin Drug Discov Devel.* 2002;5:414-21.
96. Steindl TM, Schuster D, Laggner C, Chuang K, Hoffmann RD, Langer T. Parallel screening and activity profiling with HIV protease inhibitor pharmacophore models. *J Chem Inf Model.* 2007;47:563-71.
97. Steindl TM, Schuster D, Wolber G, Laggner C, Langer T. High-throughput structure-based pharmacophore modelling as a basis for successful parallel virtual screening. *J Comput Aided Mol Des.* 2006;20:703-15.
98. Steindl TM, Schuster D, Laggner C, Langer T. Parallel screening: a novel concept in pharmacophore modeling and virtual screening. *J Chem Inf Model.* 2006;46:2146-57.
99. van de Waterbeemd H, Gifford E. ADMET in silico modelling: towards prediction paradise? *Nat Rev Drug Discov.* 2003;2:192-204.
100. Schnecke V, Bostrom J. Computational chemistry-driven decision making in lead generation. *Drug Discov Today.* 2006;11:43-50.
101. Lu XY, Chen YD, You QD. 3D-QSAR studies of arylcarboxamides with inhibitory activity on InhA using pharmacophore-based alignment. *Chem Biol Drug Des.* 75:195-203.
102. Kasnanen H, Myllymaki MJ, Minkkila Aet al. 3-Heterocycle-phenyl N-alkylcarbamates as FAAH inhibitors: design, synthesis and 3D-QSAR studies. *ChemMedChem.* 5:213-31.
103. Ekins S, Mestres J, Testa B. In silico pharmacology for drug discovery: applications to targets and beyond. *Br J Pharmacol.* 2007;152:21-37.
104. Moda TL, Montanari CA, Andricopulo AD. Hologram QSAR model for the prediction of human oral bioavailability. *Bioorg Med Chem.* 2007;15:7738-45.
105. Cramer RD, 3rd, Patterson DE, Bunce JD. Recent advances in comparative molecular field analysis (CoMFA). *Prog Clin Biol Res.* 1989;291:161-5.
106. Klebe G, Abraham U, Mietzner T. Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. *J Med Chem.* 1994;37:4130-46.
107. Potemkin V, Grishina M. Principles for 3D/4D QSAR classification of drugs. *Drug Discov Today.* 2008;13:952-9.
108. Ekins S, Bravi G, Binkley Set al. Three- and four-dimensional-quantitative structure activity relationship (3D/4D-QSAR) analyses of CYP2C9 inhibitors. *Drug Metab Dispos.* 2000;28:994-1002.
109. Weaver DC. Applying data mining techniques to library design, lead generation and lead optimization. *Curr Opin Chem Biol.* 2004;8:264-70.
110. Gola, Joelle, Obrezanovaet al. *ADMET property prediction : The state of the art and current challenges*. Wiley-Vch: Weinheim, ALLEMAGNE, 2006; Vol. 25, p 9.
111. Matthew WBT, Sean BH. Support Vector Machines for ADME Property Classification. *QSAR & Combinatorial Science.* 2003;22:533-548.
112. Berith F, Per BB, Christian V, Søren BP, Hanne HFR. <I>In Silico</I> Classification of Solubility using Binary <I>k</I>-Nearest Neighbor and Physicochemical Descriptors. *QSAR & Combinatorial Science.* 2007;26:452-459.
113. Hou T, Wang J, Zhang W, Xu X. ADME evaluation in drug discovery. 7. Prediction of oral absorption by correlation and classification. *J Chem Inf Model.* 2007;47:208-18.

114. Ratle F, Gagné C, Terrettaz-Zufferey A-L, Kanevski M, Esseiva P, Ribaux O. Advanced clustering methods for mining chemical databases in forensic science. *Chemometrics and Intelligent Laboratory Systems*. 2008;90:123-131.
115. Yamashita F, Hashida M. In silico approaches for predicting ADME properties of drugs. *Drug Metab Pharmacokinet.* 2004;19:327-38.
116. Beautrait A, Leroux V, Chavent Met al. Multiple-step virtual screening using VSM-G: overview and validation of fast geometrical matching enrichment. *J Mol Model.* 2008;14:135-48.

« "bigger is better" and "faster is best". »

CHAPITRE 2 - Amélioration des performances par utilisation des grilles de calcul : VSM-G

Les grilles de calcul sont une nouvelle Technologie de l'Information permettant la collecte et le partage de l'information, la mise en réseau d'experts et la mobilisation de ressources en routine ou en urgence. Elles ouvrent de nouvelles perspectives de réduction des coûts et d'accélération de la recherche *in silico* de médicaments. Ce chapitre présente les grilles de calcul et leur utilisation dans un protocole multi étapes de CVHD. Nous présenterons ici l'implémentation et l'utilisation de la plateforme de CVHD (VSM) sur grille de calcul (VSM-G) et les résultats que nous avons obtenus avec un jeu de ligands de la base de données ZINC sur trois structures du récepteur de l'hormone LXRß.

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

Efficiency of a hierarchical protocol for high throughput structure-based virtual screening on GRID5000 cluster Grid

Leo Ghemtio[1], Emmanuel Jeannot[2,3], B Maigret[1]

1: Nancy Université, LORIA, Equipe ORPAILLEUR, Campus scientifique, BP 239, 54506 Vandoeuvre-lès-Nancy Cedex, France
2: Nancy Université, LORIA, Equipe ALGORILLE, Campus scientifique, BP 239, 54506 Vandoeuvre-lès-Nancy Cedex, France
3: Bordeaux Université, LABRI, Equipe Runtime, 351, Cours de la Libération, 33405 Talence Cedex, France
Open Access Bioinformatics 2010:2 1-13

Sommaire

I. Contexte

La présente étude a été publiée dans « Journal of Open Access Bioinformatics » en Janvier 2010. En tant que premier auteur, j'ai réalisé les principales étapes d'initiation de l'étude, de recherche, de calcul, d'interprétation et d'écriture du manuscrit en collaboration avec Emmanuel Jeannot de l'équipe Algorille du LORIA. Nous nous sommes employés dans cette étude à élaborer une méthode automatique de CVHD organisée de façon séquentielle dans un entonnoir de CVHD multi étapes combinant les méthodes basées sur les ligands à celles basées sur la structure de la cible et utilisant une grille de calcul pour diminuer le coût et le temps des calculs.

Le CVHD, qui est une des premières étapes du processus de découverte de nouveaux médicaments quand la cible biologique a déjà été identifiée, témoigne de l'importance et des avancées réalisées dans la simulation des systèmes biologiques assistée par ordinateur, appliquée aux molécules et macromolécules biologiques. Cependant, la modélisation précise des systèmes biologiques extrêmement complexes, qui peuvent conduire à une accélération spectaculaire et à une rationalisation du processus de découverte de nouveaux médicaments, lent et consommateur de ressources est encore loin de la faisabilité des systèmes des ordinateurs modernes actuels. Sans surprise, les approches empiriques rapides exploitant les données sur les propriétés des petites molécules organiques et souvent liées aux données de structure-activité collectées sont désormais plus utilisées dans la conception de nouveaux médicaments comparées aux méthodes de simulation moléculaire tridimensionnelle très gourmandes en temps de calcul. L'imprécision et les limites d'application des méthodes qui utilisent les données sur les ligands les rendent moins populaires que les simulations tridimensionnelles du comportement du ligand dans le site actif de la protéine. Ainsi, la gestion à grande échelle dans le contexte du haut débit des données de tels systèmes complexes exige des installations informatiques massivement parallèles et des algorithmes appropriés. Pour pouvoir cribler rapidement de larges bases de composés contenant des millions de molécules de façon fiable et à faible coût, une méthodologie d'amarrage moléculaire multi étapes a été mise sur pied dans notre laboratoire (VSM : Virtual Screening Manager). VSM est constitué par une série de différentes méthodes de CVHD basées sur les ligands et sur la structure, organisées séquentiellement dans une stratégie d'entonnoir. Les techniques s'étendent de méthodes simples, uniquement basées sur la forme et pouvant s'appliquer à des millions de molécules, aux plus sophistiquées prenant en compte des

propriétés physico-chimiques et ne pouvant s'appliquer qu'à un nombre restreint de molécules, combinant la rapidité des premiers et la précision des derniers. À chaque étape du processus, le filtre renonce aux composés inopportuns. Les filtres les plus simples et les plus rapides sont utilisés au début dans le processus de filtrage. Les méthodes consommant plus de temps sont utilisées aux dernières étapes. Ainsi, au cours de ce chapitre, nous allons optimiser le temps et le coût de l'amarrage moléculaire basé sur la structure par VSM en couplant son utilisation à une grille, la grille de calculs GRID5000.

GRID5000 est une plate forme scientifique pour l'étude à grande échelle des systèmes parallèles et distribués. Il vise à fournir une plate forme expérimentale fortement reconfigurable, contrôlable et monitorable pour ses utilisateurs. Le but initial était d'atteindre 5000 processeurs dans la plate forme. Il a été ré-estimé à 5000 cœurs et a été atteint pendant l'hiver 2008-2009. L'infrastructure de GRID5000 est géographiquement distribuée sur des sites différents, initialement 9 en France. Porto Allégre, au Brésil, devient maintenant officiellement le 10ème site.

Le but de ce chapitre est de décrire VSM-G (Virtual Screening Manager for Computational Grids), qui est une plate forme de déploiement à grande échelle des méthodes de CVHD sur la grille de calculs expérimentale GRID5000, ainsi que les conditions nécessaires à un tel déploiement, les composantes de l'entonnoir de criblage et l'automatisation de tout le protocole. Ainsi, nous avons utilisé VSM-G pour le CVHD de la base de données ZINC contre 3 structures du récepteur LXRß. La première étape est un filtre basé sur les propriétés qui caractérisent un médicament potentiel et ADMETOX des molécules de la base à filtrer. Les deux étapes suivantes de la procédure sont la correspondance géométrique rapide MSSH (Harmonique Sphérique Superficielle Moléculaire)/SHEF (Filtre de coefficient Harmonique Sphérique), suivie du programme d'amarrage moléculaire flexible GOLD. A la fin de l'entonnoir, nous avons utilisé la simulation par dynamique moléculaire des meilleures protéines et structures de ligand avec NAMD afin de raffiner les résultats obtenus et de les reclasser en fonction de l'énergie libre d'interaction des complexes protéines-ligands.

II. Efficiency of a hierarchical protocol for high throughput structure-based virtual screening on GRID5000 cluster Grid

III. Commentaires

Dans ce chapitre, on a décrit l'utilisation d'une grille de calcul expérimentale (GRID5000), couplée à un protocole de criblage multi étapes en entonnoir (VSM) combinant les approches basées sur les ligands à celles basées sur la structure de la protéine, permettant de sélectionner des molécules de la base de données ZINC potentiellement actives sur la cible biologique LXRβ. Il existe différents types de grille de calcul et il y a beaucoup de façons de les classifier, selon leurs buts, leurs technologies, leurs propriétés ou leurs matériels par exemple. La complexité des ressources matérielles sous-jacentes doit être gérée et cachée par des couches logicielles fournissant des services pour des utilisateurs. Le meilleur type de grille de calcul utilisable pour la découverte in silico de nouvelles molécules thérapeutiques est la grille de cluster. Il y a différents types d'infrastructures et de technologie de grilles de cluster dans le monde. La grille européenne (EGEE), la grille nationale de Taiwan (TWGrid), la grille régionale Française (AuverGrid) sont de larges infrastructures de production, basées sur la technologie EGEE, fournissant un environnement puissant de calculs et de gestion de grandes quantités de données qui peuvent être utilisées dans le cadre du CVHD.

Le travail décrit dans ce chapitre a été une expérience utile dans l'identification des limitations et des goulots d'étranglement de l'infrastructure GRID5000. La plate forme VSM-G a été développée pour soumettre les calculs sur la grille de calculs avec un faible taux d'échec des soumissions. D'autre part, le gestionnaire de ressources a significativement limité le taux de soumission des calculs. Une autre source significative d'inefficacité est venue de la difficulté du système d'information de la grille de calcul de fournir toutes les informations appropriées au gestionnaire de ressources pour la distribution des calculs sur la grille de calculs. En conséquence, la planification de travail était une tâche consommatrice de temps lors de la gestion des données, en raison des limitations rencontrées par le système d'information, les éléments de calculs et le gestionnaire de ressources. L'expérience a montré

comment les grilles de calcul ont une capacité énorme pour mobiliser de très grandes ressources CPU vers des buts bien ciblés pendant un temps bien précis. La grille expérimentale GRID5000 a démontré que les grilles de calcul peuvent être utilisées dans le processus de découverte de médicaments. Les grilles de calcul couplées au CVHD ont permis l'implémentation de l'entonnoir de CVHD sur une infrastructure de grilles de clusters. Ici, la comparaison de forme par MSSH/SHEF est la deuxième étape du processus de filtrage venant après le filtrage initial basé sur les propriétés physico-chimiques qui caractérisent un médicament potentiel et sa toxicité éventuelle. Cette deuxième étape est suivie de l'amarrage moléculaire par GOLD, la dernière étape consistant à effectuer sur le petit nombre de molécules finalement retenu les simulations de dynamique moléculaire. Une évaluation grossière des ressources exigées est d'environ 3144 nœuds répartis sur 9 sites pendant 82 jours pour cribler une base de données de 600 000 molécules provenant de la base de données ZINC sur 3 structures de la cible biologique LXR β. L'expérience a produit une grande quantité de données à analyser. Des résultats, on a extrait 45 composés environ possédant les interactions clés et un bon classement. Ces composés sont actuellement en train de subir des analyses plus spécifiques afin d'identifier les composés tête de série pour des essais en laboratoire.

L'utilisation des grilles de calculs et du protocole hiérarchique de filtrage en entonnoir permet d'évaluer le potentiel d'affinités de millions de molécules sur plusieurs conformations d'une cible biologique avec un faible coût et d'identifier les composés les plus prometteurs pour les essais in vitro.

« Ignorance is the curse of God, knowledge the wing wherewith we fly to heaven. »

William Shakespeare

CHAPITRE 3 - Découverte de connaissances pour le criblage virtuel

Dans ce chapitre, nous présenterons l'utilisation de la découverte des connaissances dans les bases de données biologiques, chimiques, et structurales liées aux ligands et à leurs cibles, comme une méthode de CVHD. Dans ce contexte, nous décrirons les différentes étapes nécessaires à la mise en œuvre d'un processus de KDD (Knowledge Discovery in Database). Nous présenterons ici l'approche KDD avec différents algorithmes de fouille de données et les résultats que nous avons obtenus avec un jeu de ligands du récepteur de l'hormone LXRß.

Ce chapitre a fait l'objet de deux publications à des conférences internationales.

Model-driven Data Integration for Mining Protein-Ligand and Protein-Protein Interactions in a Drug Design Context

Leo Ghemtio, Emmanuel Bresso, Michel Souchet, Bernard Maigret, Malika Smaïl-Tabbone and Marie-Dominique Devignes
Journées Ouvertes Biologie Informatique Mathématiques – JOBIM, Lille (France) 2009

A KDD approach for designing filtering strategies to improve virtual screening

Leo Ghemtio, Malika Smaïl-Tabbone, Marie-Dominique Devignes, Michel Souchet, Bernard Maigret

LORIA UMR 7503, CNRS, Nancy-Université, and INRIA Research Centre Nancy Grand-Est, BP239, 54506 Vandoeuvre-les-Nancy cedex, France
International Conference on Knowledge Discovery and Information Retrieval, Madeira (Portugal) 2009

Sommaire

I. Contexte

La présente étude a fait l'objet de deux publications: l'une à la conférence JOBIM 2008 et l'autre à la conférence KDIR 2009. En tant que principal auteur, j'ai réalisé les principales étapes d'initiation de l'étude, de recherche, de calcul, d'interprétation et d'écriture des manuscrits en collaboration avec les autres membres de l'équipe impliqués dans l'étude. Dans cette étude, nous avons conçu et implémenté une base de données regroupant des informations hétérogènes sur plusieurs familles de protéines dont la protéine LXRß, ses ligands potentiels et les interactions protéines-protéines et protéines-ligands collectées à partir de différentes publications ou bases de données. La base de données P3LI (Protein-Protein and Protein-Ligand Interaction) a été développée dans le cadre de cette étude, ceci afin de pouvoir y découvrir des connaissances qui permettent de caractériser les ligands actifs sur une cible déterminée, en l'occurrence LXRß.

La quantité énorme de données rassemblées dans des bases de données excède aujourd'hui de loin notre capacité à les réduire et les analyser sans l'utilisation de techniques d'analyse automatisées. La découverte de connaissances KDD pour « Knowledge Discovery from Databases ») vise à identifier de façon explicite des unités de connaissances valides, nouvelles, utilisables et compréhensibles à partir de grandes quantités de données. La fouilles de données est au cœur de l'approche KDD et utilise des algorithmes qui explorent les données, développent des modèles et découvrent des motifs significatifs qui peuvent être traduits en unités de connaissances. Aujourd'hui, de grandes quantités de données sont disponibles dans des bases de données hétérogènes sur les protéines et leurs ligands. Ces informations, et les connaissances qui peuvent y être découvertes, peuvent conduire à concevoir des filtres, applicable aux bases de données de molécules, permettant d'accélérer et d'améliorer le processus de découverte de nouveaux médicaments. Un exemple de connaissances à découvrir concerne la capacité d'une molécule à être un 'bon' ou un 'mauvais' ligand pour une cible biologique. Pour cela il est nécessaire de disposer d'exemples bien documentés dans la littérature. Ainsi, l'approche KDD peut se présenter comme une nouvelle technique de CVHD permettant, à partir de règles d'association ou de motifs caractérisant les ligands actifs, de filtrer rapidement une nouvelle base de données de molécules, dont on ne connaît pas l'activité sur la cible considérée avec un faible coût si on le compare aux autres méthodes de CVHD. Cependant, les bases de données sont nombreuses, et hétérogènes dans leurs formats, ceci ne facilitant pas leur utilisation. Ainsi, avant le travail

de découverte, il est indispensable d'effectuer un travail d'intégration et de structuration des données, par exemple dans une base de données relationnelles. Il devient alors plus facile de construire de façon itérative des jeux de données appropriés pour la fouille de données, et adaptables à un vaste panel d'algorithmes allant des arbres de décision à la recherche de motifs fréquents et aux règles d'association.

Tout au long de ce chapitre, on détaillera la mise en place de la base de données P3LI. Toutes les étapes de conception, d'intégration, d'interrogation et d'extraction de connaissances dans la base de données seront présentées. Les connaissances découvertes à partir des données biologiques, physico-chimiques, structurales, stockées sur la protéine LXRß, ses ligands potentiels et leurs interactions seront utilisées pour filtrer une collection de molécules provenant de la base de données ZINC, afin de ne garder que les molécules les plus susceptibles d'être des ligands potentiels de LXRß en accord avec le modèle extrait.

II. Model-driven Data Integration for Mining Protein-Ligand and Protein-Protein Interactions in a Drug Design Context

III. A KDD approach for designing filtering strategies to improve virtual screening

IV. Commentaires

Dans ce chapitre j'ai décrit un protocole de découverte de connaissances dans une base de données contenant des données biologiques, physico-chimiques et structurales sur les protéines, les ligands et leurs interactions. Les unités de connaissances ainsi extraites peuvent être utilisées comme un filtre, afin de sélectionner dans une collection de données provenant de la base de données ZINC, les molécules les plus susceptibles d'être des ligands potentiels de la cible biologique LXRβ. La base de données implémentée P3LI (Protein-Protein and Protein-Ligand Interaction) est une base de données relationnelle, comportant aussi bien des données sur différentes familles de protéines que sur leurs ligands potentiels et sur les interactions entre ces protéines et leurs ligands. Les données stockées dans P3LI sont des informations biologiques, structurales et des descripteurs moléculaires 1D, 2D, 3D, ou 4D permettant de caractériser ou de décrire les protéines, les molécules et leurs interactions. Ainsi, deux algorithmes de découverte de connaissances ont été appliqués à cette base de données relationnelle afin d'y extraire les règles d'association ou de constituer des arbres de décision permettant d'identifier un 'bon' ou un 'mauvais' ligand pour la protéine LXRβ. L'étape finale de découverte de connaissances à partir d'une base de données est de vérifier si les modèles produits (règles d'association ou arbres de décision) par les algorithmes de découverte de connaissances sont applicables à des jeux de données plus larges. Tous les modèles trouvés par les algorithmes de découverte de connaissances ne sont pas nécessairement valides. C'est pourquoi il est nécessaire d'effectuer une évaluation du modèle sur un jeu de données test qui n'a pas été utilisé pour créer le modèle. Ainsi, si le modèle construit ne permet pas d'identifier les éléments recherchés, il est nécessaire de réévaluer le modèle, de changer le prétraitement ou l'algorithme de fouille de données. Si le modèle construit permet d'identifier les éléments recherchés, on peut interpréter le modèle construit et le transformer en unité de connaissance.

Le travail décrit dans ce chapitre a été une expérience utile dans la conception de modèle de base de données permettant l'intégration de grande quantité de données hétérogènes et la mise en place de règles et la comparaison de différents algorithmes pouvant permettre le filtrage de larges bases de données avant une étude plus approfondie par des techniques plus exigeantes en temps et en argent. Une évaluation grossière des ressources exigées pour l'étape de filtrage par les règles générées par le KDD dans le cadre de la recherche de ligands potentiels à la cible protéique LXRβ est de 1 nœud d'un ordinateur et environ 2 minutes pour générer un modèle à partir d'une base de données de 300 molécules provenant des études de relation structure activité (SAR).

L'utilisation du KDD a permis d'évaluer le potentiel d'affinités de millions de molécules sur une cible biologique avec un faible coût et d'identifier les composés les plus prometteurs pour les étapes suivantes de la recherche *in silico* de nouveaux médicaments. Le KDD est un domaine en plein expansion avec une grande applicabilité dans différents domaines de recherche. Elle est pressentie comme la nouvelle technologie de base de données pendant les prochaines années. Le besoin en outils de KDD automatisés provoque une explosion du nombre et du type d'outils commerciaux ou dans le domaine public (Weka*, Coron*). On prévoit que les systèmes de base de données commerciaux de l'avenir incluront des outils de KDD sous la forme d'interfaces intelligentes de base de données.

**http://www.cs.waikato.ac.nz/ml/weka/*
**http://coron.loria.fr/site/index.php*

« Great things are done by a series of small things brought together »

Vincent van Gogh

CHAPITRE 4 - Application 1 : Évaluation d'un filtre à base de connaissances

Dans le chapitre précédent, nous avons présenté la mise en œuvre d'une méthodologie de découverte de connaissances dans une base de données relationnelles (P3LI) contenant des données hétérogènes biologiques, physico-chimiques, et structurales sur les protéines, les ligands et leurs interactions. Dans ce chapitre, nous présenterons l'utilisation de ces connaissances comme un filtre pour le CVHD. Celui-ci sera comparée à la méthode de CVHD multi étapes (VSM-G) qui a été mise sur pied dans notre laboratoire afin d'optimiser les méthodes classiques de CVHD, ainsi qu'à la méthode basée sur les pharmacophores 3D, présentée comme une des méthodes les plus prometteuses de nos jours dans le CVHD. Ces 3 méthodes seront utilisées pour cribler un jeu de ligands de la base de données ZINC sur trois structures du récepteur de l'hormone LXRß.

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

Comparison of three pre-processing filters efficiency in virtual screening: Identification of new putative LXRß regulators as a test case

Leo Ghemtio, Marie-Dominique Devignes, Malika Smaïl-Tabbone, Michel Souchet[#], Vincent Leroux, Bernard Maigret

Nancy Université, LORIA, Groupe ORPAILLEUR, Campus scientifique, BP 239, 54506 Vandoeuvre-lès-Nancy Cedex, France
#: Present address : Harmonic Pharma, 615 rue du Jardin Botanique, 54600 Villers les Nancy, France
Journal of Chemical Information and Modeling 2010 Article ASAP

Sommaire

I. Contexte

La présente étude a été publiée dans « Journal of Chemical Information and Modeling » en Septembre 2010. En tant que principal auteur, j'ai réalisé les principales étapes d'initiation de l'étude, de recherche, de calcul, d'interprétation et d'écriture du manuscrit en collaboration avec les autres membres de l'équipe impliqués dans l'étude et de Michel Souchet de l'entreprise Harmonic Pharma. Dans cette étude, nous nous sommes employés à évaluer l'efficacité et le coût de 3 méthodes de criblage. L'une est basée sur l'utilisation d'un entonnoir de criblage permettant d'aller des méthodes les plus rapides à celles les plus consommatrices en temps de calculs tout en filtrant progressivement la base de molécules à cribler. La deuxième méthode est basée sur l'utilisation des pharmacophores 3D définis à partir des interactions entre la cible et un ligand actif connu. La dernière méthode est l'utilisation de la connaissance extraite de la base de données P3LI comme des contraintes permettant de filtrer une base de données de molécules. Ces méthodes de criblage ont été testées sur un jeu de ligands de la base de données ZINC avec trois structures du récepteur d'hormone LXRß.

Le criblage virtuel basé sur la structure de la cible qui utilise des algorithmes d'amarrage moléculaire est devenu un des outils les plus utilisés dans les processus de découverte de nouvelles molécules thérapeutiques. D'énormes progrès ont été réalisés permettant d'appliquer ces outils à plusieurs cibles biologiques. Dans les premières phases de recherche de nouveaux médicaments, l'approche standard est de tester tous les composés disponibles dans l'entreprise ou dans le commerce sur la ou les cibles biologiques considérées avec le ou les algorithmes d'amarrage moléculaire disponibles. Cependant, avec le nombre croissant de structures cristallographiques aujourd'hui disponibles, une telle approche est très coûteuse en temps de calcul et en argent. C'est pourquoi le prétraitement des bases de données de ligands afin de définir les composés prioritaires pour les algorithmes complexes d'amarrage moléculaire est une étape importante qui peut permettre d'optimiser le CVHD. Ainsi, après avoir éliminé par des méthodes de chemoinformatique les molécules qui n'ont pas des propriétés de médicaments potentiels, différentes méthodes peuvent être utilisées afin de filtrer les molécules non éliminées.

Le but de ce chapitre est d'abord de décrire l'utilisation des connaissances provenant de la base de données P3LI comme une méthode de filtrage de larges bases de données pouvant s'insérer dans un entonnoir de CVHD multi étapes afin de permettre d'optimiser le CVHD, de

la comparer ensuite à deux autres méthodes basées sur la structure. La première est VSM-G, méthode qui est développée au sein de notre laboratoire et qui utilise un entonnoir de criblage virtuel multi étapes afin de restreindre les molécules qui seront soumises aux algorithmes d'amarrage moléculaire. La seconde méthode est basée sur les pharmacophores 3D permettant de caractériser la structure cristallographique du complexe protéine-ligand. Celle-ci sera utilisée pour restreindre la base de molécules à celles qui respectent un certain nombre de propriétés décrites par le pharmacophore 3D. La troisième méthode est l'utilisation de l'approche KDD comme un filtre pour le CVHD. Ainsi, les connaissances découvertes par l'approche KDD à partir des données biologiques, physico-chimiques, structurales sur la protéine et ses ligands potentiels seront utilisées pour filtrer une nouvelle collection de molécules. Ces trois méthodes de criblage ont été testées sur un jeu de ligands de la base de données ZINC avec trois structures du récepteur d'hormone LXRß et, dans certain cas, plusieurs conformations des ligands.

II. Comparison of three pre-processing filters efficiency in virtual screening: Identification of new putative LXRß regulators as a test case

III. Commentaires

Ce chapitre nous a permis de décrire la comparaison de trois méthodes permettant d'optimiser le criblage virtuel des molécules provenant de la base de données ZINC, afin de déterminer celles potentiellement actives sur la cible biologique LXRβ. Nous avons essayé de mettre sur pied une méthode optimale de pré-traitement des données et d'optimiser l'enrichissement en molécules actives par les algorithmes de CVHD, en essayant trois approches différentes. Dans la mise en place d'un protocole de CVHD, la technique utilisée doit nécessiter une faible quantité en CPU et en ressources, ainsi qu'une base de données de taille physique optimale. Cette étude est un instantané de certaines des techniques disponibles les plus populaires utilisées dans le traitement des grandes quantités de molécules. Tous les programmes utilisés l'ont été avec des paramètres par défaut. L'ajustement de certains de ces paramètres peut permettre d'augmenter les taux d'enrichissement observés.

Le travail décrit dans ce chapitre a été une expérience utile dans la recherche des procédés permettant l'optimisation des méthodes de CVHD. Elle a permis de démontrer que, dans le cadre d'un CVHD des molécules d'une large base de données sur la protéine LXRβ, le KDD constitue une méthode à part entière de sélection de composés potentiellement actifs sur la cible biologique, aussi efficace que les autres méthodes (VSM-G, pharmacophore 3D) dont l'efficacité a déjà été démontrée. Ce qui fait qu'elle peut parfaitement s'insérer dans un entonnoir de CVHD multi-étapes comme l'une des composantes de l'entonnoir. Étant donné que chacune des méthodes de filtrage décrites dans ce chapitre a ses propres mérites et limites, ceci permettrait de combiner les avantages des approches différentes et devrait fournir une façon de surmonter les limites diverses rencontrées par chacun de ces algorithmes de CVHD. Ceci permettrait de réduire les coûts et les taux de faux-positifs qui sont actuellement les obstacles principaux au CVHD. Les résultats présentés ici ont des implications importantes pour tous ceux qui voudraient s'engager dans des expériences de CVHD. Avec la quantité croissante d'algorithmes académiques et commerciaux de criblage virtuel disponibles pour les chercheurs, le choix d'un algorithme peut avoir un impact très significatif dans l'expérience considérée. Ainsi, l'étude réalisée nous montre que le choix de l'algorithme est un paramètre important dans un processus de criblage virtuel, qui peut avoir des conséquences significatives dans les résultats. Celui-ci doit être adapté aux données à traiter. On constate aussi que le criblage virtuel peut être amélioré par la mise en place des protocoles ou par

l'utilisation de ressources permettant de réduire le coût et le temps des calculs, tout en ayant un fort taux d'enrichissement des molécules sélectionnées.

L'utilisation des trois méthodes de criblage virtuel décrites plus haut a permis d'évaluer le coût et l'efficacité de chacune de ces méthodes seules, ou en combinaison, lors de la recherche d'affinités de millions de molécules sur plusieurs conformations d'une cible biologique. Cette approche a conduit à l'identification de composés prometteurs pour les essais in vitro.

IV. Annexes

Annexes I: 46 Consensus compound

« L'apparition des résistances aux ARV devient un problème si crucial qu'il est considéré désormais comme un objectif prioritaire de recherche et d'action de la lutte anti-sida. »

Pillay et al 2007

CHAPITRE 5 - Application 2 : Plateforme bioinformatique centrée sur une base de données pour l'étude des résistances du VIH aux ARV

Ce chapitre présente la conception d'une plateforme bioinformatique centrée sur une base de données HIV-PDI qui permette de mettre en relation les données de patients infectés par le VIH présentant des résistances aux ARV avec les modifications que cela induit sur la structure de la protéase et les implications sur la thérapie. Cette plateforme bioinformatique sera utilisée pour l'analyse et le traitement des résistances du HIV aux ARV. Cette étude a été réalisée en partenariat avec le CIRCB au Cameroun dans le cadre du Challenge Exploration grant N° 52034 (Round I) financé par la fondation Bill & Melinda Gates.

Ce chapitre a été soumis pour publication.

HIV-PDI (Protein Drug Interactions) database: An integrated resource for studying HIV drug resistance based on 3D Protein-Drug interactions

Ghemtio L[1], Smaïl-Tabbone M[1], Djikeng A[2#,] Devignes MD[1], Keminse L[3], Ndiaye B[1], Petronin F[1], Fokam J[3], Maigret B[1], Ouwe-Missi-Oukem-Boyer O[3]

1: Nancy Université, LORIA, Groupe ORPAILLEUR, Campus scientifique, BP 239, 54506 Vandœuvre-lès-Nancy Cedex, France
2: The J. Craig Venter Institute (JCVI), 9712 Medical Center Drive, Rockville, MD 20850, USA.
3: Centre International de Référence Chantal Biya (CIRCB) pour la Recherche sur la Prévention et la Prise en charge du VIH/SIDA, BP 3077, Yaoundé, Cameroun
#Current address: Biosciences eastern and central Africa - International Livestock Research Institute (BecA ILRI) Hub, P.O. Box 30709, Nairobi, Kenya
BMC Medical Genomics 2010

Sommaire

I. Contexte

La présente étude a été soumise pour publication. En tant que auteur principal, j'ai réalisé les principales étapes d'initiation de l'étude, de recherche, de calcul, d'interprétation et d'écriture du manuscrit en collaboration avec les autres de membres de l'équipe impliqués dans l'étude. Elle a été réalisée en partenariat avec le CIRCB au Cameroun dans le cadre du Challenge Exploration grant N° 52034 (Round I) financé par la fondation Bill & Melinda Gates et le Dr Appolinaire Djikeng du J. Craig Venter Institute. L'objectif général de ce projet est de mettre sur pied une plateforme de gestion, de traitement et d'analyse de l'information biologique, clinique et épidémiologique des patients atteints de VIH et naïfs de tout traitement ou sous traitement ARV, en utilisant autour d'une base de données, les méthodes de bioinformatique, d'extraction de connaissances, d'étude statistique et de modélisation moléculaire pour étudier les interactions entre les ARV et les séquences mutées ou non du VIH.

L'apparition des résistances aux ARV devient un problème si crucial qu'il est considéré désormais comme un objectif prioritaire de recherche et d'action de la lutte anti-sida. Ces résistances sont dues principalement à des mutations affectant le patrimoine génétique du virus et, en particulier, les gènes ciblés par les ARV, tout particulièrement ceux de la transcriptase inverse et de la protéase virale. Ceci pose déjà de sérieux problèmes de santé publique en Occident où les souches résistantes aux ARV sont responsables de 11% des nouvelles infections, les pays en voie de développement, et tout particulièrement les pays d'Afrique, ne sont pas épargnés par ce phénomène comme le montrent plusieurs études récentes. Les perspectives moléculaires et cliniques de lutte contre ces résistances montrent clairement le besoin d'identifier, le plut tôt possible, la présence de résistances pour choisir le médicament le plus adapté. Cela évitera ainsi la diffusion de sous-types viraux et virulents et permettra de développer de nouvelles substances et des méthodes permettant d'anticiper les évolutions futures du virus suite à de nouveaux traitements.

Cette étude s'intègre dans un projet plus général appelé P3LI (Protéine-Protéine et Protéine-Ligand Interactions) permettant d'établir des relations séquence-structure-fonction pour des classes particulières de protéines. Il est important de souligner que le projet P3LI est déjà soutenu par la Région Lorraine dans le cadre du contrat de plan Etat-Region 2007-2013*.

Des plateformes intégratives, combinant les données sur les séquences avec les informations cliniques et/ou épidémiologiques, sont déjà en développement dans plusieurs

*http://bioinfo.loria.fr

centres de recherche de pays émergents. Cependant, aucun de ces outils ne met en évidence l'effet des mutations connues sur la structure des cibles protéiques. La base de données Los Alamos (http://www.hiv.lanl.gov/content/index), qui inventorie les mutations des gènes du VIH conférant des résistances aux ARV, commence à peine à s'intéresser à l'aspect tridimensionnel de leurs interactions avec les ARV. De plus, cette base de données effectue la cartographie des mutations seulement sur deux des principaux gènes cibles des ARV que sont la protéase et la transcriptase inverse. De même, la base de Stanford (http://hivdb.stanford.edu/index.html) n'a pas encore intégré l'effet des mutations au niveau tridimensionnel bien qu'elle recense déjà les mutations de toutes les protéines cibles actuelles, y compris les plus récentes comme l'intégrase et la protéine d'enveloppe.

Ainsi, le but de ce chapitre est l'établissement d'une plateforme bioinformatique autour d'une base de données HIV-PDI (PDI : Protein Drug Interaction) qui permette de mettre en relation des données de patients infectés par le VIH et présentant des résistances aux ARV avec les modifications que cela induit sur la structure des enzymes du VIH et les implications sur la thérapie. Ceci permettra d'étudier l'effet des modifications de la structure tridimensionnelle des enzymes mutées sur l'interaction avec le ligand grâce aux observations obtenues par modélisation et dynamique moléculaire. Ceci nous permettra de proposer de nouvelles molécules à fort potentiel thérapeutique pour les nouvelles mutations observées.

II. HIV-PDI (Protein Drug Interactions) database: An integrated resource for studying HIV drug resistance based on 3D protein-drug interactions

HIV-PDI (Protein Drug Interactions) database: An integrated resource for studying HIV drug resistance based on 3D protein-drug interactions

Ghemtio L[1§], Smaïl-Tabbone M[1], Djikeng A[2#], Devignes MD[1], Keminse L[3], Ndiaye B[1], Petronin F[1], Fokam J[3], Maigret B[1], Ouwe-Missi-Oukem-Boyer O[3§]

[1]Nancy Université, LORIA, Groupe ORPAILLEUR, Campus scientifique, BP 239, 54506 Vandœuvre-lès-Nancy Cedex, France

[2]The J. Craig Venter Institute (JCVI), 9712 Medical Center Drive, Rockville, MD 20850, USA

[3]Centre International de Référence Chantal Biya (CIRCB) pour la Recherche sur la Prévention et la Prise en charge du VIH/SIDA, BP 3077, Yaoundé, Cameroun

[#]Current address: Biosciences eastern and central Africa - International Livestock Research Institute (BecA ILRI) Hub, P.O. Box 30709, Nairobi, Kenya

[§]Corresponding authors

Email addresses:

GL: leo.ghemtio@loria.fr

STM: Malika.Smail@loria.fr

DA: A.Djikeng@cgiar.org

DMD: Marie-Dominique.Devignes@loria.fr

KL: kermesse84@yahoo.fr

NB: Birama.Ndiaye@loria.fr

PF: petroninflorent@hotmail.fr

FJ: josephfokam@gmail.com

MB: bernard.maigret@loria.fr

OO: ooukem@yahoo.fr

Abstract

Background

Fighting against resistance to Antiretroviral drugs (ARVs) in HIV-infected patients is one of the major challenges today in AIDS research. It is now widely accepted that the knowledge on HIV genotypes present in patients presenting a resistance and their consequences on the binding of ARVs should help to understand, and further to overcome, HIV resistance to a given treatment. Therefore, identification of the critical interactions lost further to one or several HIV mutations, and consequently the modifications of other molecular factors, could be indicators to propose appropriate ARVs escaping the resistance.

Results

In this respect, besides the usual data found in most databases concerning mostly patients, clinical, biological and chemical information, adding the detailed knowledge of three-dimensional structures of protein-drug complexes and their interactions could help avoiding successfully the trap of resistance. This paper introduces the HIV-PDI (Protein-Drug Interactions) database that has been designed to provide such an integrated and robust framework able to collect most of the abundant and ever-growing data related to HIV drug resistance and to be used uses for physicians and biologist's decision making concerning the most appropriate treatment in front of resistant patients.

Conclusion

HIV-PDI includes clinical information about patients, resistance to given ARVs treatments, HIV proteins structures and mutations, HIV protein/ARV drugs and their three-dimensional interactions. In its present preliminary version, up to now limited to the protease data, the HIV-PDI database currently contains entries for 2029 patients (covering different treatment conditions), for 2540 protease target variants along with 5393 potential and approved drugs directed against this target. The database is coupled to visualization and analysis tools of 3D Protein-Drug interactions including data mining programs. Each entry can be retrieved through multiple methods including target name, drug name or function and drug therapeutic classification. The HIV-PDI database can be used in order to help understanding the appearance of resistance and further to promote novel drug and treatment developments, based on existing data about drug interactions.

Background

Although several antiretroviral (ARVs) drugs can currently target different HIV proteins in combination and treat HIV infected individuals, rapid emergence of resistance to ARVs has become a major obstacle to effective long-term treatment and management of HIV infections.[1, 2] Numerous retrospective and prospective studies have shown that the presence of a resistant HIV strain before initiation of ARV treatment could have an impact on the outcome of the therapy.[3, 4]. Thus, resistance to ARV drugs is increasingly becoming a crucial problem, and is now considered as a top priority for HIV/AIDS research and actions.[5, 6] These resistances documented so far are mainly due to mutations affecting the genetic makeup of the virus, and in particular the genes targeted by ARVs.[3, 4]

It is now acknowledged that resistance due to mutations implies modifications of structural factors that affect the binding of ligands to the active site of viral targets. In fact X-rays crystallographic studies have clearly shown that particular mutations on the target proteins affect the geometry of their active site, thus reducing the binding affinity of ARVs [7-10]. The structural bases of ARV resistance should be studied by carefully inspecting the interactions between the mutated target (HIV variant) and its ligand (the drug) at the interatomic level [11-13]. It seems therefore necessary to integrate such three-dimensional (3D) related information's within the data to be found in a database devoted to HIV studies. Hence it has become important to establish relations between the structural modifications affecting a viral protein following a given mutation, and the efficiency of ARVs targeting this protein. Such knowledge would allow the development of more effective molecules or the suggestion of adequate ARVs to be used in first or second line treatments [14, 15].

Several databases [16-27] have been developed for the collection and storage of information on known mutations, resistances, patients clinical data,additional metadata on HIV proteins and correlated information for AIDS treatment (see Table 1). However, despite the availability of a lot of information in these databases, there is a serious need for a comprehensive HIV/AIDS data repository. Such a repository should in addition contain information on structural modifications and 3D interactions between HIV proteins (drug targets) and ARVs with references on resistance or susceptibility to ARVs. The integration of such metadata would provide new insights into the molecular basis of HIV drug resistance. Consequently, such a database could open other avenues for the evaluation of new drugs that

are still able to target mutated HIV proteins with a relatively high affinity, thus providing new and more adapted treatments to a given case of HIV resistance to ARVs.

Database name	URL	Focus	Coverage
HIV Drug Resistance Database	http://hivdb.stanford.edu/	Evolutionary and drug-related sequence variation in the human immunodeficiency virus (HIV) reverse transcriptase (RT) and protease enzymes	Sequences on HIV-1 isolates from more than 7,000 individuals and from about 500 laboratory isolates containing mutations generated by virus passage or site-directed mutagenesis. About 20,000 drug susceptibility results from tests performed on more than 2,000 virus isolates.
HIV Sequence Database	http://hiv-web.lanl.gov/	All published and many unpublished HIV and related SIV DNA and translated amino acid sequences.	>95,000 annotated entries.
UK HIV Drug Resistance Database	http://www.hivrdb.org.uk/	Resistance tests performed as part of routine clinical care throughout the UK.	More than 51,000 test results. Most of these (around 90%) in the form of viral gene sequences.
IAS-USA International AIDS Society-USA	http://www.iasusa.org/resistance_mutations/mutations_figures.pdf	Survey of new data on HIV-1 drug resistance published or presented at recent scientific meetings to maintain a current list of mutations associated with antiretroviral drug resistance	More than 200 drug resistance mutations data.
HIV-1, Human Protein Interaction Database	http://www.ncbi.nlm.nih.gov/RefSeq/HIVInteractions/	The HIV-1, human protein interaction presented here are based on literature. This dataset is available in report pages per HIV-1 protein and is also integrated into Entrez Gene report pages for HIV-1 and human proteins.	More than 6,000 documenting interaction of HIV-1 proteins with those of the host cell is crucial to understanding the process of HIV-1 replication and pathogenesis.

Table 1: HIV Databases. More HIV Databases are available in Additional file 3: HIV Database List.

To design and implement an integrated HIV database for the management of HIV infection, treatment and drug resistance, our working hypotheses were that (1) for a given patient, resistance to a treatment is due to a loss of affinity of the delivered ARVs compound targeting HIV proteins, (2) this loss of affinity is the consequence of structural modifications and possibly loss of interaction within their active sites after mutations, and (3) restoring the decreasing affinities of a given treatment can be achieved by identifying and compensating the lost interactions.

In the present paper, we report the setting-up of a HIV-PDI (protein-drug interaction) database integrating all chemical, pharmacological, biological, clinical, epidemiological and structural data collected from HIV infected patients. In this preliminary issue of the HIV-PDI database, we focused on the HIV protease, which is currently one of the best characterized [16-19] HIV targets both in term of structure and function for HIV drug development and treatment [16, 20]. This new database was conceived to serve as an integrated resource for studying HIV drug resistance at the structural level of the protein-drug interaction, with a special emphasize on the active site of the HIV drug target.

Results and Discussion

HIV-PDI conceptual model

The variety of existing data and their relationships were integrated into a single Entity Relationship (ER) as illustrated on Figure 1 while the entire conceptual model describing the entities which were used in our database and their relations is depicted on Figure 2.

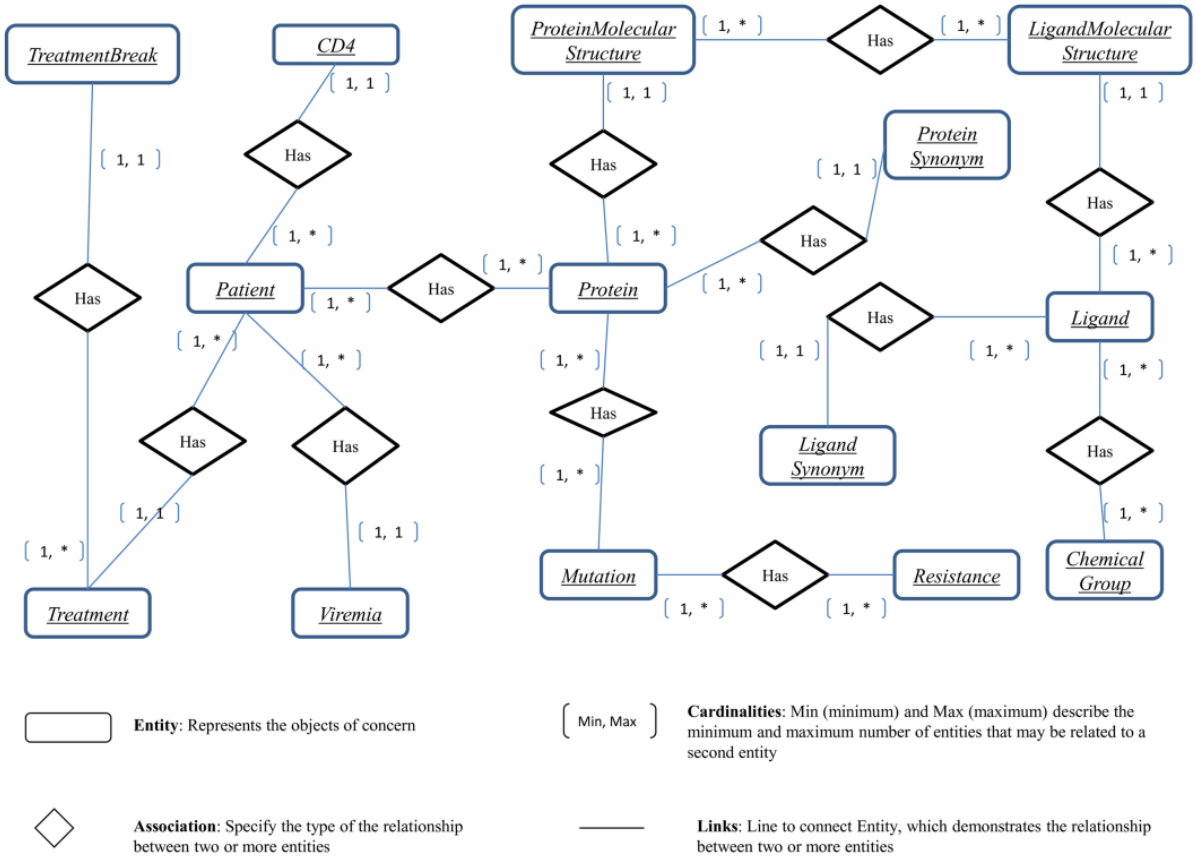


Fig. 1: ER of HIV-PDI database without Attributes to have a simplified overview of the diagram. The variety of existing data and their relationships were integrated into a single Entity Relationship (ER).

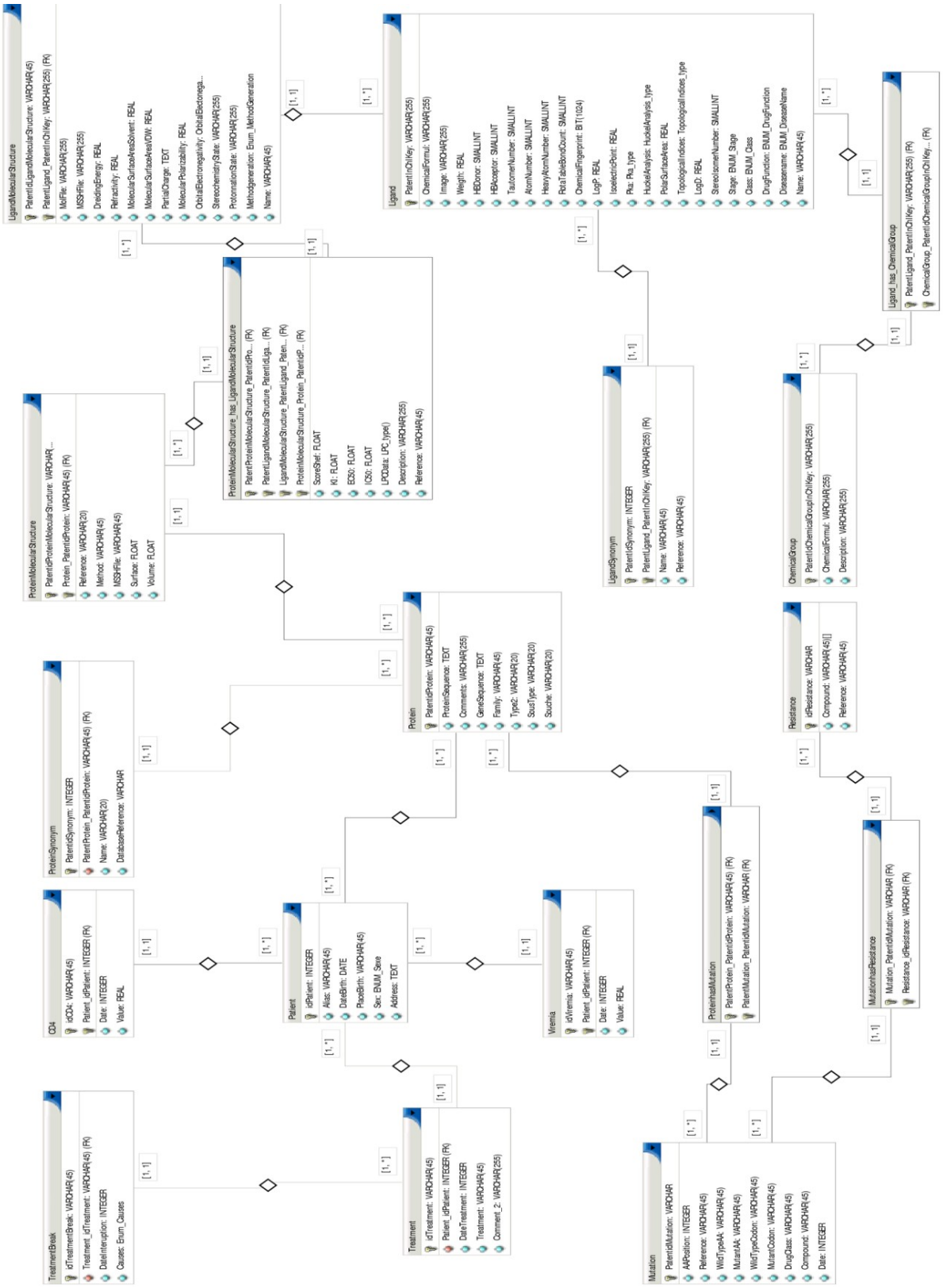


Fig. 2: Designed tables and their relationships in the functional relational HIV-PDI database. The entire model describing the entities which were used in our database and their relations.

The entity *Protein* represents a protein sequence (only the protease in the present version) WT or mutant found in a patient, and which has a given sequence at the time of sequencing. It will be identified by its unique key (specific to this database) and associated with an instance of the entity *Patient*. This association has a temporal descriptor since it is important for tracking the apparition and/or disappearance of such a specific mutation in a patient, taking into account its treatment or treatment break. The entity *Protein* is also associated with one or more instances of the entity *Mutation* since the sequence of a *Protein* instance may contain one or more mutations, and conversely a mutation can be related to several *Protein* instances.
The entity *Ligand* is associated with any ligand or drug and contains information characterizing the chemical and pharmaceutical properties of the compound.
To represent the structural aspects of proteins and drugs, the entity *Protein Molecular Structure* is associated with the entity *Ligand Molecular structure* in order to set the tridimensional structure of the protein/ligand complexes (PDB entry when available or molecular docking predictions), as well as the molecular interactions occurring to stabilize the complex. Other descriptors that characterize the complex like RMSD score of Protein-Ligand surface complementarities of harmonics surface and affinity data (EC50, KI, and IC50) are also available. Of course, one given protein can be associated with more than one protein conformer in order to consider protein flexibility and one given ligand can be associated with more than one ligand conformer in order to consider ligand flexibility.
The entity *Resistance* represents the fact that a ligand is no more interacting with the mutated variant of a viral protein here a protease. Each instance of this entity is thus associated with an instance of the entity *Mutation* which is itself documented by established observations found in other databases or provided by various health facilities participating in the collection of data from HIV patients. The relevant mutations can be obtained through the association between the entity *Protein* and the entity *Mutations*.
The entity *Patient* is associated with entities representing temporal clinical information such as the entity *Treatment* (drugs and regimen), *Viremia* and *CD4*.

An important aspect of our approach is that a conceptual data model was employed for communicating with domain experts when validating the database design. By transferring user requirements to a specific graphical data model, conceptual design allows to show and discuss all the data elements items and their relationships at a glance.

HIV-PDI current available records

Upon completion of the data entry in the HIV-PDI database, available records are summarized in the Table 2. These records are associated with the following main tables allowing a large body of queries:

Ligand table

Presently, this table contains 5,393 records corresponding each to a given compound. For example, a SQL query to the database, such as: "*SELECT patentinchikey, hbdonor, tautomernumber, atomnumber, logp, polarsurfacearea, stage, class, drugfunction, diseasename, name FROM ligand;*" will produce results that are partially represented in Additional file 4: Sample selection from the Ligand table.

Protein table

It contains 2,540 records. For example, a SQL query to the database, such as: "*SELECT patentidprotein, patient_idpatient, proteinsequence, genesequence, family, type2, souche FROM protein;*" will produce results that are partially represented in Additional file 5: Sample selection from the Protein table.

Patient table

It holds 2,029 records. SQL query to the database can be performed such as "*SELECT idpatient, alias, datebirth, placebirth, sex, address FROM patient;*" will produce results that are partially represented in Additional file 6: Sample selection from the Patient table.

Mutation table

It holds 16,193 records allowing SQL queries to the database such as "*SELECT patientidmutation, aaposition, reference, wildtypeaa, mutantaa, drugclass, compound FROM mutation;*".will produce results that are partially represented in Additional file 7: Sample selection from the Mutation table

Treatment table

It holds 13,629 records. As above, several SQL queries are possible, such as "*SELECT idtreatment, patient_idpatient, datetreatment, treatment, comment_2 FROM treatment;*" will produce results that are partially represented in Additional file 8: Sample selection from the Treatment table.

	Ligand	*Protein*	*Patient*	*Mutation*	*Viremia*	*Treatment*
Number of entries	5.393	2.540	2.029	16.193	28.447	13.629

Table 2: entries currently available in the HIV-PDI database

Database Graphic User Interface

The Graphic User Interface (GUI) (Fig. 3) was implemented to simplify access to the database by enabling physicians and biologists who are non-specialist users to send simple and logic queries through several graphical menus. The GUI was written mostly in Python and was implemented on a Linux server running Apache with the management system PostgreSQL. Currently the GUI is composed by different categories of data visualization interface. The main interface (Fig. 3) allows access to all data present in the database by several categories of request. All entities present in the database can be explored as a result of the request done in the main interface. The database can be explored directly by one request on one entity or kind of information like Mutation, Ligand, Patient, etc. Complex requests are also possible by using a combination of request on different entities. The complex request is validated, only when all the criteria selected to use in different entities for querying the database are all filled to compose a complex query. A complex request can be submitted to extract the information in the database according all items contained in the complex request. All entity levels are linked, which allows easy transition for example between i) associating structural information and patients, ii) protein sequences and therapies, and/or iii) drug resistance and treatment.

HIV-PDI example of use

In order to extract information and/or to find correlations between various data stored in the database, it is possible to perform several analyses from the GUI and to extract some information or knowledge by transforming the data into tables.

Cheminformatics analysis

Handling cheminformatics queries from the database, such as fetching ligands with a given substructure, searching or comparing molecules by their 3D shapes, chemical groups or functions could help the user to understand the role of chemical moieties in the problems of resistance. For that purpose, many conformation-independent molecular descriptors were pre-

calculated and can be accessed as mapped attributes directly from the GUI. For instance, substructure analysis was implemented at the chemical component level, which is directly linked to the ligand through a mapped attribute. Currently implemented in the GUI are the fundamental cheminformatics routines such as substructure searching, Tanimoto similarity (using fingerprints) calculations in the form of either overloaded or normal operators. For this purpose, in the Main Interface (Fig 3), they are search possibilities on ligand categories in order to upload the file containing the required properties to be used for further analyses in the database. For example it is possible to upload the chemical or 3D structure of any ligand as well as its InChIKey, or Chemaxon fingerprints. This allows to find all compounds similar to a given one in the database, or to identify all compounds that have been selected by chemical groups. In Figure 4 there is an example on how to find a ligand according to its molecular structure, to visualize as result (Fig 4A) the InChIKey of selected ligand ranked according to their level of similarities to the uploading molecular coordinates, and one InChIKey can be selected in order to visualize all data available in the database related to this particular InChiKey (Fig 4B). More advanced analysis can be carried out as well, such as fetching ligands with similar shape to the query ligand by uploading in the Main interface in the category Ligand the MSSH coefficient file of the query ligand.

Fig. 3: Example of a screenshot of HIV-PDI PHP-PostgreSQL interface designed namely. The Main Interface is used for request all data within the database. More graphical user interfaces are available in Additional file 2: All HIV-PDI graphical interfaces.

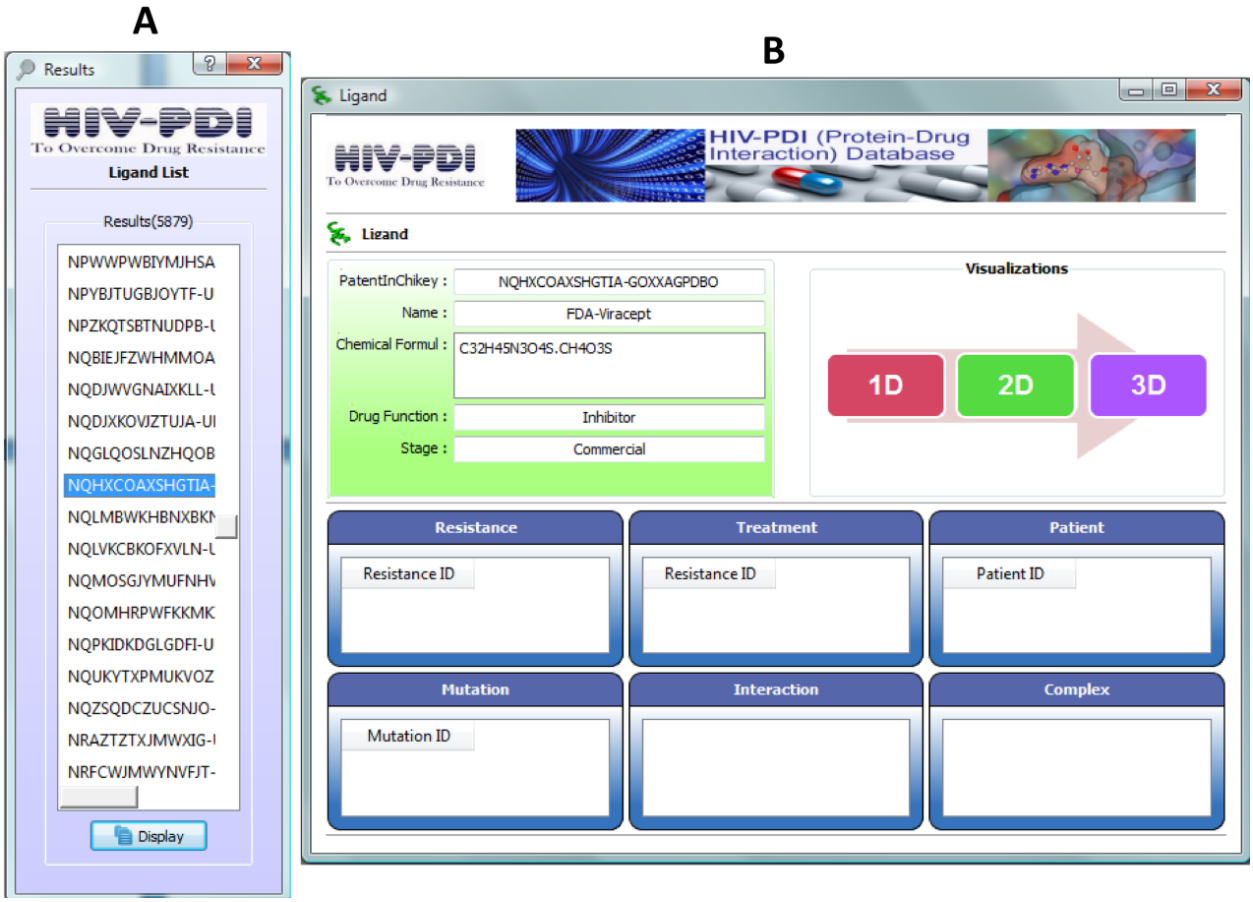


Fig. 4: Example of request chronology (A), (B), to extract and visualize a ligand according certain property: Uploading the file that contains the interesting property in Ligand category on Main Interface. (A): InChiKey of result ranked by their level of similarity. (B): Selection of one InChiKey to visualize all available properties of this ligand

Mutation analysis

The abstracts of all primary citations for mutant and WT PDB structures, which can provide valuable information, are stored in the HIV-PDI database as well and can be retrieved through the GUI. It will be also possible to check the influence of given mutations on the 3D structures of the target proteins, to compare these 3D structures and to highlight their differences. The GUI can also be used together with VMD (Visual Molecular Dynamics) for visualizing the mutated 3D structures and to characterize some aspects of their differences. Mutations from various other sources such as IAS-USA, Stanford, ANRS (Agence Nationale de Recherche sur le SIDA) and other public resources are implemented, i.e. mapped onto protein structures with the help of the sequence-to-structure mapping. These data could be used to investigate whether variations found in protein binding sites have the potential to disrupt or deteriorate ligand binding, thus leading to pathological status or drug resistance.

The HIV-PDI database can be used to identify the mutations or other molecular factor that are important in the sensibility decrease of the HIV protease to ARVs.

- HIV Protease structural information's

  The protease functions as a homodimer, which is composed of two identical 99 amino-acid chain, each chain containing the characteristic Asp-Thr-Gly active site sequence at position 25 to 27 (Fig 5) [17, 18]. The two subunits are linked by a four-stranded antiparallel β-sheet involving both the amino and the carboxyl termini of each subunit. Upon binding, both subunits form a long cleft where the catalytically important aspartic acids are located in a coplanar configuration on the floor of the cleft. In addition the protease contains a so called "flap structure" [21] in each subunit, an antiparallel β-hairpin with a β-turn that extends over the substrate binding site. Consistent structural differences are present between the bound and free states of the protein (Fig 6) [17, 18]. In all of the liganded forms, the flaps are pulled in toward the bottom of the active site ("closed" form), whereas the structures for the unbound protease all adopt a "semi open" conformation with the flaps shifted away from the dual Asp25-Thr26-Gly27 catalytic triads but still substantially closed over the active site and in contact with each other. The most populated states are the closed, semi-open and open. The nonflap residues show only slight variation. It is generally thought that most ligands, particularly the peptide substrate, can only access the active site through the open conformation.[17, 18]

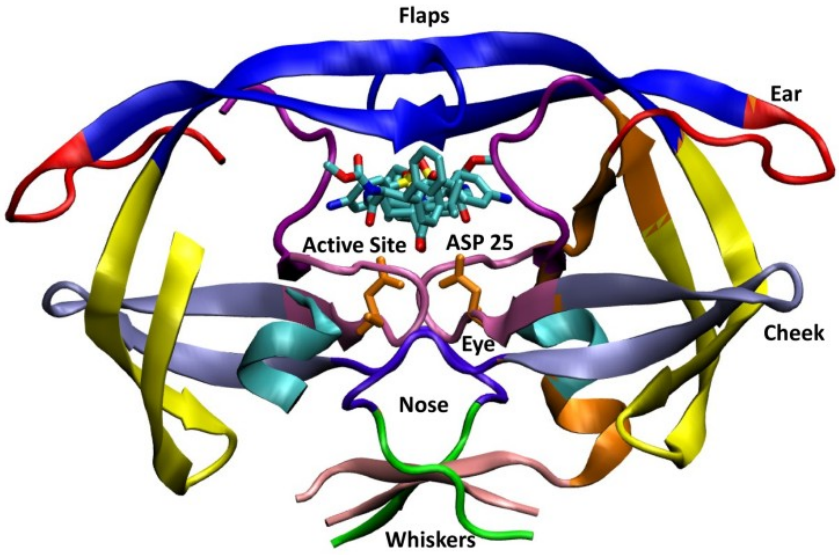


Fig. 5: Topology[21] of Crystal Structure of HIV-1 protease 2qmp. The convention for the terminology of the topology of HIV protease involves the following: Flap (43–58), Ear (35–42), Cheek (Cheek Turn 11–22 and Cheek Sheet 59–75), Eye (23–30), and Nose (6–10).

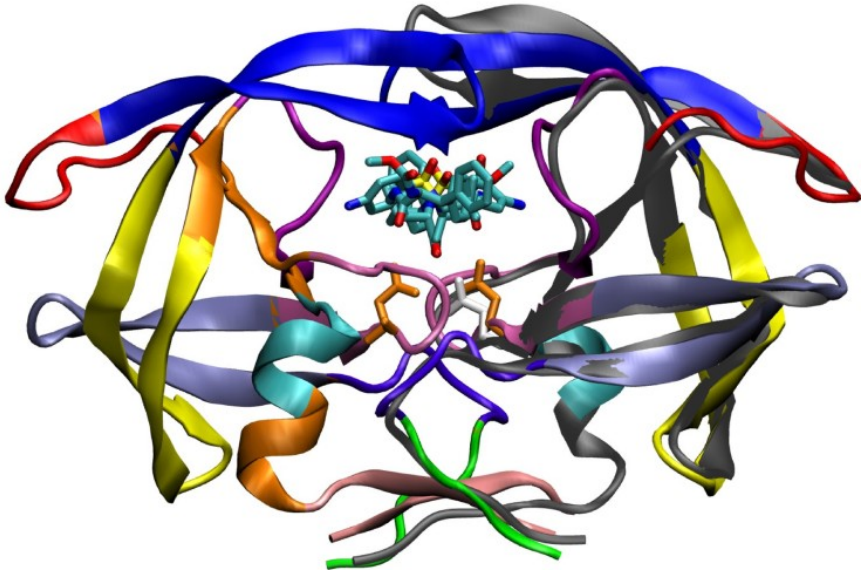


Fig. 6: Comparaison by superposition of wild-type apo protease (PDB id: 3phv) in grey and the wild-type (PDB id: 2qmp). The ASP 25 of apo form is in white, and not in the same place than the ASP 25 of complex form. The flaps of apo form are also more open than the flap of complex protease.

- HIV Protease mutation information's

  As detailed in Table 3, different types of mutations may occur in the protease. But all these mutations cannot be considered in the same manner [22-27] according to their localization within the protein structure. Therefore, the critical substitutions in or around the active site are firstly checked because their important resistance potential have been confirmed by several studies [22-27], while the role of secondary substitutions in other parts of the protein in the development of resistance is not completely understood. Some substitutions are known to be compensatory, or enhancing the replicative efficiency of the virus, and it is obvious that there is a striking overlap of resistance conferring mutations among most of the available protease inhibitors. The flap modification by mutation may modify the binding of the protease substrate to the active site and thereby inhibit activity of HIV protease inhibitors [18, 28, 29]. Various groups have identified anticorrelated motion between the flap and ear (residues 35-42)[21] regions through normal mode analysis and molecular dynamics simulations. Restricting movement of the ear region has been shown to concurrently limit the conformational sampling of the flap [18]. Therefore, as describing the effect of mutations on the protease structure could help to understand the resistance phenomenon, our analyses were focused firstly on the protease mutations already described as critical and on all the other ones that appear in the active site, flap and ear regions.

- Patient mutated HIV protease analyses

  The use of the HIV-PDI database is exampled here for a patient presenting a resistance revealed by virological failure (increased viremia) after a first line treatment with Amprenavir (APV) as antiprotease. The mutations at time=0 are illustrated in the 3D structure (constructed by homology modeling) of the HIV-1 sequence of this patient sequenced at time=0, and modifications of the 3D structure of HIV-1 protease due to these mutations are revealed by the comparison of the mutated protein with the crystal structure of HIV-1 protease WT (3ekv) complex with APV available in the PDB. Figure 7a shows the presence of several structural differences between the two structures: the flap of the mutant is more open than the flap of the WT structure, and the ear, whiskers, nose, eye and the two active sites are clearly different. The mutations are present on the flap, ear region, which are critical for the binding action of substrate. The presence of well known mutations (see Table 3), added to the fact that most of these mutations are located in the critical region for the substrate binding can explain the appearance of resistance. For

example, this hypothesis can be confirmed in Figure 7b, were the superposition of 3ekv crystal structure with the crystal structure of APV in complex with a drug resistant HIV-1 protease variant (I50L/A71V) is illustrated. The presence of mutation I50L on flap region provokes the modification of the conformation of Flap, with the modification of ear, and active site region. In this case the presence of the critical mutation I50L can explain the resistance.

	Position(s) in protease	
Drug	Critical substitutions	Secondary substitutions
Saquinavir	48, 90	10, 36, 63, 71
Ritonavir	82, 84	20, 36, 46, 54, 63, 71, 90
Indinavir	46, 82	10, 20, 24, 32, 54, 63, 71, 84, 90
Nelfinavir	30	46, 63, 71, 88, 90
Amprenavir	50	10, 46, 47

Drug = protease inhibitors

Table 3: HIV-1 resistance to protease inhibitors [22-25, 27]

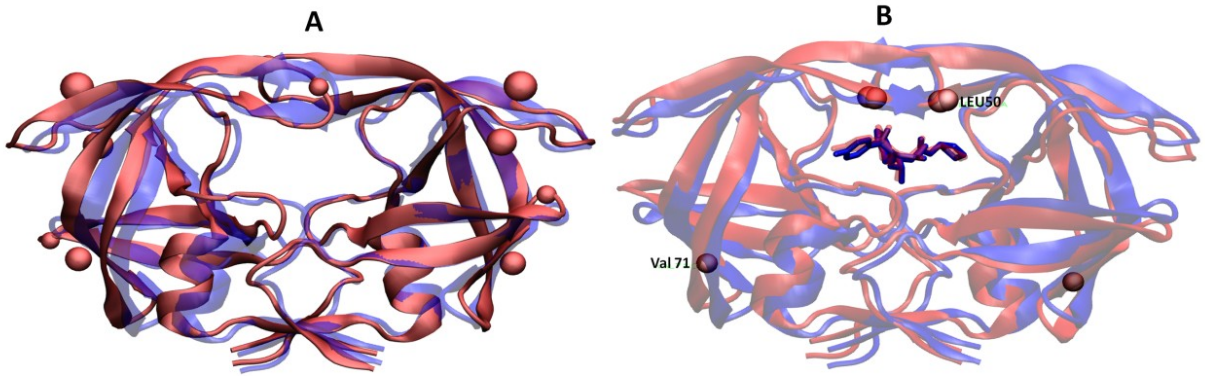


Fig. 7: Effect of mutation in a 3D structure of HIV-1 protease. (A): Superposition to reveal structural modifications by mutations on HIV-1 protease 3D structure of patient 39546 sequencing at time t=0 in HIV-PDI (in red and the mutations are represented by the ball) against crystal structure of Wild Type HIV-1 protease (PDB id: 3ekv in purple) available in PDB. (B): Superposition of I50L Drug-Resistant (in red and the mutations are represented by the ball) HIV-1 Protease Mutant (PDB id: 3em3) against crystal structure of Wild Type HIV-1 protease (PDB id: 3ekv in blue) available in PDB

HIV protease cavity analysis

The MSSH coefficient, cavity volume and cavity surface of each HIV-1 protease is stored in the HIV-PDI database. These data can be used to check the effect of mutations on the binding cavity characteristics, which could explain the absence of binding of a given drug in the active site of a mutated HIV-1 protease. An example of such kind of study is presented

here using the WT 3ekv PDB X-ray structure, the 3em3 mutated PDB X-ray structure and the patient 39546 homology-modeled structure of HIV-1 protease. For these three targets, we extracted from the HIV-PDI database the spherical harmonic coefficients, volume and surface data in order to analyze the impact of mutations on these parameters and to see if these data could be related to the lost of affinity by 3em3 and patient 39546 HIV-1 protease to APV. As illustrated in Figure 8, the binding site cavities are different in the three structures, showing that the patient mutated protease has the smallest cavity compared to the two others. Such a difference could explain the fact that the two mutated HIV-1 proteases do not have the same affinity to APV in comparison to the WT HIV-1 protease. This confirms that the use of spherical harmonic data for the analysis of binding site cavities could be a valuable tool for understanding mutations-induced resistance problems.

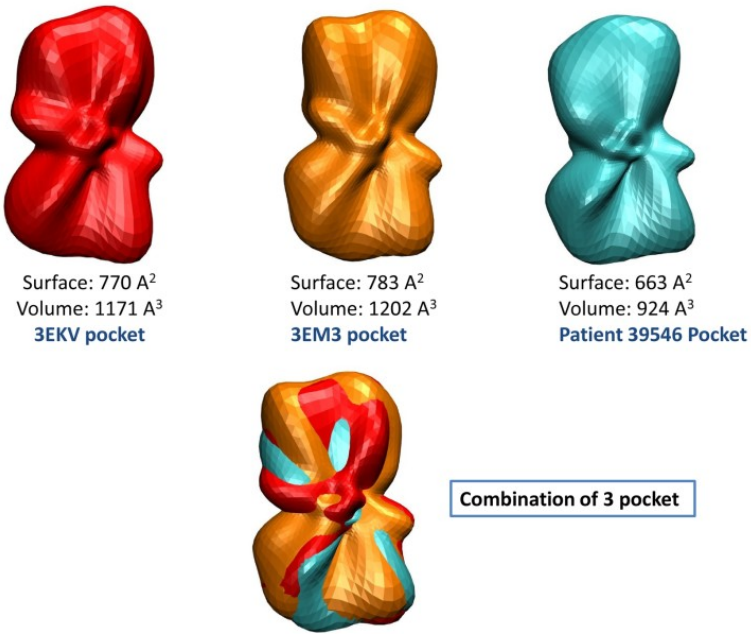


Fig. 8: Volume and surface representation. Volume and surface representation of the cavity of wild type HIV-1 protease (PDB id: 3ekv) and HIV-1 protease mutant (PDB id: 3em3) available in PDB against a patient 39546 HIV-1 protease 3D structure.

Protein/ligand interactions analysis

Another way to help understanding the resistance phenomena is to analyze, in the mutated proteins that appear during the infection, the loss of affinity for inhibitors by disrupting favorable binding interactions [10, 30] with important residues of the protease.

Such analyses would provide atomic details about the chemical groups and interaction types that contribute to the stability of the protein/ligands complexes.

From them, it is possible to check the differences in such interactions due to mutations. In fact, resistance is expected to be related to disturbances of the previously favorable ligand protein interaction in the binding pocket. Several recent papers have already highlighted the interest of such interaction data to understand the resistance problem [11, 31-46].

Such an analysis can be conducted by comparing the interaction differences between the WT HIV-1 protease complexes with APV (PDB code 3ekv) and the two mutant structures complexed with the same molecule and corresponding to the example presented in the above sections (the PDB crystal structure 3em3 and the structure obtained for the patient 39546 at time = 0). When we extract all the interactions available in the HIV-PDI database for these three structures according to residues Asp_{25}, Gly_{27}, Asp_{29}, Asp_{30}, Gly_{48}, Ile_{50}, which were identified as important for drug binding and its stability, we were able to clearly identify loss interactions due to the effect of mutation in the protease active site between the drugs and these important residues binding [10, 30] (see Interaction list in Additional file 1: Interaction list of WT, mutate and patient protease complexes with Amprenavir). Their analysis shows also that the number of interactions is different for the three structures. For example, in the two mutants, some interactions are missing between APV and the critical Asp_{25} residue in chain A and B and this loss of interaction could already explain the resistance of these 2 mutants to this drug. As detailed in Figure 9a and 9b, several types of interactions have been lost in the mutated complexes. These interactions include hydrophobic/hydrophilic, Van der Walls, hydrogen and polar bonds. They concerned residues of both the flap and the active site regions necessary to enable the protease inhibition [10, 30], therefore explaining the decrease of drug sensibility by mutated HIV-1 protease.

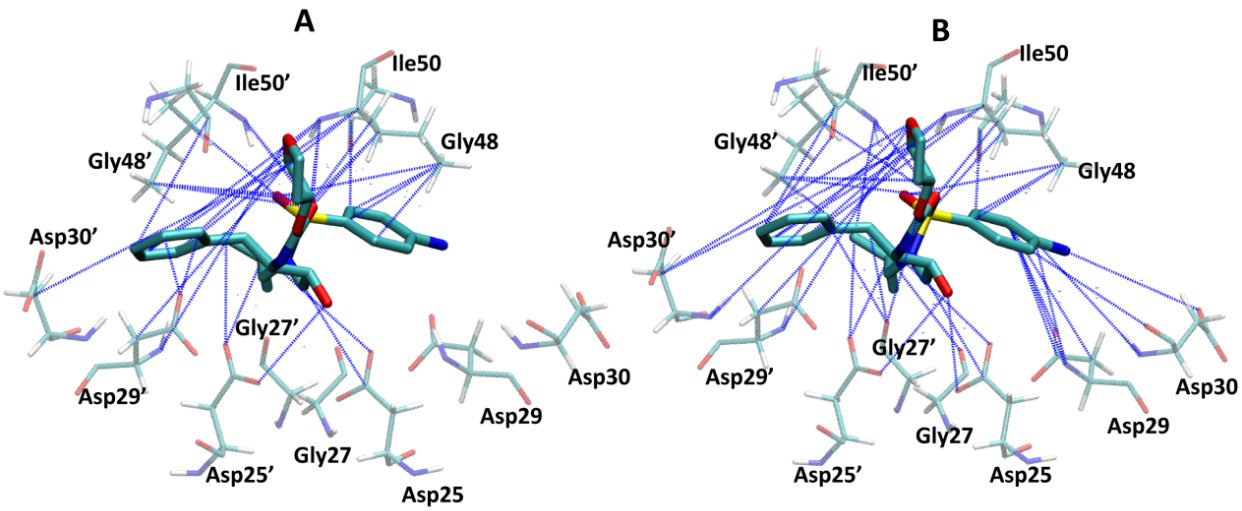


Fig 9: Visualization of interactions lost by mutated complex protease compared to Wild Type protease interaction between Amprenavir (APV, anti protease drug) and main important residues of protease (ASP 25, GLY 27, ASP 29, ASP 30, GLY 48, ILE 50)[10, 30] for drug binding. (A): Interactions lost in 3em3 complex with APV compared to 3ekv Wild Type in the active site of the Protease. (B): Interactions lost in structure of HIV-1 protease for patient 39546 complex with APV compared to 3ekv Wild Type in the active site of the Protease.

Resistance analysis

Drug resistances that were identified as the results of mutations stored in HIV-PDI database can be used to identify the treatment that induced a specific resistance, as well as, in certain cases, the treatments that allowed surmounting certain resistance in a patient. Such an analysis is illustrated in Figure 10 to identify the treatment which causes the resistance to APV in the patient 39546 at time=0 and what treatment can be used to help surmount the resistance. When clinical data are extracted from the HIV-PDI database for this patient, the treatment history with the HIV protein sequence and mutation (see Fig 10) can be drawn. As described in Figure 10, treatment history of a patient gives information about the fluctuation of viremia rate according to the treatment and apparition of mutations along of time.

Going back to the example of the patient 39546 (Fig 10), no treatment was taken at time=-7 (ie 7 week before the sequencing). The viral load was 2.3 copies of viral genome/ml and there was no mutation data. When the sequencing was done at time=0, the patient was put under a treatment which is the combination of APV, ritonavir, stavudine and lamivudine. The sequencing enabled identification of several mutations (L15V/E35D/R41K, I50L/V82L)

whereas the viral load reached 2.8 copies of viral genome/ml. At time=+12 the patient was still under the same treatment, but the viral load increased to 5.8 copies of viral genome/ml, and no new sequencing data was available. The increased viremia can be explained by the fact that at time 0, the patient had the critical mutation to APV (I50L), ritonavir (V82L) and other mutations on important region for substrate binding site. So during the 12 weeks, the resistance form of virus has increased and the patient has become resistant to the current treatment. This can be confirmed by the fact that, between 20 and 28 weeks after the change of first treatment by a new treatment (nelfinavir, zidovudine, lamivudine), the viral load droped from 5.8 to 1.9 copies of viral genome/ml. It's possible to conclude that the change of the initial treatment (amprenavir, ritonavir, stavudine and lamivudine) to a new treatment (nelfinavir, zidovudine, lamivudine) with mutations (L15V/E35D/R41K/I50L/V82L) could help to surmount resistance.

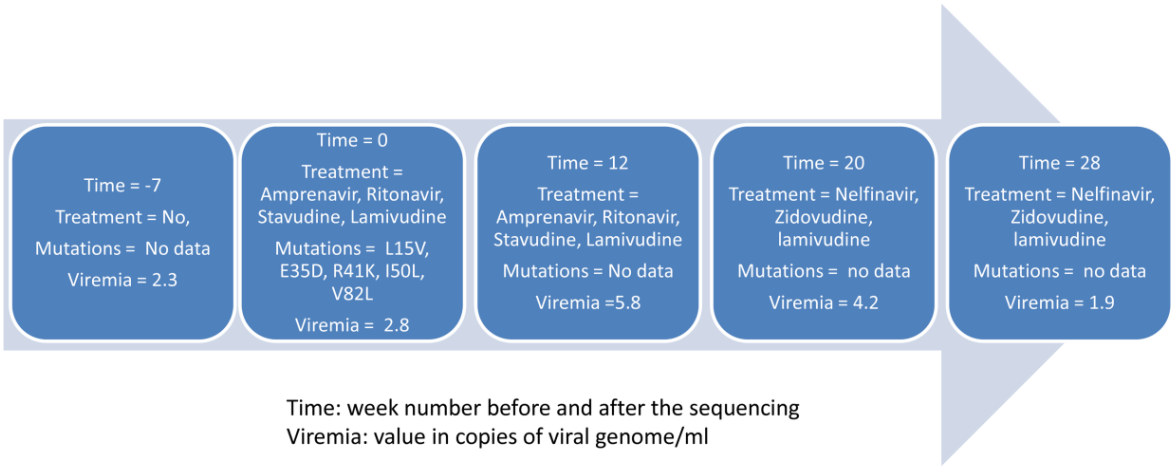


Fig. 10: Treatment history of patient 39546. The treatment history of a patient gives information about the fluctuation of viremia rate according to the treatment and apparition of mutations along of time.

Conclusions

In this paper, we designed and implemented a HIV-PDI database platform to address the problems of drug resistance with a central focus on the 3D structure of the target-drug interaction. Clinical and biological data, structural and physico-chemical information and 3D interaction data concerning the targets (HIV proteins) and the drugs (ARVs) were meticulously included in our database and combined with tools dedicated to study HIV mutations and their consequences on the efficacy of drugs. Specifically, the HIV-PDI

database is thereby useful as a system of data collection allowing interpretation on the basis of all available information, thus helping in possible decision-makings.

Tables of the developed relational database contain experimental measurements data, clinical data, and structural data between protease and anti-protease drugs and, in general, physico-chemical properties and molecular descriptors. These data represent a large body of possible searches within the HIV-PDI database in order to detect the possible structural elements characterizing the resistance phenomenon observed in a given patient.

Some examples of information or knowledge that were derived from the database are listed below:

- List all patient, protein, drug, mutations, resistance, 3D protein-drug complex, 3D interactions and all clinical data of a patient
- List all mutations and treatment which cause resistance
- Finding treatment which cause specific mutation
- Finding treatment which help to surmount specific resistance
- Finding the 3D interactions related to the loss cause the resistance of mutated protein.

The introduction of such 3D structural analysis about the protein/ligand complexes in a database related to HIV and the consequences of mutations on the stability of the protein/ligand complexes make the main difference with the others known databases (see Table 1).

Also those data consist in an infrastructure for much related information or knowledge. When those data are transformed into information or knowledge, there will be an opportunity for innovation which is the main purpose of research and experimentation. Table's interpretation outputs as information or knowledge depend on the level of usefulness of the outputs. It is possible to get a lot of new information that may be very hard to obtain without developed databases.

Possible advantages of the HIV-PDI database are listed below compared to the other known HIV database (see Table 1):

- It's a relational database
- Centralized in the same place different kinds of information and data
- It will be useful for researchers who want to adapt the treatment to overcome specific resistance

- Visualize and analyze the 3D structural modification due by the mutation
- Visualize the evolution of clinical data of patient

The HIV-PDI database provide for a patient, oriented queries, efficient follow up of HIV patients treatment, review of treatment outcome, documentation of drug resistance and the identification of treatment alternatives in cases of drug resistance leading to therapeutic failure. The ultimate goal is that an integrated bioinformatics platform will help clinicians/physicians to identify the consequence of mutations which causes the loss of affinity to delivered ARVs compound. Such information would be further used in the selection of new drugs based on their ability to compensate or restore the decreasing affinities with the drug targets. This information system can also be used for studies dedicated to the design and/or identification of new molecules or ARVs able to circumvent the increasing HIV drug resistance.

Future work will include adding the data for all other HIV protein target family and drugs, more chemoinformatics modules, such as the shape analysis functionality will be greatly improved since geometrical and physicochemical features can be compared at the same time. Hence, future work will include clustering of ligands using the shape descriptors together with target information and activity classes from drug databases to predict putative targets for new small molecule structures [47]. We will develop statistical model or knowledge database extraction algorithm that predicts virological response or mutation/resistance to therapy from molecular interaction, HIV-1 genotype and other clinical information [10, 11, 48]. Structural alignments of ligand-binding protein domains are could be add in HIV-PDI. In future, these alignments could be used to predict functional residues in homologous structures or for modeling, particularly modeling of binding sites, characterization and structural analysis of HIV-1 conservation[49]. In addition, ligand interaction fingerprints could be created to cluster ligands or simply to detect conserved 3D interactions with protein residues.

Methods

Database design

According to physicians and biologists' requirements, we had to provide a centralized way to store experimental and theoretical results obtained from patient's samples, proteins and drugs. This storage had to offer a unified access to information (HIV protein sequence and structure, drug structure, 3D interactions of protein-drug and their physico-chemical

properties and biological properties, patient information with its clinical data like CD4 “dosage of CD4 epitopes”, Viremia “level of virus particles in blood”, Mutation, Resistance, Treatment, and Treatment Break) enclosed in various textual sources (scientific articles, patents, etc.), HIV databases (see Table 1), or other databases (Protein Data Bank (PDB) [50], Swiss-Prot [51], GenBank [52], European Molecular Biology Laboratory (EMBL) [53] data sources, etc..). Also, we had to connect, complete, and update the existing sources of information. There is a rich sample of experimental, clinical, theoretical and computational resources stemming from various scientific domains such as biology, genomics, biochemistry, bio-computing, and pharmacology, which were conceived to support research on HIV proteins and ligands. The HIV-PDI database had to give access to these available resources via a simple interface, in avoiding information redundancy. The sources of information had to be stored in the HIV-PDI database to facilitate its traceability. In addition, the HIV-PDI database had to supply experimental and computational results rather than interpreted data. Due to the fact that the discrepancies in interpretations of data are often subject to discussions until a consensus is adopted, the HIV-PDI database had to store the experimental and computational results obtained in referenced scientific publications as well as the description of the method used to produce the data. In case of contradictory data, all experimental results should be reported with the accurate references so that interpretation of data is left under the user’s responsibility. Finally, the HIV-PDI database should give access to simple queries such as information or data on ligands, ligand’s conformations, ligand chemical group, proteins, proteins conformations, protein-ligand complexes and 3D interactions types, patients information and clinical data (CD4, Viremia, Mutation, Resistance, Treatment, and Treatment Break) or to complex queries like selection of ligands that have one special chemical group, treatment taken by one or several patients, mutations found in one or several patients, resistances observed in one or several patients. The HIV-PDI database should also give access to even more complex queries including the selection of proteins or/and ligands that are in complex and presenting one or several types of interactions, the selection of patient’s treatments that induce mutation and whose mutation has lead to resistance, the identification of treatments that have allowed to surmount resistance to previous treatments, and the visualization of the 3D interactions of a mutated protein-drug complex compared to a wild type protein-drug complex.

The HIV-PDI database will be extensively used for similarity searching, diversity/similarity analysis of compound libraries. Clinical data, mutation, resistance and molecular descriptors [54-57] of proteins and drugs stored in the HIV-PDI database will also be used for the development of statistical models such as Quantitative Structure Activity Relationships (QSAR) [58-61]/Quantitative Structure-Property Relationships (QSPR) [60, 61] and data mining analysis such as artificial Neural Networks (NN) [60, 62]. These models can be used for the prediction of receptors binding, treatment, resistance or properties of compounds from their molecular descriptors or for defining rules for virtual screening [63, 64]. All these analyses above could help us for chemoinformatics, mutations, resistances, and 3D interactions study on HIV protease-drug.

The HIV-PDI database was designed with the relational database model so as to take into account the complexity of the data anticipated to be entered and extracted in search for correlations between the recorded data from patients, physico-chemical, structural and biological data. We therefore used the classical methodology suitable for the design of a relational database [65-67]. The Entity Relationship (ER) approach was chosen to data requirements specification and conceptual modeling of the HIV-PDI database. The ER of HIV-PDI database has then been translated into a set of relational tables, which were normalized by removing redundancies and prevented from data anomalies. This entire phase was automatically performed through the use of free available Computer-Aided Software Engineering (CASE) tools DB Designer [68].

Data description and sources

Drugs

- Drug information

  Ligand: This table was constructed with drug IUPAC (International Union of Pure and Applied Chemistry) International Chemical Identifier (InChiKey) [69] and several 1D/2D molecular descriptors (name, chemical formula, pka, hydrogen bond donor, etc.). Ligand molecular descriptors represent structural and physicochemical features of compounds. InChiKey is a textual identifier for chemical substances, designed to provide a standard and human-readable way to encode molecular information and to facilitate such information's search in databases and on the web.

LigandMolecularStructure: This table is used to store 3D molecular descriptors of drug structure such as 3D molecular coordinates, molecular surface spherical harmonic coefficients [70-72], molecular polarizability, stereochemistry state, etc.

ChemicalGroup: This table is used to store chemical group of all drug compounds.

LigandSynonym: This table stores all known existing names of a given drug.

- Drug sources

  We have used four main sources to collect data about compounds defined as HIV protease inhibitors: i) commercial drugs approved by the US FDA (Food and Drug Administration), ii) compounds in clinical phases, iii) compounds from Life Chemical [73] database and defined as HIV like compounds and iv) compounds described in articles published in various international peer-reviewed journals. The collected structures were stored in the HIV-PDI database and for each molecule they were used to generate multiple conformers and to compute a set of molecular descriptors. Conformational sampling was performed and stored for each compound in the database using the Omega [74] software. Drugs information is computed from structures collected and stored in the HIV-PDI database by Chemaxon [75] and InChI [69] packages tools

Proteins

- Protein information

  Protein: This table is used to store the 1D/2D descriptors, sequences and known information on HIV protease proteins such as protein sequences, gene sequences family, type, etc.

  ProteinMolecularStructure: This table stores 3D molecular descriptors of proteins such as 3D coordinates, pocket surface, molecular surface of spherical harmonic coefficients, etc.

  ProteinSynonym: This table stores all known existing names of a given protein.

- Proteins sources

  Swiss-Prot and PDB were the main sources used for extraction of protein data when available. The wild type (WT) HIV-1 protease sequence was collected from the Stanford HIVdb (Swiss-Prot [51] identifier of WT is Q9WFL7, and PDB identifier of WT is 2qmp) where it has been adopted as the most collectively accepted reference for comparison with all the mutated sequences.

  When for some mutations no experimental X-ray structure was available, 3D models of the mutated protein were computed with the Modeler homology program. [76, 77] These

models were used as starting points for sampling the target conformational space in order to take into account the protein flexibility. For that purpose, after placing the protein in proper explicit water bath, short 1 ns. Molecular Dynamics (MD) simulations were performed using the NAMD MD software.[78] For each protein we therefore obtained a set of conformational states to be used later for analysis of the mutation consequences on the structure and on the ligand bindings. The Molecular Surface Spherical Harmonic (MSSH) program was used to compute MSSH coefficient, volume and surface of protein active site. [70, 79]

Protein-Drug Interactions

- Protein-Drug interaction information

  ProteinMolecularStructure_has_LigandMolecularStructure: This table stores 3D molecular descriptors of complex proteins-drug such as lists of 3D interactions, Root Mean Score Deviation (RMSD) [70] of spherical harmonic coefficients, affinity data (Median Effective Concentration "EC50", receptor affinity "KI", and Median Inhibition Concentration "IC50") between protein and drug.

- Protein-Drug interaction sources

  Data on protein-drug interactions were obtained from the 3D structures of HIV protease-drug complexes between proteases and ligands stored in the database. For compounds without an experimental protease 3D structure complex found in the PDB database, we have used the flexible docking [63, 64, 80] program Glide [81] to generate a 3D model of the protease-drug complex. Activity of protein-drug complexes was extracted from the scientific publications. The list of 3D interactions was extracted with the LPC [82] software from Protein-Drug complex file. The RMSD [70] of spherical harmonic coefficients was computed between MSSH (Molecular Surface Spherical Harmonic) coefficient of protein and drug by SHEF (Spherical Harmonic coEfficient Filter) program. [79]

Patients

- Patient information

  Patient: This table stores all information extracted from patients for instance the name, the sex, the alias, etc.

CD4: This table is used to store data (date, value in cells/µl, etc.) on the CD4 T Lymphocytes count of patients.

Viremia: This table is used to store data (date, value in copies of viral genome/ml, etc.) on the plasma viral load of patients.

Treatment and *Treatment Break*: These tables were used to store data (date of treatment, treatment specificity, interruption schedules and reasons, etc.) on the patient's treatment.

Mutation: This table is used to store data (amino acid position, mutated amino acid, WT amino acid, mutated codons, etc.) on the patient's viral mutation(s).

Resistance: This table is used to store resistance data on the patient's HIV resistance.

- Patient sources

  Clinical data on HIV infected patients was extracted from the Stanford HIVdb. The patient's identification number connects these data all together. List of known mutations were extracted from Stanford HIVdb, Los Alamos, Swiss-Prot, PDB, and International AIDS Society-USA (IAS-USA)[83] databases (see Table 1) and free available bibliographic. [22, 23] The whole sequences of mutated proteins were deducted by comparison against the protease WT reference sequence.

Database implementation

The relational database was implemented in PostgreSQL [84] (Version 8.3.7) on the GNU/Linux x86 64 bits operating system. Data collection from remote data sources and integration were performed with wrappers developed as python scripts. The Structured Query Language (SQL) [65, 85] was used to access to the data in the HIV-PDI relational database. It allows users to describe the data that they wish to see. SQL also allows users to define the data in a database, and to manipulate these data. As everybody not used SQL and as the various analyses will not can beings satisfied by SQL only, then a graphical user interface including query forms and visualization tools was designed using python and software library QT designer. [86]

Authors' contributions

All authors have jointly developed the research concept and collaborated on the writing of the manuscript. As the main author LG has initiated the study, carried out the computational

analyses, has interpreted the results, and drafted the manuscript. All authors revised the manuscript and approved its final version.

Acknowledgements

We thank The Bill & Melinda Gates Foundation for their financial support through the Grand Challenge Exploration grant N° 52034 (Round I). Ghemtio Léo was supported by grants from INRIA (Institut National de Recherche en Informatique et en Automatique), CNRS (Centre National pour la Recherche Scientifique) and The Bill & Melinda Gates Foundation. Keminse Lionel and Fokam Joseph were supported by grant from The Bill & Melinda Gates Foundation. Ndiaye Birama was supported by grants from INRIA. Petronin Florent was supported by grants from CNRS. We thank Openeye and Chemaxon for providing free access to their software according to an academic license. This work was supported in part by Region Lorraine within the framework of the PRST MISN (MBI operation).

References

1. Dau B, Holodniy M: **Novel targets for antiretroviral therapy: clinical progress to date.** *Drugs* 2009, **69:**31-50.
2. Temesgen Z, Warnke D, Kasten MJ: **Current status of antiretroviral therapy.** *Expert Opin Pharmacother* 2006, **7:**1541-1554.
3. Paar C, Palmetshofer C, Flieger K, Geit M, Kaiser R, Stekel H, Berg J: **Genotypic antiretroviral resistance testing for human immunodeficiency virus type 1 integrase inhibitors by use of the TruGene sequencing system.** *J Clin Microbiol* 2008, **46:**4087-4090.
4. Shafer RW: **Genotypic testing for human immunodeficiency virus type 1 drug resistance.** *Clin Microbiol Rev* 2002, **15:**247-277.
5. Mascolini M, Larder BA, Boucher CA, Richman DD, Mellors JW: **Broad advances in understanding HIV resistance to antiretrovirals: report on the XVII International HIV Drug Resistance Workshop.** *Antivir Ther* 2008, **13:**1097-1113.
6. Pillay D: **The priorities for antiviral drug resistance surveillance and research.** *J Antimicrob Chemother* 2007, **60 Suppl 1:**i57-58.
7. Garriga C, Perez-Elias MJ, Delgado R, Ruiz L, Najera R, Pumarola T, Alonso-Socas Mdel M, Garcia-Bujalance S, Menendez-Arias L: **Mutational patterns and correlated amino acid substitutions in the HIV-1 protease after virological failure to nelfinavir- and lopinavir/ritonavir-based treatments.** *J Med Virol* 2007, **79:**1617-1628.
8. Kovalevsky AY, Chumanevich AA, Liu F, Louis JM, Weber IT: **Caught in the Act: the 1.5 A resolution crystal structures of the HIV-1 protease and the I54V mutant reveal a tetrahedral reaction intermediate.** *Biochemistry* 2007, **46:**14854-14864.
9. Tie Y, Kovalevsky AY, Boross P, Wang YF, Ghosh AK, Tozser J, Harrison RW, Weber IT: **Atomic resolution crystal structures of HIV-1 protease and mutants V82A and I84V with saquinavir.** *Proteins* 2007, **67:**232-242.

10. Wang YF, Tie Y, Boross PI, Tozser J, Ghosh AK, Harrison RW, Weber IT: **Potent new antiviral compound shows similar inhibition and structural interactions with drug resistant mutants and wild type HIV-1 protease.** *J Med Chem* 2007, **50:**4509-4515.
11. Hou T, Zhang W, Wang J, Wang W: **Predicting drug resistance of the HIV-1 protease using molecular interaction energy components.** *Proteins* 2009, **74:**837-846.
12. Kontijevskis A, Prusis P, Petrovska R, Yahorava S, Mutulis F, Mutule I, Komorowski J, Wikberg JE: **A look inside HIV resistance through retroviral protease interaction maps.** *PLoS Comput Biol* 2007, **3:**e48.
13. Shuman CF, Markgren PO, Hamalainen M, Danielson UH: **Elucidation of HIV-1 protease resistance by characterization of interaction kinetics between inhibitors and enzyme variants.** *Antiviral Res* 2003, **58:**235-242.
14. Beerenwinkel N, Sing T, Lengauer T, Rahnenfuhrer J, Roomp K, Savenkov I, Fischer R, Hoffmann D, Selbig J, Korn K, et al: **Computational methods for the design of effective therapies against drug resistant HIV strains.** *Bioinformatics* 2005, **21:**3943-3950.
15. Ghosh AK, Chapsal BD, Weber IT, Mitsuya H: **Design of HIV protease inhibitors targeting protein backbone: an effective strategy for combating drug resistance.** *Acc Chem Res* 2008, **41:**78-86.
16. Anderson J, Schiffer C, Lee SK, Swanstrom R: **Viral protease inhibitors.** *Handb Exp Pharmacol* 2009**:**85-110.
17. Hong L, Zhang XC, Hartsuck JA, Tang J: **Crystal structure of an in vivo HIV-1 protease mutant in complex with saquinavir: insights into the mechanisms of drug resistance.** *Protein Sci* 2000, **9:**1898-1904.
18. Lexa KW, Damm KL, Quintero JJ, Gestwicki JE, Carlson HA: **Clarifying allosteric control of flap conformations in the 1TW7 crystal structure of HIV-1 protease.** *Proteins* 2009, **74:**872-880.
19. Perryman AL, Lin JH, Andrew McCammon J: **Optimization and computational evaluation of a series of potential active site inhibitors of the V82F/I84V drug-resistant mutant of HIV-1 protease: an application of the relaxed complex method of structure-based drug design.** *Chem Biol Drug Des* 2006, **67:**336-345.
20. Bierman WF, van Agtmael MA, Nijhuis M, Danner SA, Boucher CA: **HIV monotherapy with ritonavir-boosted protease inhibitors: a systematic review.** *AIDS* 2009, **23:**279-291.
21. Perryman AL, Lin JH, McCammon JA: **HIV-1 protease molecular dynamics of a wild-type and of the V82F/I84V mutant: possible contributions to drug resistance and a potential new target site for drugs.** *Protein Sci* 2004, **13:**1108-1123.
22. Bennett DE, Camacho RJ, Otelea D, Kuritzkes DR, Fleury H, Kiuchi M, Heneine W, Kantor R, Jordan MR, Schapiro JM, et al: **Drug resistance mutations for surveillance of transmitted HIV-1 drug-resistance: 2009 update.** *PLoS One* 2009, **4:**e4724.
23. Johnson VA, Brun-Vezinet F, Clotet B, Gunthard HF, Kuritzkes DR, Pillay D, Schapiro JM, Richman DD: **Update of the Drug Resistance Mutations in HIV-1.** *Top HIV Med* 2008, **16:**138-145.
24. Boden D, Markowitz M: **Resistance to human immunodeficiency virus type 1 protease inhibitors.** *Antimicrob Agents Chemother* 1998, **42:**2775-2783.
25. Kolli M, Stawiski E, Chappey C, Schiffer CA: **Human immunodeficiency virus type 1 protease-correlated cleavage site mutations enhance inhibitor resistance.** *J Virol* 2009, **83:**11027-11042.

26. Parera M, Fernandez G, Clotet B, Martinez MA: **HIV-1 protease catalytic efficiency effects caused by random single amino acid substitutions.** *Mol Biol Evol* 2007, **24:**382-387.
27. Svicher V, Ceccherini-Silberstein F, Erba F, Santoro M, Gori C, Bellocchi MC, Giannella S, Trotta MP, Monforte A, Antinori A, Perno CF: **Novel human immunodeficiency virus type 1 protease mutations potentially involved in resistance to protease inhibitors.** *Antimicrob Agents Chemother* 2005, **49:**2015-2025.
28. Ishima R, Louis JM: **A diverse view of protein dynamics from NMR studies of HIV-1 protease flaps.** *Proteins* 2008, **70:**1408-1415.
29. Hornak V, Okur A, Rizzo RC, Simmerling C: **HIV-1 protease flaps spontaneously open and reclose in molecular dynamics simulations.** *Proc Natl Acad Sci U S A* 2006, **103:**915-920.
30. Seibold SA, Cukier RI: **A molecular dynamics study comparing a wild-type with a multiple drug resistant HIV protease: differences in flap and aspartate 25 cavity dimensions.** *Proteins* 2007, **69:**551-565.
31. Alcaro S, Artese A, Ceccherini-Silberstein F, Ortuso F, Perno CF, Sing T, Svicher V: **Molecular dynamics and free energy studies on the wild-type and mutated HIV-1 protease complexed with four approved drugs: mechanism of binding and drug resistance.** *J Chem Inf Model* 2009, **49:**1751-1761.
32. Ghosh AK, Leshchenko-Yashchuk S, Anderson DD, Baldridge A, Noetzel M, Miller HB, Tie Y, Wang YF, Koh Y, Weber IT, Mitsuya H: **Design of HIV-1 protease inhibitors with pyrrolidinones and oxazolidinones as novel P1'-ligands to enhance backbone-binding interactions with protease: synthesis, biological evaluation, and protein-ligand X-ray studies.** *J Med Chem* 2009, **52:**3902-3914.
33. Hamacher K: **Relating sequence evolution of HIV1-protease to its underlying molecular mechanics.** *Gene* 2008, **422:**30-36.
34. Hou T, McLaughlin WA, Wang W: **Evaluating the potency of HIV-1 protease drugs to combat resistance.** *Proteins* 2008, **71:**1163-1174.
35. Jayaraman S, Shah K: **Comparative studies on inhibitors of HIV protease: a target for drug design.** *In Silico Biol* 2008, **8:**427-447.
36. Jenwitheesuk E, Samudrala R: **Prediction of HIV-1 protease inhibitor resistance using a protein-inhibitor flexible docking approach.** *Antivir Ther* 2005, **10:**157-166.
37. Lapins M, Eklund M, Spjuth O, Prusis P, Wikberg JE: **Proteochemometric modeling of HIV protease susceptibility.** *BMC Bioinformatics* 2008, **9:**181.
38. Lapins M, Wikberg JE: **Proteochemometric modeling of drug resistance over the mutational space for multiple HIV protease variants and multiple protease inhibitors.** *J Chem Inf Model* 2009, **49:**1202-1210.
39. Liu F, Kovalevsky AY, Tie Y, Ghosh AK, Harrison RW, Weber IT: **Effect of flap mutations on structure of HIV-1 protease and inhibition by saquinavir and darunavir.** *J Mol Biol* 2008, **381:**102-115.
40. Meiselbach H, Horn AH, Harrer T, Sticht H: **Insights into amprenavir resistance in E35D HIV-1 protease mutation from molecular dynamics and binding free-energy calculations.** *J Mol Model* 2007, **13:**297-304.
41. Murphy MD, Marousek GI, Chou S: **HIV protease mutations associated with amprenavir resistance during salvage therapy: importance of I54M.** *J Clin Virol* 2004, **30:**62-67.
42. Paulsen D, Elston R, Snowden W, Tisdale M, Ross L: **Differentiation of genotypic resistance profiles for amprenavir and lopinavir, a valuable aid for choice of therapy in**

protease inhibitor-experienced HIV-1-infected subjects. *J Antimicrob Chemother* 2003, **52:**319-323.
43. Sherman W, Tidor B: **Novel method for probing the specificity binding profile of ligands: applications to HIV protease.** *Chem Biol Drug Des* 2008, **71:**387-407.
44. Turner D, Schapiro JM, Brenner BG, Wainberg MA: **The influence of protease inhibitor resistance profiles on selection of HIV therapy in treatment-naive patients.** *Antivir Ther* 2004, **9:**301-314.
45. Van Marck H, Dierynck I, Kraus G, Hallenberger S, Pattery T, Muyldermans G, Geeraert L, Borozdina L, Bonesteel R, Aston C, et al: **The impact of individual human immunodeficiency virus type 1 protease mutations on drug susceptibility is highly influenced by complex interactions with the background protease sequence.** *J Virol* 2009, **83:**9512-9520.
46. Verkhivker G: **Computational proteomics analysis of binding mechanisms and molecular signatures of the HIV-1 protease drugs.** *Artif Intell Med* 2009, **45:**197-206.
47. Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK: **Relating protein pharmacology by ligand chemistry.** *Nat Biotechnol* 2007, **25:**197-206.
48. Wang D, Larder B, Revell A, Montaner J, Harrigan R, De Wolf F, Lange J, Wegner S, Ruiz L, Perez-Elias MJ, et al: **A comparison of three computational modelling methods for the prediction of virological response to combination HIV therapy.** *Artif Intell Med* 2009, **47:**63-74.
49. Ceccherini-Silberstein F, Malet I, D'Arrigo R, Antinori A, Marcelin AG, Perno CF: **Characterization and structural analysis of HIV-1 integrase conservation.** *AIDS Rev* 2009, **11:**17-29.
50. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank.** *Nucleic Acids Res* 2000, **28:**235-242.
51. **http://www.expasy.ch/sprot/.**
52. **http://www.ncbi.nlm.nih.gov/Genbank/.**
53. **http://www.ebi.ac.uk/embl/.**
54. Bajorath J: **Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening.** *J Chem Inf Comput Sci* 2001, **41:**233-245.
55. Bajorath J: **Chemoinformatics methods for systematic comparison of molecules from natural and synthetic sources and design of hybrid libraries.** *Mol Divers* 2002, **5:**305-313.
56. Godden JW, Bajorath J: **Chemical descriptors with distinct levels of information content and varying sensitivity to differences between selected compound databases identified by SE-DSE analysis.** *J Chem Inf Comput Sci* 2002, **42:**87-93.
57. Oprea TI, Matter H: **Integrating virtual screening in lead discovery.** *Curr Opin Chem Biol* 2004, **8:**349-358.
58. de Julian-Ortiz JV: **Virtual darwinian drug design: QSAR inverse problem, virtual combinatorial chemistry, and computational screening.** *Comb Chem High Throughput Screen* 2001, **4:**295-310.
59. Tropsha A, Golbraikh A: **Predictive QSAR modeling workflow, model applicability domains, and virtual screening.** *Curr Pharm Des* 2007, **13:**3494-3504.
60. Kirchmair J, Distinto S, Schuster D, Spitzer G, Langer T, Wolber G: **Enhancing drug discovery through in silico screening: strategies to increase true positives retrieval rates.** *Curr Med Chem* 2008, **15:**2040-2053.

61. Cao DS, Liang YZ, Xu QS, Li HD, Chen X: **A new strategy of outlier detection for QSAR/QSPR.** *J Comput Chem* 2009.
62. Winkler DA, Burden FR: **Application of neural networks to large dataset QSAR, virtual screening, and library design.** *Methods Mol Biol* 2002, **201:**325-367.
63. Bajorath J: **Integration of virtual and high-throughput screening.** *Nat Rev Drug Discov* 2002, **1:**882-894.
64. Good AC, Krystek SR, Mason JS: **High-throughput and virtual screening: core lead discovery technologies move towards integration.** *Drug Discov Today* 2000, **5:**61-69.
65. Kabachinski J: **Databases, Tuples, and SQL.** *Biomed Instrum Technol* 2008, **42:**385-387.
66. Teorey TJ: **Database Modeling And Design.** *Morgan Kaufmann* 2006, **4:**275.
67. Thompson CB, Sward K: **Modeling and teaching techniques for conceptual and logical relational database design.** *J Med Syst* 2005, **29:**513-525.
68. **http://fabforce.net/dbdesigner4/.**
69. Coles SJ, Day NE, Murray-Rust P, Rzepa HS, Zhang Y: **Enhancement of the chemical semantic web through the use of InChI identifiers.** *Org Biomol Chem* 2005, **3:**1832-1834.
70. 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, **20:**313-328.
71. Mavridis L, Hudson BD, Ritchie DW: **Toward high throughput 3D virtual screening using spherical harmonic surface representations.** *J Chem Inf Model* 2007, **47:**1787-1796.
72. Yamagishi ME, Martins NF, 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.
73. **http://www.lifechemicals.com/.**
74. **http://www.eyesopen.com/.**
75. **http://www.chemaxon.com/.**
76. Eswar N, Eramian D, Webb B, Shen MY, Sali A: **Protein structure modeling with MODELLER.** *Methods Mol Biol* 2008, **426:**145-159.
77. Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, Shen MY, Pieper U, Sali A: **Comparative protein structure modeling using MODELLER.** *Curr Protoc Protein Sci* 2007, **Chapter 2:**Unit 2 9.
78. Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, Chipot C, Skeel RD, Kale L, Schulten K: **Scalable molecular dynamics with NAMD.** *J Comput Chem* 2005, **26:**1781-1802.
79. Cai W, Xu J, Shao X, Leroux V, Beautrait A, Maigret B: **SHEF: a vHTS geometrical filter using coefficients of spherical harmonic molecular surfaces.** *J Mol Model* 2008, **14:**393-401.
80. Miteva MA, Lee WH, Montes MO, Villoutreix BO: **Fast structure-based virtual ligand screening combining FRED, DOCK, and Surflex.** *J Med Chem* 2005, **48:**6012-6022.
81. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knoll EH, Shelley M, Perry JK, et al: **Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy.** *J Med Chem* 2004, **47:**1739-1749.

82. Sobolev V, Sorokine A, Prilusky J, Abola EE, Edelman M: **Automated analysis of interatomic contacts in proteins.** *Bioinformatics* 1999, **15:**327-332.
83. **http://www.iasusa.org/.**
84. **http://www.postgresql.org/.**
85. Jamison DC: **Structured Query Language (SQL) fundamentals.** *Curr Protoc Bioinformatics* 2003, **Chapter 9:**Unit9 2.
86. **http://www.qtsoftware.com/.**

Additional files

- **Additional file 1**
  Interaction list of WT, mutate and patient protease complexes with Amprenavir.

- **Additional file 2**
  All HIV-PDI graphical interfaces

- **Additional file 3**
  HIV Database List

- **Additional file 4**
  Sample selection from the Ligand table

- **Additional file 5**
  Sample selection from the Protein table

- **Additional file 6**
  Sample selection from the Patient table

- **Additional file 7**
  Sample selection from the Mutation table

- **Additional file 8**
  Sample selection from the Treatment table

III. Commentaires

Dans ce chapitre, j'ai décrit la conception d'une plateforme bioinformatique autour de la base de données HIV-PDI. Cette plateforme permet de mettre en relation les données de patients infectés par le VIH présentant des résistances aux inhibiteurs de protéases avec les modifications que cela induit sur la structure de la protéase et les implications sur la thérapie. Elle sert aussi à l'analyse de l'effet des modifications de la structure tridimensionnelle des protéines mutées sur l'interaction avec le ligand. Ceci peut nous permettre de proposer des molécules à fort potentiel thérapeutique pour les mutations observées. Le travail décrit dans ce chapitre a été une expérience enrichissante, car il nous a permis de mettre sur pied une plateforme bioinformatique facilement utilisable, qui permette d'analyser et de mieux comprendre l'apparition des résistances aux ARV par le virus du HIV, afin de pouvoir les surmonter.

La base de données HIV-PDI est ainsi utile en tant que base de données intégrée permettant une meilleure interprétation fondée sur toutes les informations disponibles des résistances aux ARV. Elle aide ainsi à la prise de décisions sur les nouveaux traitements. Un aspect important de notre approche est que le modèle de données conceptuel a été employé pour communiquer avec des experts du domaine pour valider le modèle de la base de données. Le retour d'information des experts du domaine est en effet un élément essentiel lors du développement de la base de données dans des domaines comme la biologie où les informations sont très hétérogènes et complexes. Le second aspect important est la prise en compte des informations sur les modifications structurales des enzymes du VIH et les interactions 3D entre des enzymes du VIH (cible des médicaments) et les ARV avec des références sur la résistance ou la sensibilité aux ARV. L'intégration de telles données fournit de nouveaux aperçus dans l'interprétation et l'analyse des résistances du VIH aux ARV. Par conséquent, une telle base de données ouvre de nouvelles voies pour l'évaluation des nouveaux médicaments qui peuvent ciblé des protéines de VIH ayant subi une ou plusieurs mutations avec une forte affinité, fournissant ainsi des traitements nouveaux et plus adaptés à un cas donné de résistance du VIH aux ARV.

L'utilisation de cette plateforme bioinformatique a permis une meilleure analyse et, dans certains cas, une meilleure compréhension de l'apparition des mutations et des résistances dans le cas de la protéase HIV et de ses antirétroviraux. Elle permet aussi, par l'analyse des

données d'interactions 3D, de l'évolution des mutations et des résistances au cours du temps, de proposer des stratégies afin de pouvoir surmonter les résistances à un traitement.

IV. Annexes

Annexes II: Additional file 1 (Interaction list of WT, mutate and patient protease complexes with Amprenavir)

Annexes III: Additional file 2 (All HIV-PDI graphical interfaces)

Annexes IV: Additional file 3 (HIV Database List)

Annexes V: Additional file 4 (Sample selection from the Ligand table)

Annexes VI: Additional file 5 (Sample selection from the Protein table)

Annexes VII: Additional file 6 (Sample selection from the Patient table)

Annexes VIII: Additional file 7 (Sample selection from the Mutation table)

Annexes IX: Additional file 8 (Sample selection from the Treatment table)

« Si on commence avec des certitudes, on finit avec des doutes. Si on commence avec des doutes, on finit avec des certitudes. »

Francis Bacon

Conclusions et perspectives

Sommaire

I. Conclusions et perspectives

Le travail réalisé au cours de cette thèse nous a donné l'opportunité d'utiliser différentes stratégies applicables au criblage virtuel. Nous en avons testé les performances sur deux projets pharmacologiquement différents: la régulation du taux de cholestérol et le VIH. Dans le chapitre 1 de ce rapport, nous avons décrit l'état de l'art des méthodes utilisables pour analyser l'affinité d'une molécule organique pour une cible de nature protéique. En particulier, nous avons montré l'intérêt de l'amarrage moléculaire pour lequel il est primordial de considérer la flexibilité de la protéine, du ligand et leurs descripteurs moléculaires. Nous avons également introduit la notion de dynamique moléculaire lors de l'amarrage moléculaire afin de considérer l'information de variabilité structurale de la cavité et du ligand. Par ailleurs, nous avons décrit un protocole de CVHD multi étapes (VSM), combinant la rapidité des méthodes basées sur les ligands à la précision des méthodes basées sur la structure de la cible, de façon à réduire le coût en argent et en temps de calculs lors du CVHD de bases de données de ligand de grande taille. Leurs limites d'utilisation ont été démontrées. Le deuxième chapitre a été dédié au développement et à l'utilisation de la plateforme VSM sur la grille de calcul GRID5000 (VSM-G) dans le cadre de l'identification de composés potentiellement actifs sur la cible LXRß, afin de diminuer les coûts (temps et argent) des programmes de criblage virtuel basés sur la structure. Après avoir décrit la plateforme GRID5000 et son utilisation couplée à celle de VSM (VSM-G), nous avons mis en évidence l'enrichissement des résultats obtenus et le gain en temps de calcul réalisé, comparé à l'utilisation d'unité de calcul standard (ordinateur de bureau). La plateforme VSM-G nous permet d'envisager le criblage virtuel sur une chimiothèque de plusieurs millions de composés. Ensuite, le chapitre 3 a porté sur l'utilisation de processus de découverte de connaissances dans les bases de données biologiques, chimiques, et structurales liées aux ligands et à leurs cibles, comme une méthode de CVHD. Nous avons décrit les différentes étapes nécessaires à la mise en œuvre d'un processus de KDD. De plus, nous avons décrit l'utilisation de cette méthodologie avec deux algorithmes de fouille de données (règles d'association et arbres de décision), dans le cadre de l'identification de composés potentiellement actifs sur la cible LXRß. Le KDD a aussi été utilisée dans le chapitre 4 comme une méthode de CVHD comparée à d'autres méthodes telles que VSM-G ou la méthode basée sur les pharmacophores 3D pour cribler un jeu de ligands de la base de données ZINC sur des structures du récepteur de l'hormone LXRß. Cette comparaison a prouvé l'intérêt d'utiliser des filtres fondés sur des unités de connaissance extraites des données dans le processus de CVHD d'une grande base de données

moléculaires. Le cinquième et dernier chapitre de nos travaux à concerné l'application du module de base de données intégrée P3IL pour l'étude de résistance aux ARV du VIH. Dans un premier temps, nous avons mis sur pied une plateforme bioinformatique autour d'une base de données HIV-PDI permettant de mettre en relation les données de patients infectés par le VIH présentant des résistances aux inhibiteurs de protéases avec les modifications que cela induit sur la structure de la protéase et les implications sur la thérapie. Ensuite, dans un second temps, cette plateforme bioinformatique a été utilisée pour l'analyse et le traitement des résistances du HIV aux ARV. Ainsi, elle a permis de mieux comprendre et de pouvoir analyser l'apparition de résistance aux ARV.

Les perspectives intéressantes à tous ces travaux seraient d'associer l'utilisation des propriétés physico-chimiques à la comparaison des formes géométriques dans VSM-G, et de tester différents algorithmes de fouille de données afin de déterminer les plus adéquats et le cadre de leurs utilisations. Enfin, la mise en place d'empreintes de ligands et de cibles serait un moyen efficace permettant d'optimiser la recherche de nouveaux médicaments en ne centrant celle-ci que sur les propriétés les plus importantes pour le mécanisme d'action du complexe protéine-ligand.

En conclusion, même si aujourd'hui il existe une vaste panoplie de méthodes *in silico* rapides et efficaces pour la découverte de nouvelles molécules thérapeutiques, il est très important, dans toute stratégie de CVHD, de considérer dans son ensemble la totalité des données disponibles sur la problématique à traiter, de savoir intégrer ces données, et d'apprendre à extraire les connaissances qui y sont cachées. De façon complémentaire aux approches informatisées, il faut aussi souligner l'importance de l'expertise du chimiste en ce qui concerne la définition et la sélection des descripteurs moléculaires et de celle du biologiste quand il s'agit de qualifier l'activité ou l'inactivité des composés. Pour relever les défis du CVHD il convient donc de constituer ou de renforcer des équipes interdisciplinaires à l'intersection de la biologie, de la chimie et de l'informatique.

Ces travaux ainsi que les projets connexes auxquels j'ai participé ont donné lieu à différentes communications et 6 publications dont 1 en cours de révision et 5 publiées. La liste des travaux est indiquée ci-dessous:

II. Communications (présentation orale et poster)

Présentation orale : Ghemtio L, Maigret B, Smaïl-Tabbone M, Souchet M, Devignes MD. (2009). *A KDD Approach for designing filtering strategies to improve Virtual Screening* ***International Conference on Knowledge Discovery and Information Retrieval, Madeira (Portugal)***.

Poster: Ghemtio L, Petronin F, Maigret B, Smaïl-Tabbone M, Devignes MD. (2009). Model-driven Data Integration for predict HIV protease ligand activity in a Drug Design Context ***HIV Inhibitors Drug Discovery Chemistry Conference, San Diego (USA)***.

Poster: Ghemtio L, Bresso E, Maigret L, Smaïl-Tabbone M, Devignes MD, Souchet M. (2009). Model-driven Data Integration for Mining Protein-Ligand and Protein-Protein Interactions in a Drug Design Context ***MEDCHEM EUROPE, Berlin (Germany)***.

Poster: Ghemtio L, Maigret B, Smaïl-Tabbone M, Devignes MD, Souchet M, Leroux V. (2008). Improvement of high throughput structure-based virtual screening using filtering strategies. Application to hit discovery with Liver-X receptor ***Computer-Aided Drug Design Symposia, Steamboat Springs (USA)***.

Présentation orale : Ghemtio L, Bresso E, Souchet M, Maigret B, Smaïl-Tabbone M, Devignes MD. (2008). Model-driven data integration for mining protein-ligand and protein-protein interactions in a drug design context ***Journées Ouvertes Biologie Informatique Mathématiques, Lille (France)***.

Présentation orale : Ghemtio L, Maigret B, Beautrait A. (2007). Large-scale distributed in silico drug discovery using VSM-G ***International Conference on the Bioinformatics of African Pathogens and Disease Vectors, Nairobi (Kenya)***.

III. Publications

Ghemtio L, Smaïl-Tabbone M, Djikeng A, Devignes MD, Keminse L, Birama N, Petronin F, Fokam J, Maigret B, Ouwe-Missi-Oukem-Boyer O. (2010). HIV-PDI (Protein Drug Interactions) database: An integrated resource for studying HIV drug resistance based on 3D protein-drug interactions ***BMC Medical Genomics*** (In review).

Ghemtio L, Maigret B, Emmanuel J. (2010). Efficiency of a hierarchical protocol for high throughput structure-based virtual screening on GRID5000 cluster Grid ***Open Access Bioinformatics*** (Accepted and on line ASAP).

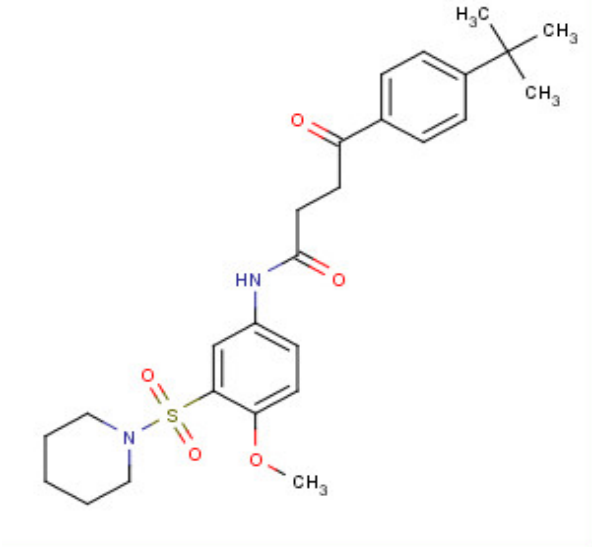
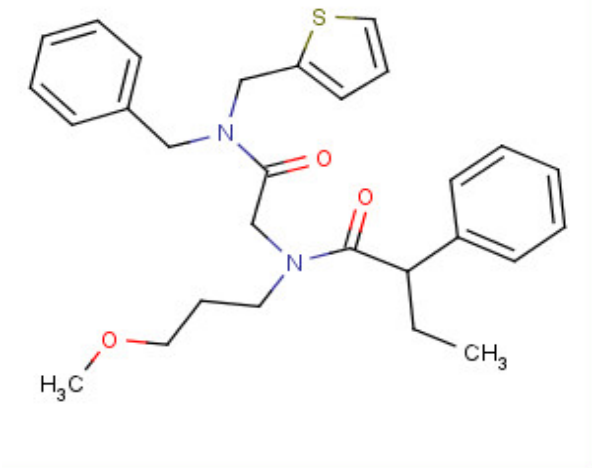
Ghemtio L, Devignes MD, Smaïl-Tabbone M, Souchet M, Leroux V, Maigret B. (2010). Comparison of three pre-processing filters efficiency in virtual screening: Identification of new putative LXRß regulators as a test case ***Journal of Chemical Information and Modeling*** (Accepted and on line ASAP).

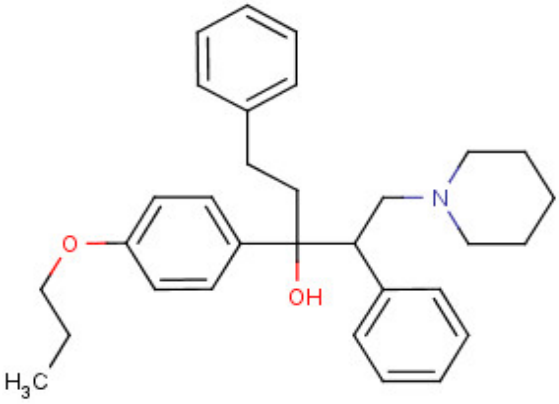
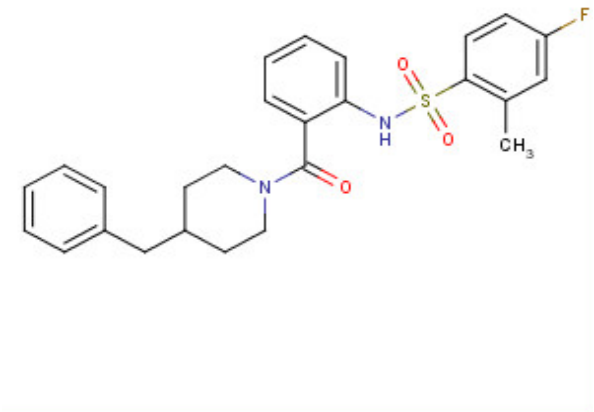
Ghemtio L, Smaïl-Tabbone M, Devignes MD, Souchet M, Maigret B. (2009). A KDD Approach for designing filtering strategies to improve Virtual Screening ***International Conference on Knowledge Discovery and Information Retrieval, Madeira (Portugal).***

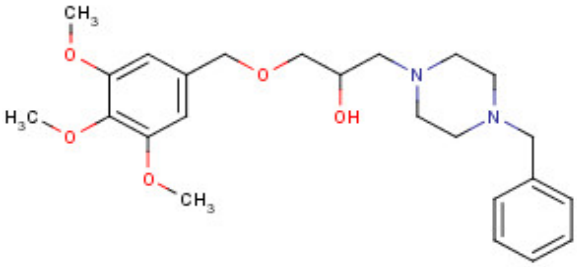
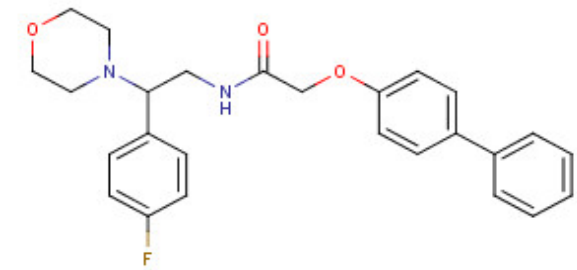
Ghemtio L, Bresso E, Souchet M, Maigret B, Smaïl-Tabbone M, Devignes MD. (2008) Model-driven data integration for mining protein-ligand and protein-protein interactions in a drug design context ***Journées Ouvertes Biologie Informatique Mathématiques – JOBIM, Lille (France)***.

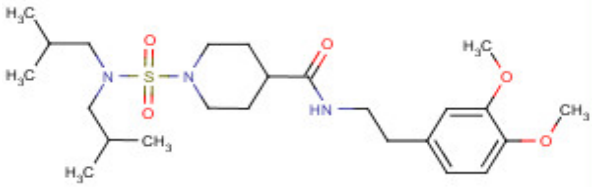
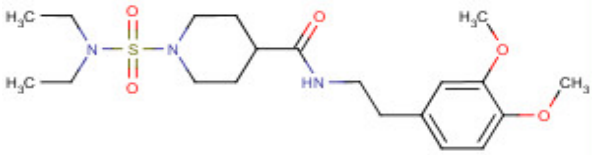
Beautrait A, Leroux V, Chavent M, Ghemtio L, Devignes MD, Smail-Tabbone M, Cai W, Shao X, Moreau G, Bladon P. et al. (2008). Multiple-step virtual screening using VSM-G: overview and validation of fast geometrical matching enrichment ***Journal of Molecular Modeling 14, 2 135-148***.

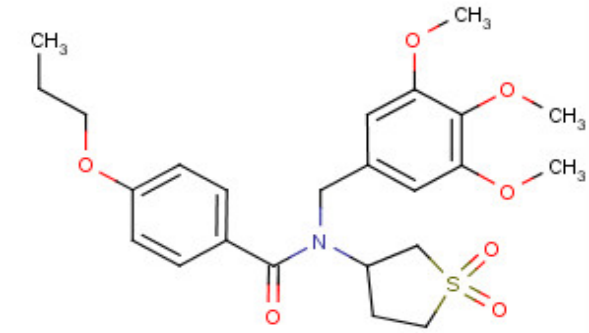
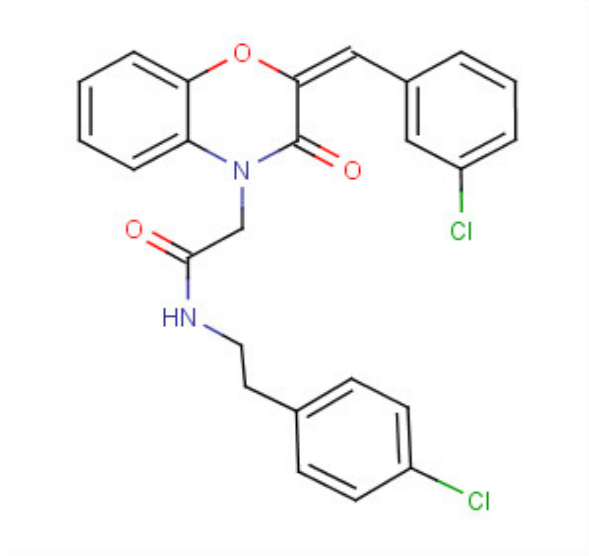
Annexes I: 46 Consensus compound

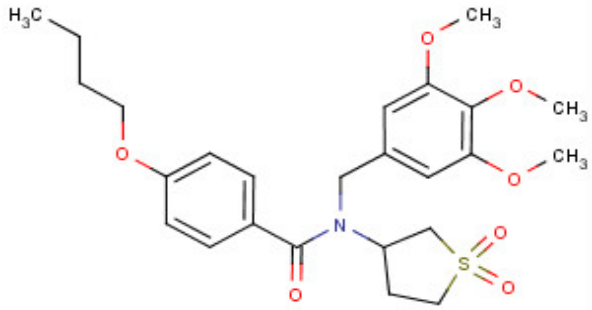
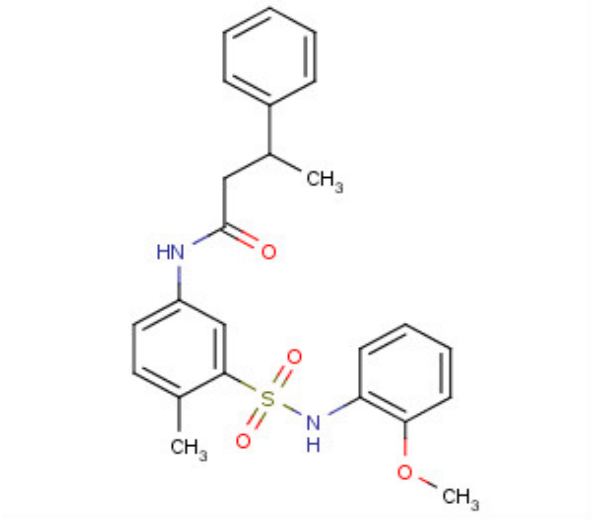
CdId	Structure	Name	Vendor
1		T5480651	enamine
2		CGX-00826190	comgenex

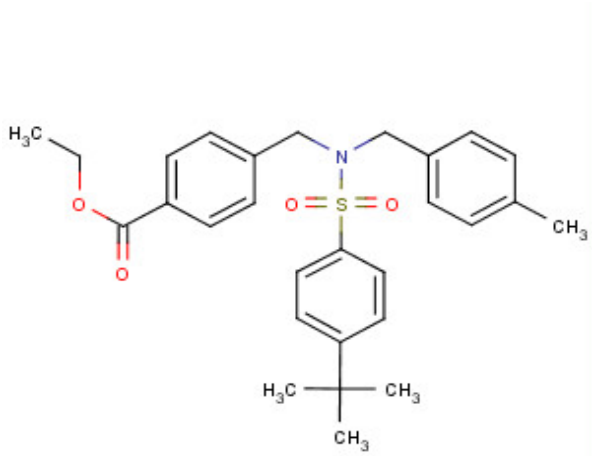
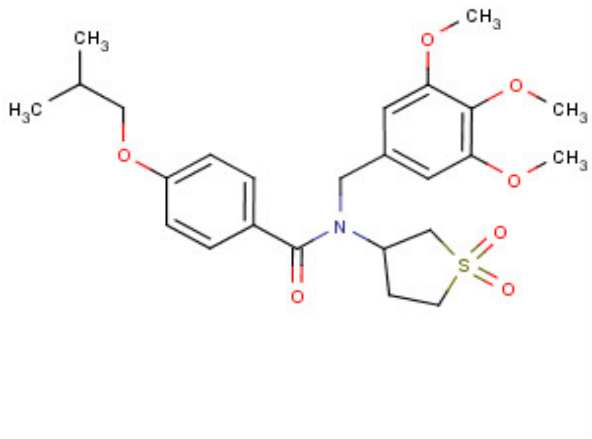
3		D025-0003	chemdiv
4		T0505-6182	enamine

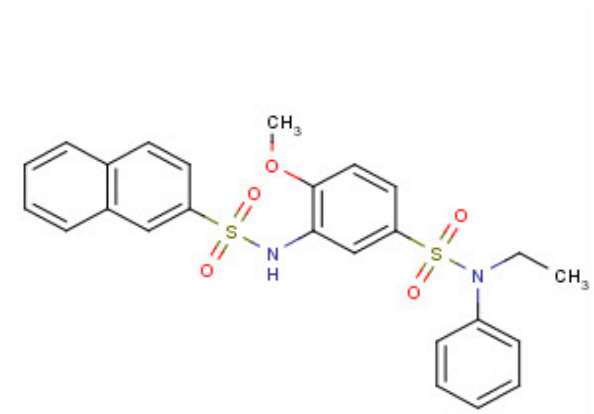
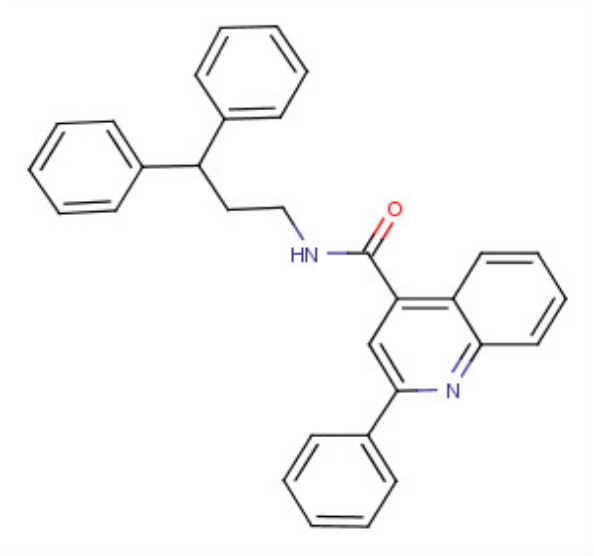
5		T0502-0990	enamine
6		T5540524	enamine

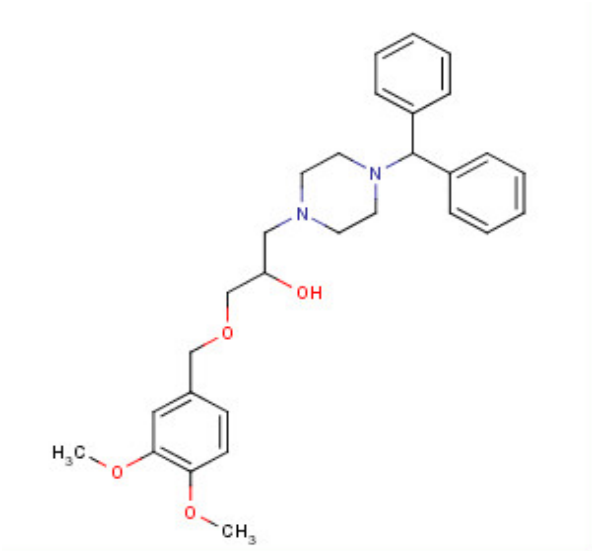
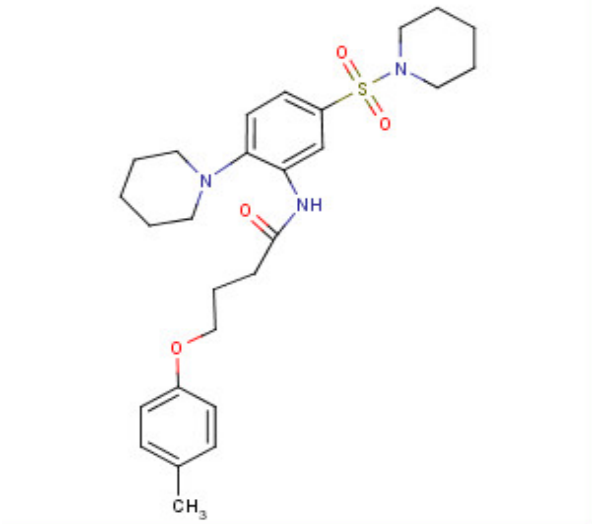
7		K786-6865	chemdiv
8		K786-0206	chemdiv

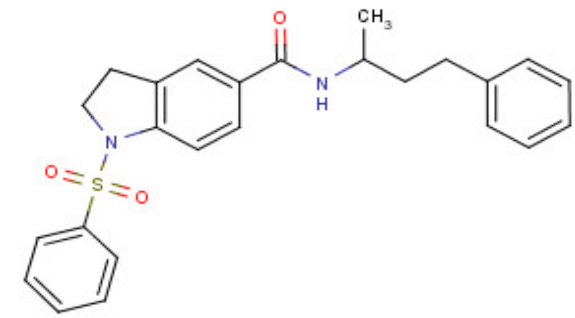
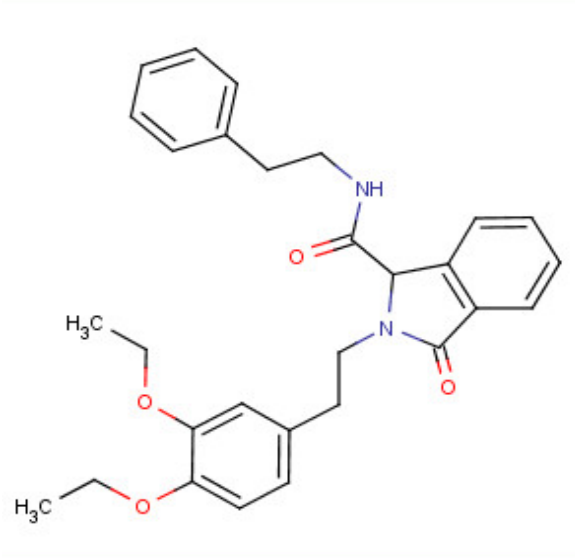
9		D111-0045	chemdiv
10		C311-0192	chemdiv

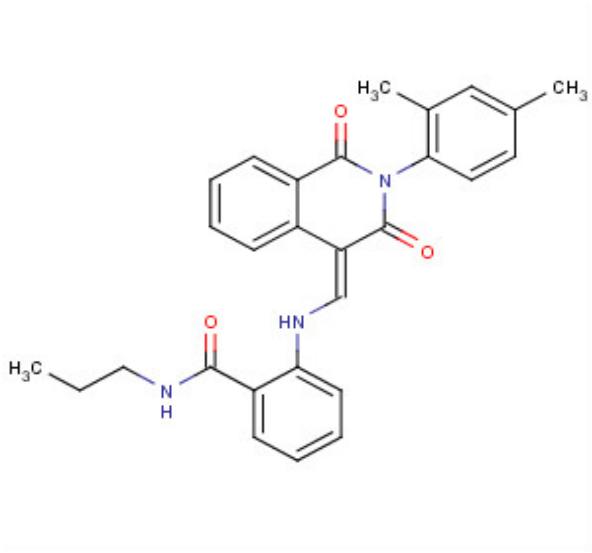
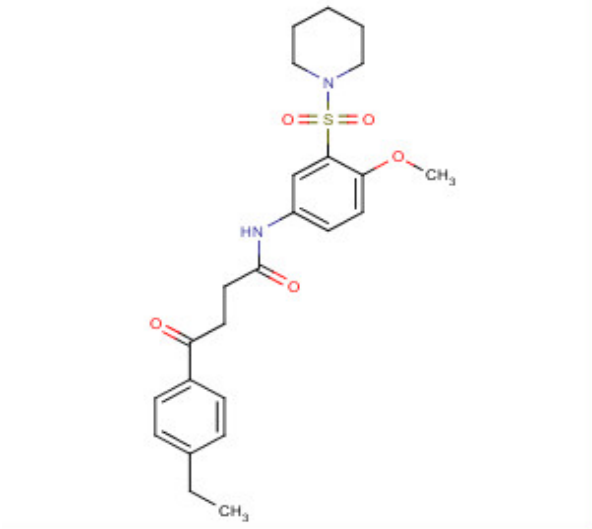
11		D111-0046	chemdiv
12		T5381598	enamine

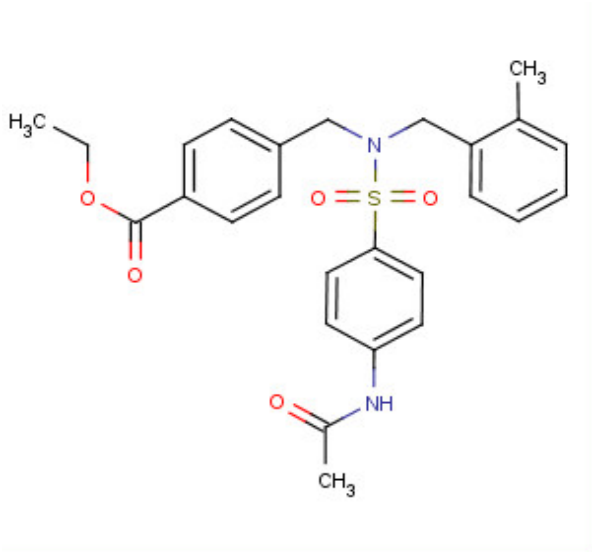
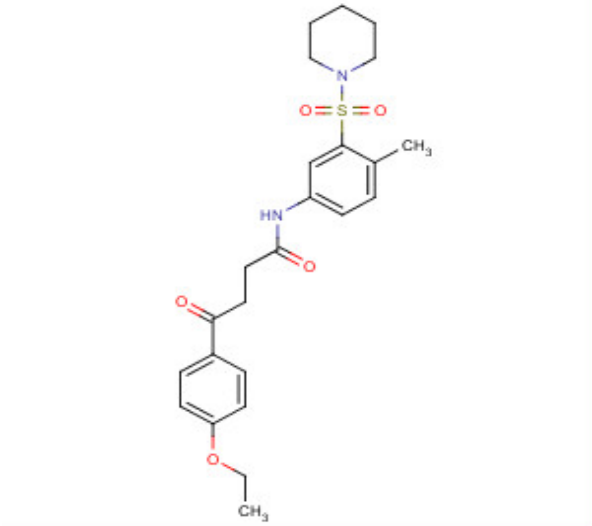
13		K783-3576	chemdiv
14		D111-0055	chemdiv

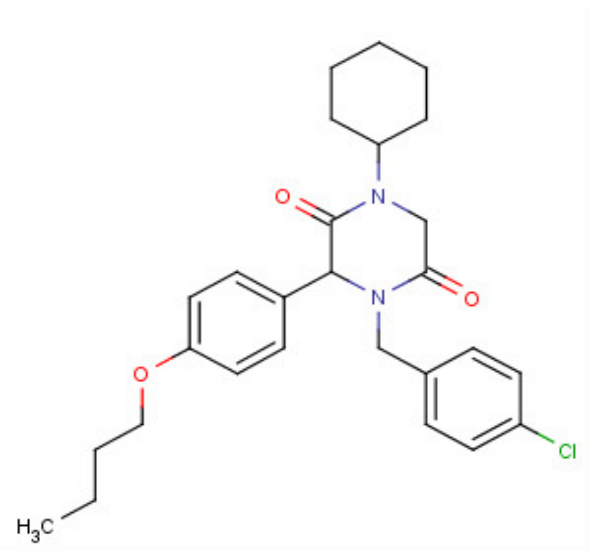
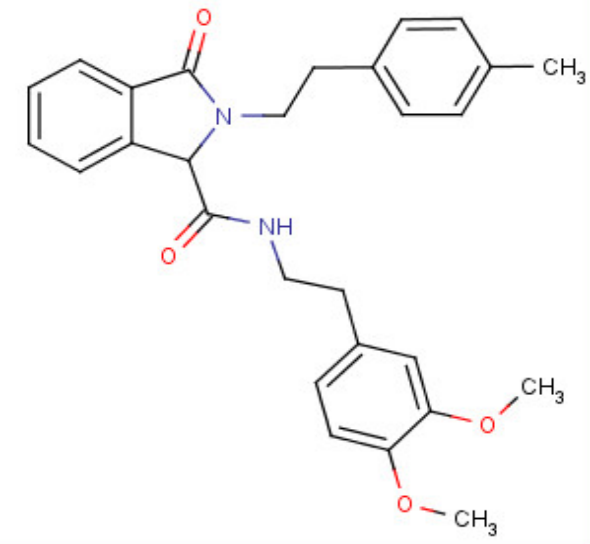
15		T5231816	enamine
16		3366-5414	chemdiv

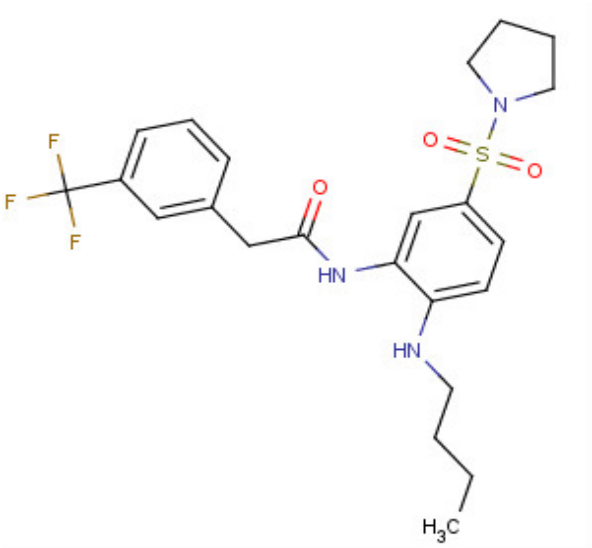
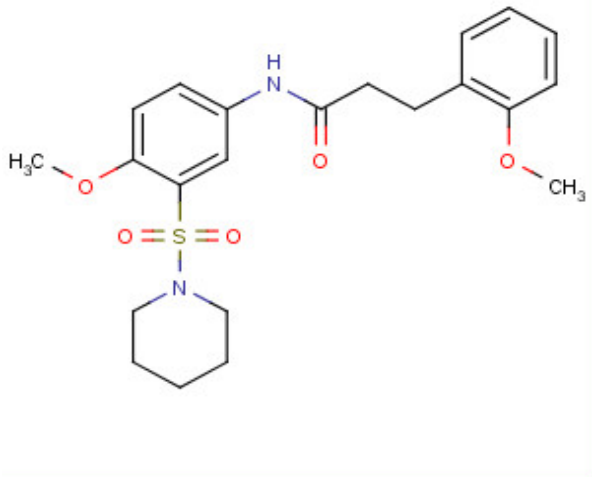
17		T5255750	enamine
18		T5245011	enamine

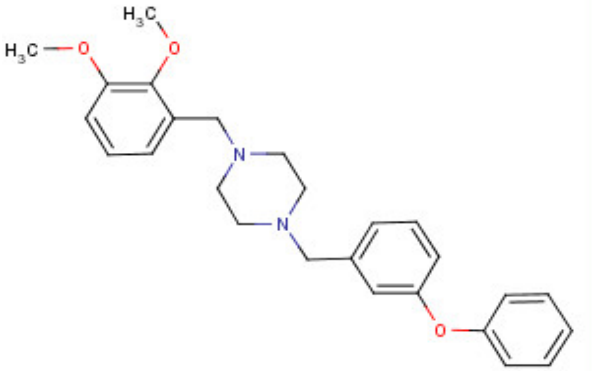
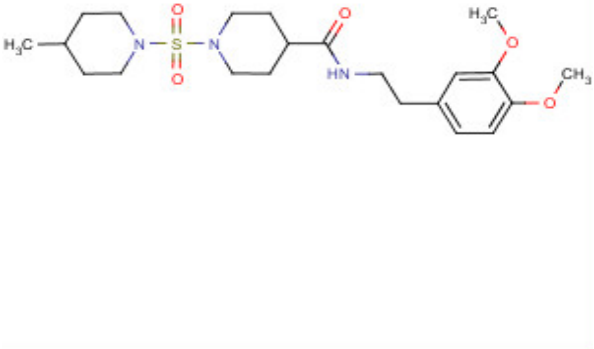
19		D256-0286	chemdiv
20		C257-0240	chemdiv

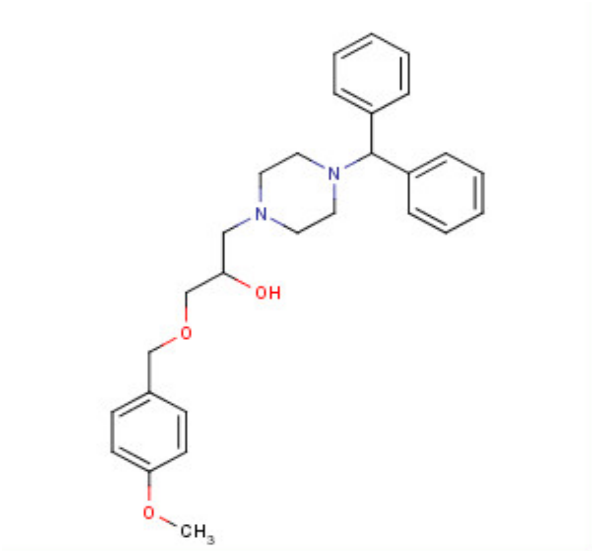
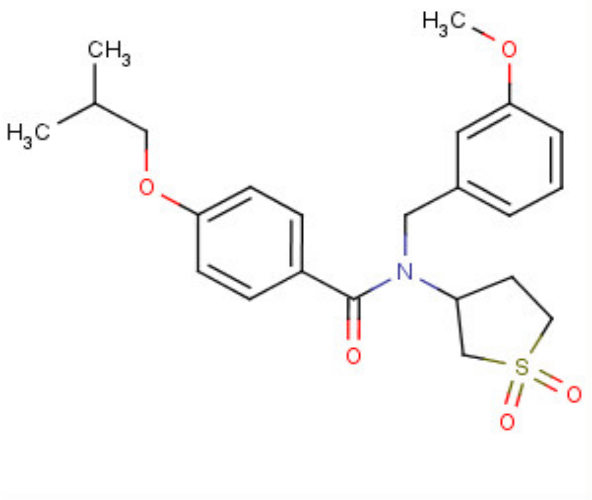
21		T5314194	enamine
22		T5506401	enamine

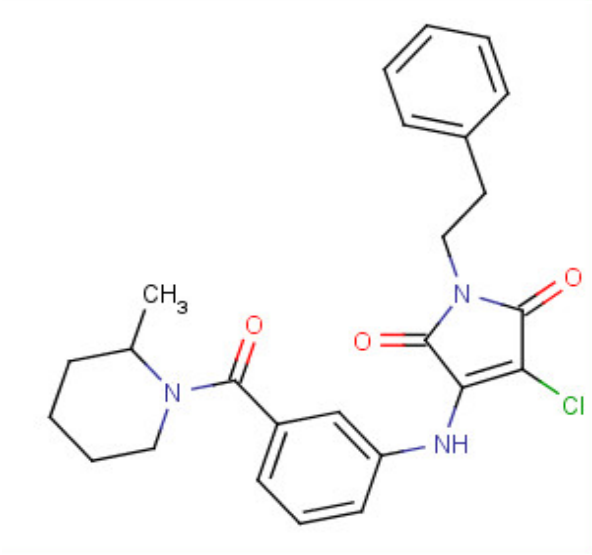
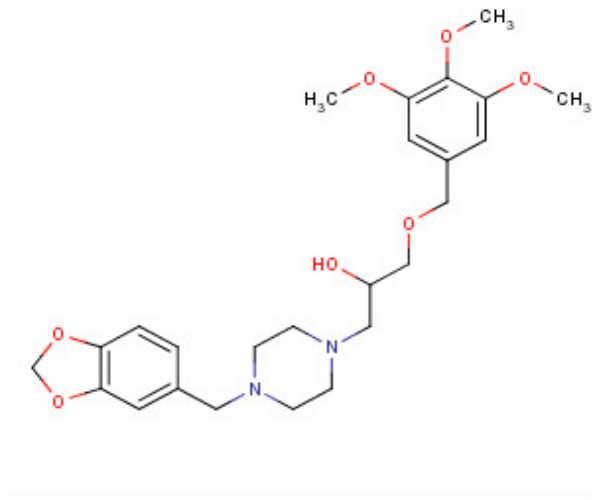
23		K783-3450	chemdiv
24		T5506424	enamine

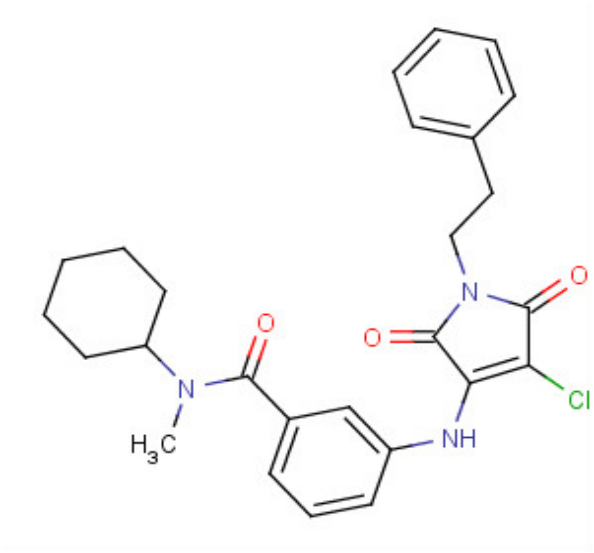
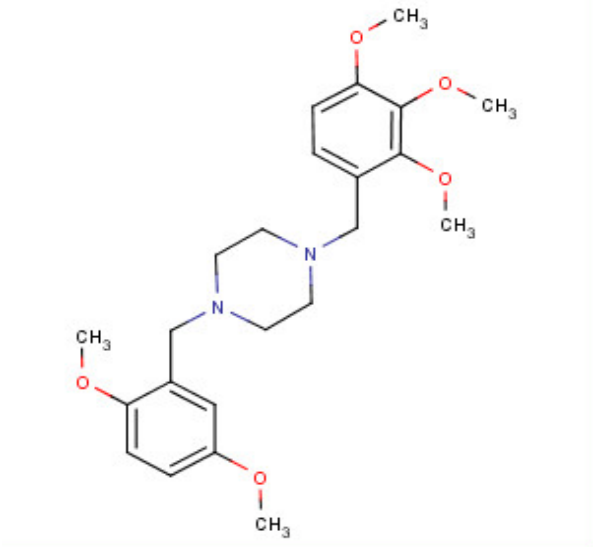
25		K784-7621	chemdiv
26		K786-2427	chemdiv

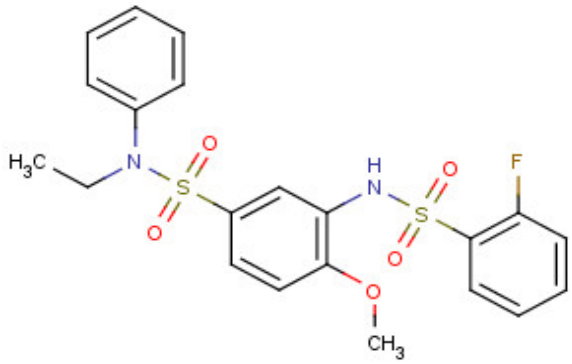
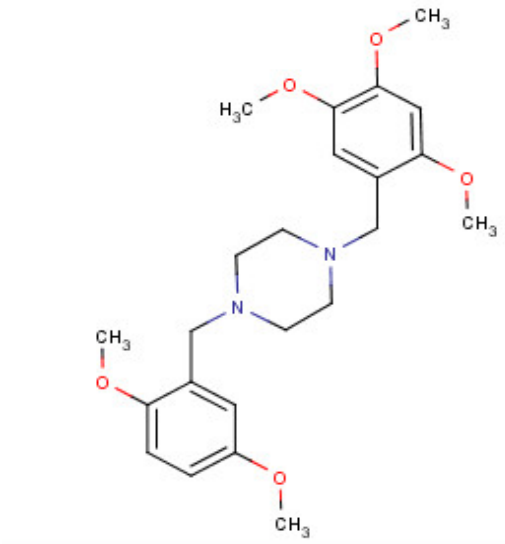
27		T5340844	enamine
28		T5565581	enamine

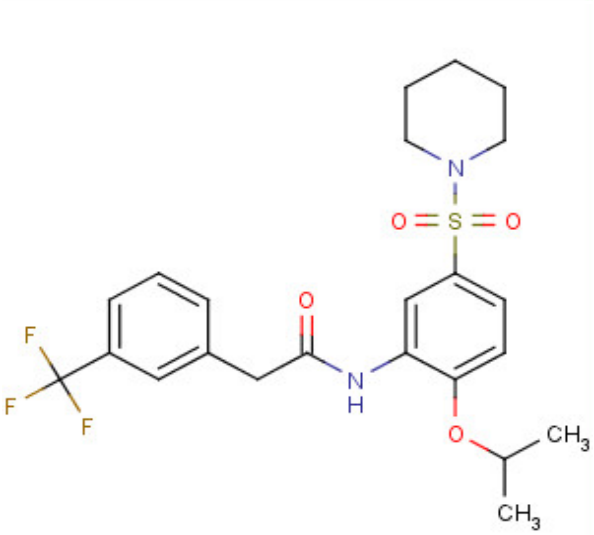
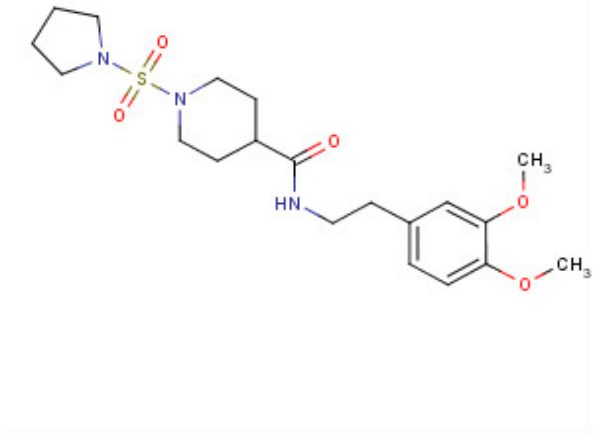
29		4092-0841	chemdiv
30		K786-2558	chemdiv

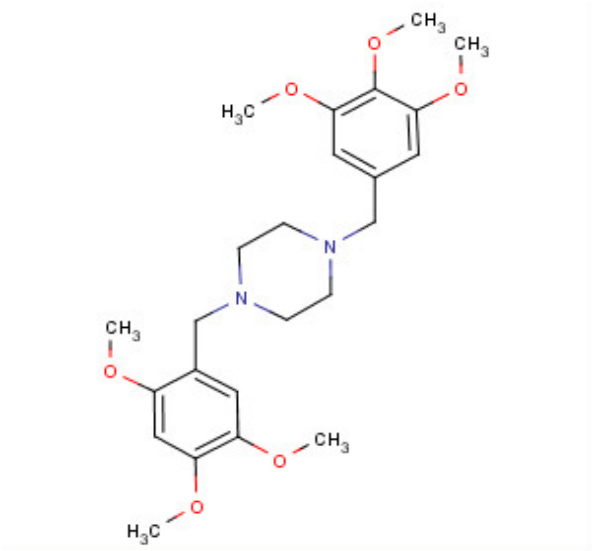
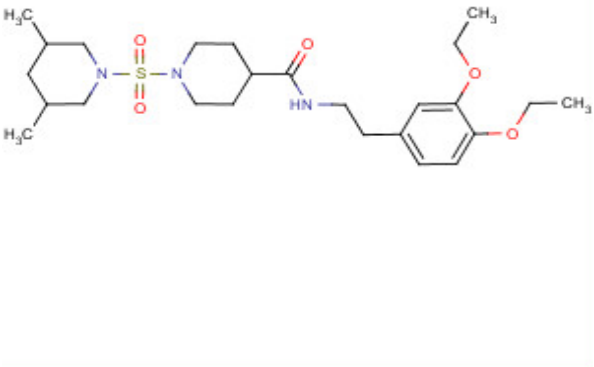
31		T5251163	enamine
32		D076-0055	chemdiv

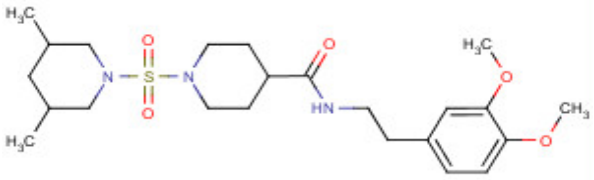
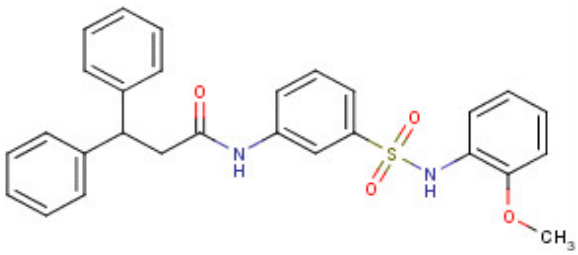
33		T5432559	enamine
34		T0505-1531	enamine

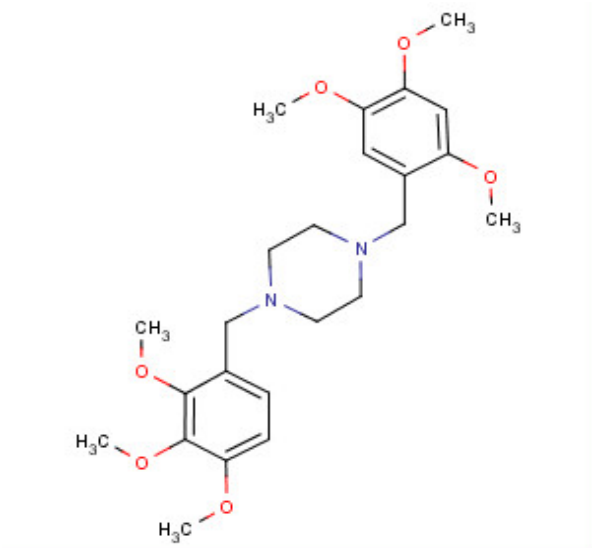
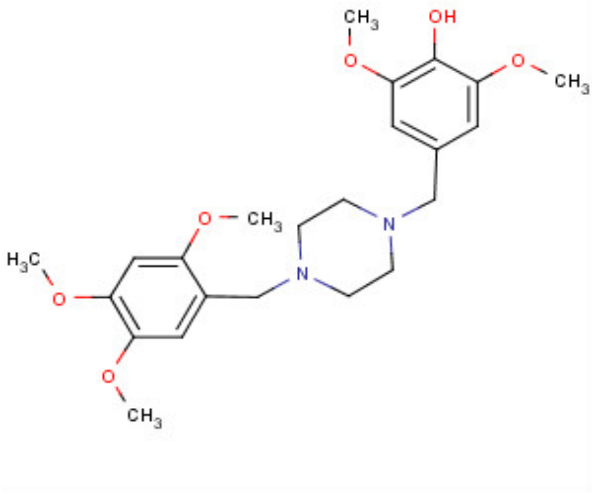
35		T5432613	enamine
36		3381-0518	chemdiv

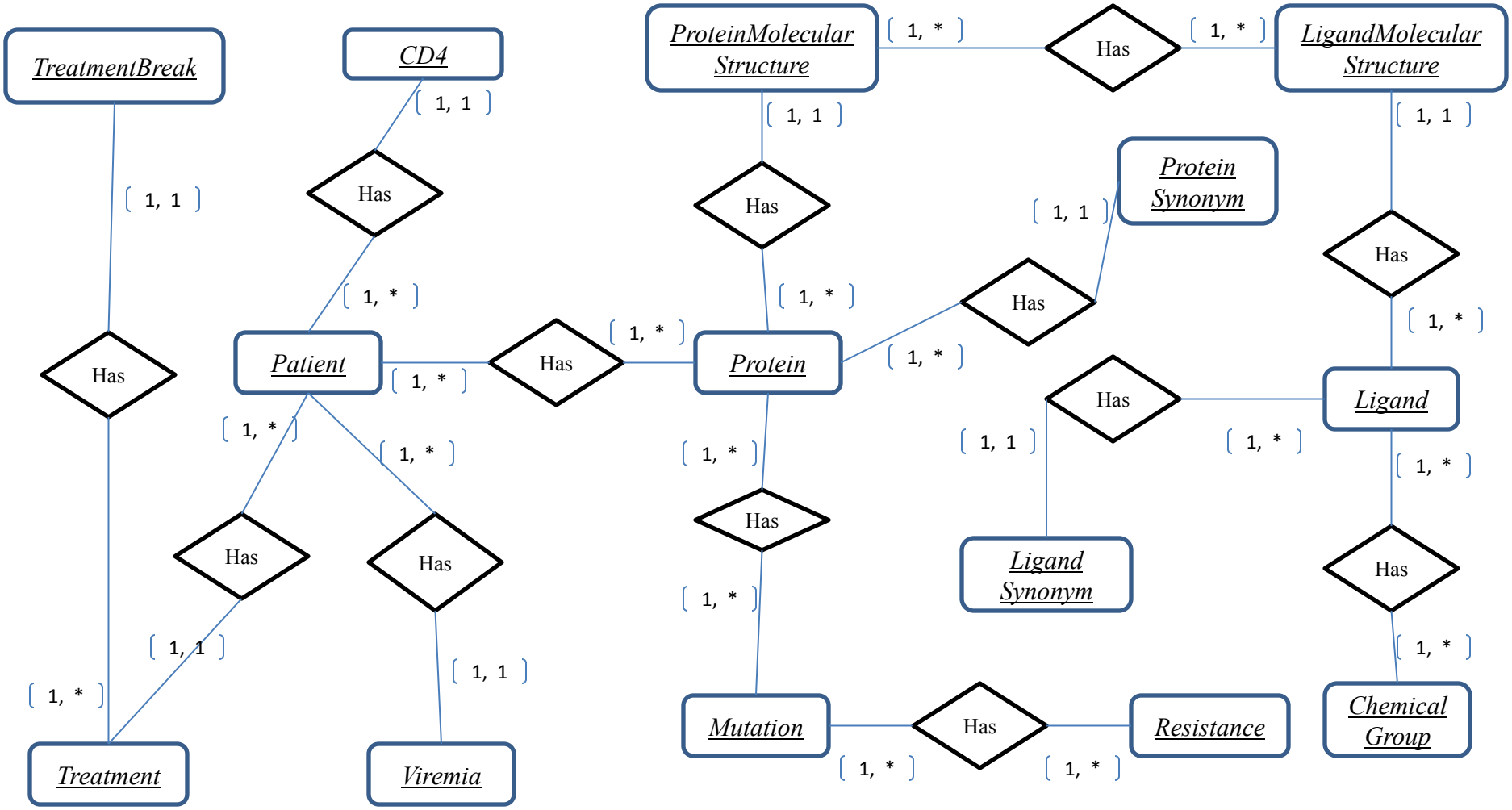
37		T5223398	enamine
38		3381-0712	chemdiv

39		T5401345	enamine
40		K786-5924	chemdiv

41		3702-0742	chemdiv
42		K786-6180	chemdiv

43		K786-6450	chemdiv
44		T0505-9439	enamine

45		3702-0928	chemdiv
46		3701-0475	chemdiv

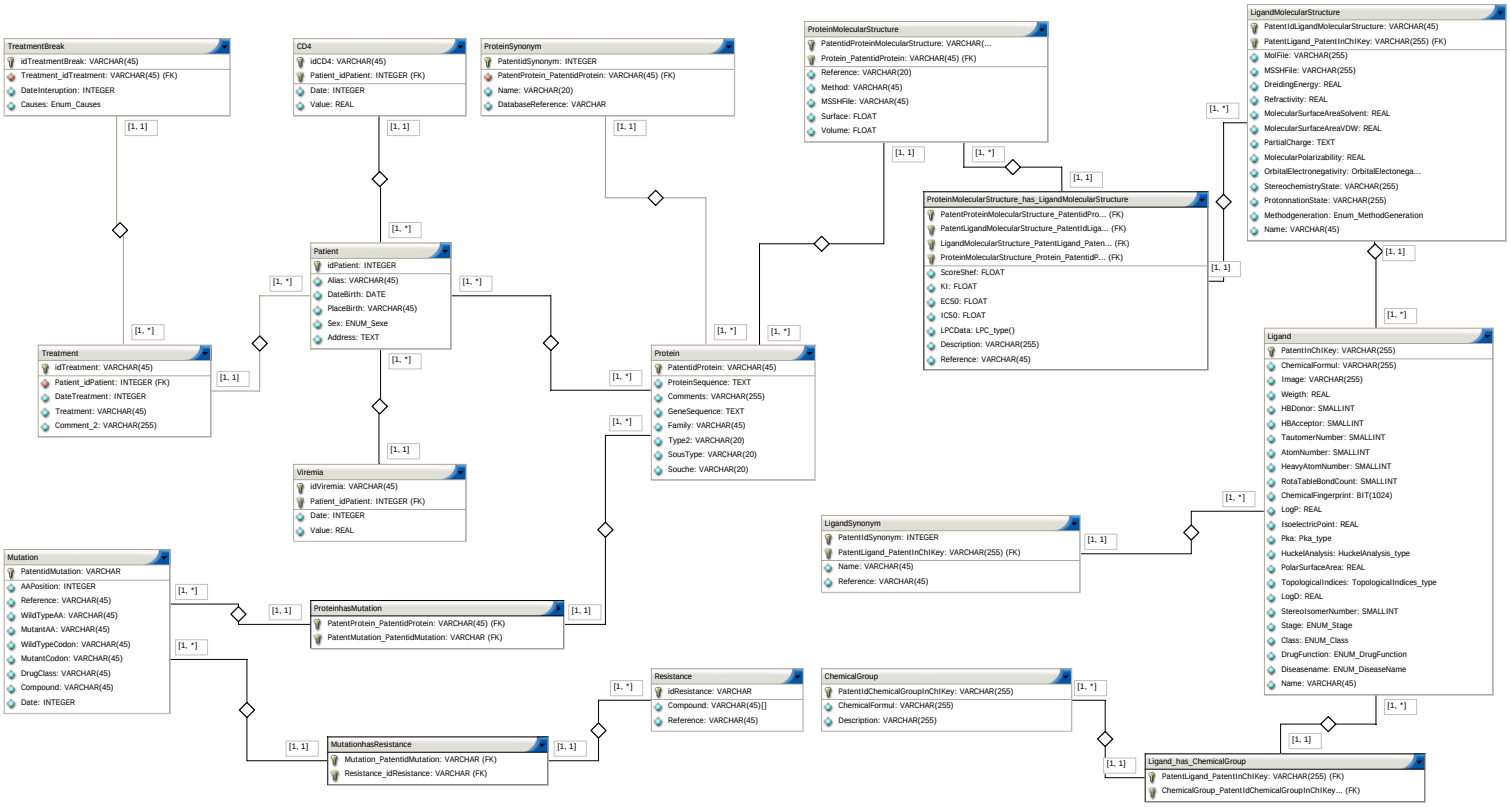


Entity: Represents the objects of concern

Min, Max — **Cardinalities**: Min (minimum) and Max (maximum) describe the minimum and maximum number of entities that may be related to a second entity

Association: Specify the type of the relationship between two or more entities

Links: Line to connect Entity, which demonstrates the relationship between two or more entities


TreatmentBreak
idTreatmentBreak: VARCHAR(45)
Treatment_idTreatment: VARCHAR(45) (FK)
DateInteruption: INTEGER
Causes: Enum_Causes
CD4
idCD4: VARCHAR(45)
Patient_idPatient: INTEGER (FK)
Date: INTEGER
Value: REAL
ProteinSynonym
PatentIdSynonym: INTEGER
PatentProtein_PatentIdProtein: VARCHAR(45) (FK)
Name: VARCHAR(20)
DatabaseReference: VARCHAR
ProteinMolecularStructure
PatentIdProteinMolecularStructure: VARCHAR(...
Protein_PatentIdProtein: VARCHAR(45) (FK)
Reference: VARCHAR(20)
Method: VARCHAR(45)
MSSHFile: VARCHAR(45)
Surface: FLOAT
Volume: FLOAT
LigandMolecularStructure
PatentIdLigandMolecularStructure: VARCHAR(45)
PatentLigand_PatentInChIKey: VARCHAR(255) (FK)
MolFile: VARCHAR(255)
MSSHFile: VARCHAR(255)
DreidingEnergy: REAL
Refractivity: REAL
MolecularSurfaceAreaSolvent: REAL
MolecularSurfaceAreaVDW: REAL
PartialCharge: TEXT
MolecularPolarizability: REAL
OrbitalElectronegativity: OrbitalElectonega...
StereochemistryState: VARCHAR(255)
ProtonnationState: VARCHAR(255)
Methodgeneration: Enum_MethodGeneration
Name: VARCHAR(45)
Patient
idPatient: INTEGER
Alias: VARCHAR(45)
DateBirth: DATE
PlaceBirth: VARCHAR(45)
Sex: ENUM_Sexe
Address: TEXT
ProteinMolecularStructure_has_LigandMolecularStructure
PatentProteinMolecularStructure_PatentIdPro... (FK)
PatentLigandMolecularStructure_PatentIdLiga... (FK)
LigandMolecularStructure_PatentLigand_Paten... (FK)
ProteinMolecularStructure_Protein_PatentIdP... (FK)
ScoreShef: FLOAT
KI: FLOAT
EC50: FLOAT
IC50: FLOAT
LPCData: LPC_type()
Description: VARCHAR(255)
Reference: VARCHAR(45)
Treatment
idTreatment: VARCHAR(45)
Patient_idPatient: INTEGER (FK)
DateTreatment: INTEGER
Treatment: VARCHAR(45)
Comment_2: VARCHAR(255)
Protein
PatentIdProtein: VARCHAR(45)
ProteinSequence: TEXT
Comments: VARCHAR(255)
GeneSequence: TEXT
Family: VARCHAR(45)
Type2: VARCHAR(20)
SousType: VARCHAR(20)
Souche: VARCHAR(20)
Ligand
PatentInChIKey: VARCHAR(255)
ChemicalFormul: VARCHAR(255)
Image: VARCHAR(255)
Weigth: REAL
HBDonor: SMALLINT
HBAcceptor: SMALLINT
TautomerNumber: SMALLINT
AtomNumber: SMALLINT
HeavyAtomNumber: SMALLINT
RotaTableBondCount: SMALLINT
ChemicalFingerprint: BIT(1024)
LogP: REAL
IsoelectricPoint: REAL
Pka: Pka_type
HuckelAnalysis: HuckelAnalysis_type
PolarSurfaceArea: REAL
TopologicalIndices: TopologicalIndices_type
LogD: REAL
StereoIsomerNumber: SMALLINT
Stage: ENUM_Stage
Class: ENUM_Class
DrugFunction: ENUM_DrugFunction
Diseasename: ENUM_DiseaseName
Name: VARCHAR(45)
Viremia
idViremia: VARCHAR(45)
Patient_idPatient: INTEGER (FK)
Date: INTEGER
Value: REAL
LigandSynonym
PatentIdSynonym: INTEGER
PatentLigand_PatentInChIKey: VARCHAR(255) (FK)
Name: VARCHAR(45)
Reference: VARCHAR(45)
Mutation
PatentIdMutation: VARCHAR
AAPosition: INTEGER
Reference: VARCHAR(45)
WildTypeAA: VARCHAR(45)
MutantAA: VARCHAR(45)
WildTypeCodon: VARCHAR(45)
MutantCodon: VARCHAR(45)
DrugClass: VARCHAR(45)
Compound: VARCHAR(45)
Date: INTEGER
ProteinhasMutation
PatentProtein_PatentIdProtein: VARCHAR(45) (FK)
PatentMutation_PatentIdMutation: VARCHAR (FK)
Resistance
idResistance: VARCHAR
Compound: VARCHAR(45)[]
Reference: VARCHAR(45)
ChemicalGroup
PatentIdChemicalGroupInChIKey: VARCHAR(255)
ChemicalFormul: VARCHAR(255)
Description: VARCHAR(255)
MutationhasResistance
Mutation_PatentIdMutation: VARCHAR (FK)
Resistance_idResistance: VARCHAR (FK)
Ligand_has_ChemicalGroup
PatentLigand_PatentInChIKey: VARCHAR(255) (FK)
ChemicalGroup_PatentIdChemicalGroupInChIKey... (FK)

HIV-PDI

HIV-PDI
To Overcome Drug Resistance

HIV-PDI
To Overcome Drug Resistance

Mutation

AA Position :
Reference : All
Wild Type AA :
Mutant AA :
Drug Class : All
Compound : All
Date :

Reset Validate

Treatment

Date :
Treatment :

Reset Validate

Resistance

Mutation ID :
Reference : All
Compound :

Reset Validate

Ligand

1/2 D

InchIkey
All
Name
Chemical Formul
Drug Function
All

3 D

Name
Method Generation
All
Molecular Structure ID

File : Type : All

Reset Validate

Protein

Protein ID : All
Protein Sequence :
Gene Sequence :
Family : All
Type : All
Souche : All
Sous-Type : All

Reset Validate

Complex

Complex ID | Protein ID | Ligand ID
All | All | All

Interactions

- Hydrophilic
- Aromatic
- Acceptor
- Neutral
- Donor
- Neutral-Donor
- Hydrophobic
- Neutral-Acceptor

Reset Validate

Patient

Patient ID: All
Alias :
Sexe : All
Age : 0,0
Address :

Reset Validate

more informations...

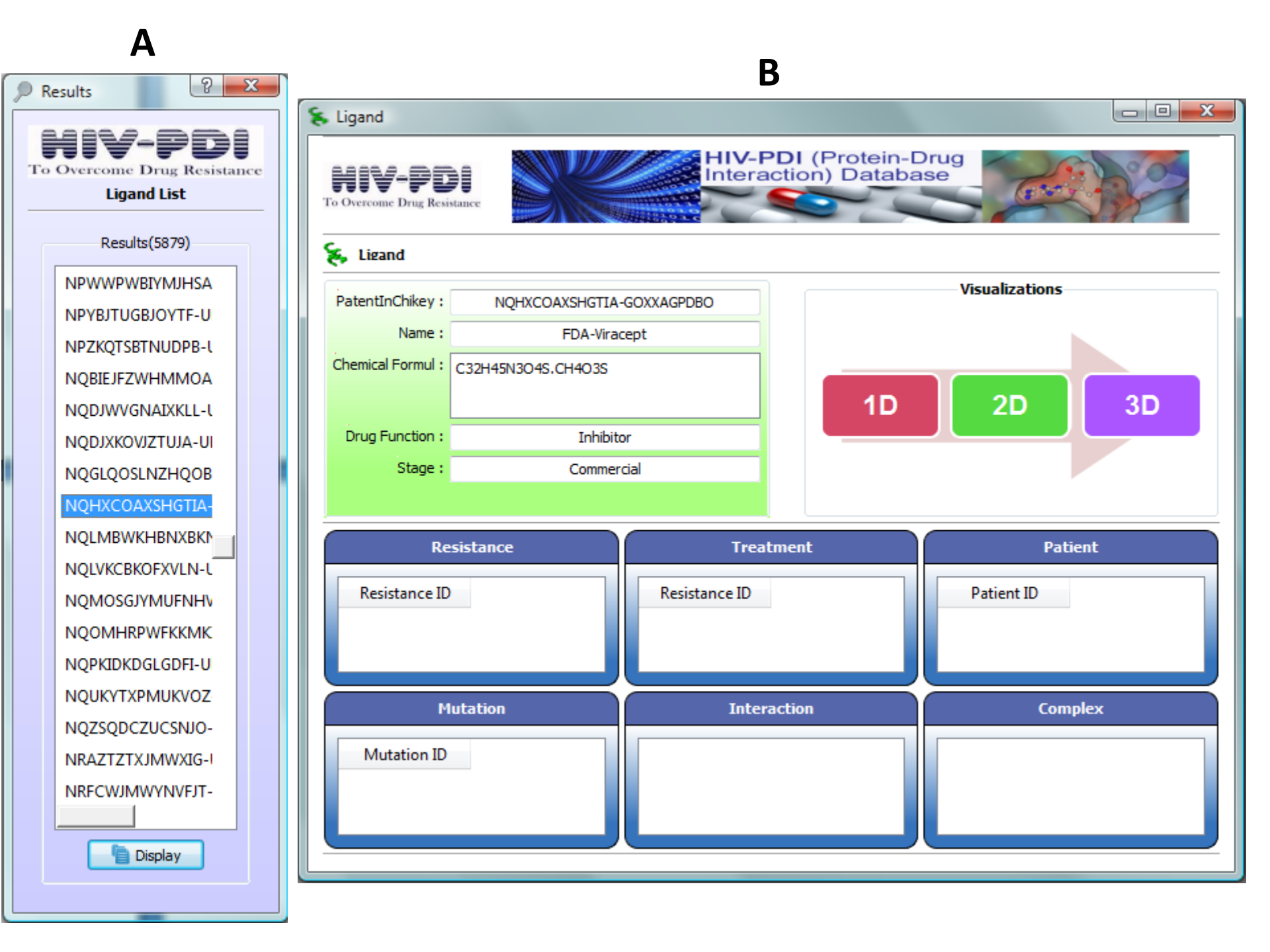
RESET ALL

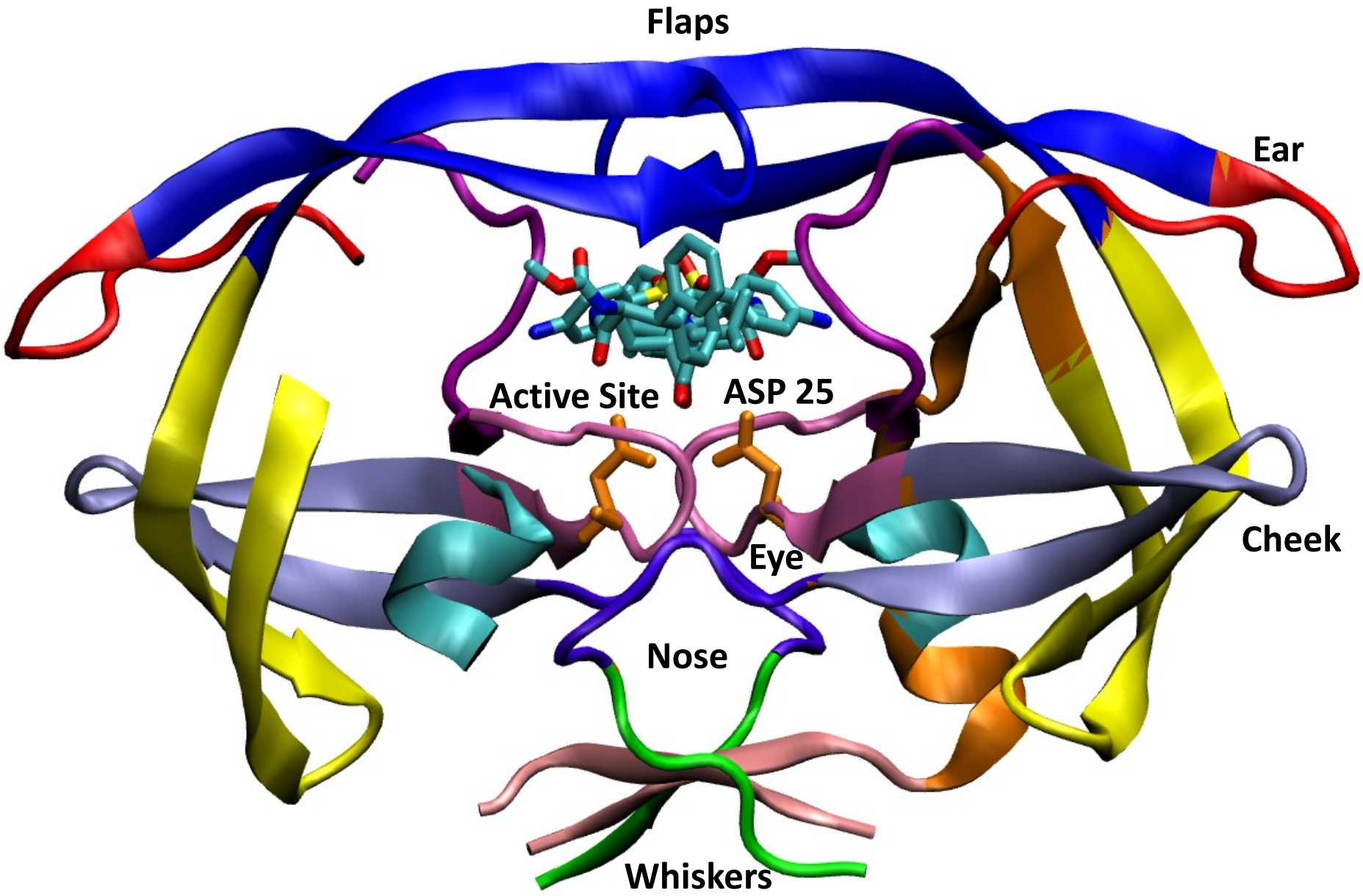
SUBMIT ALL

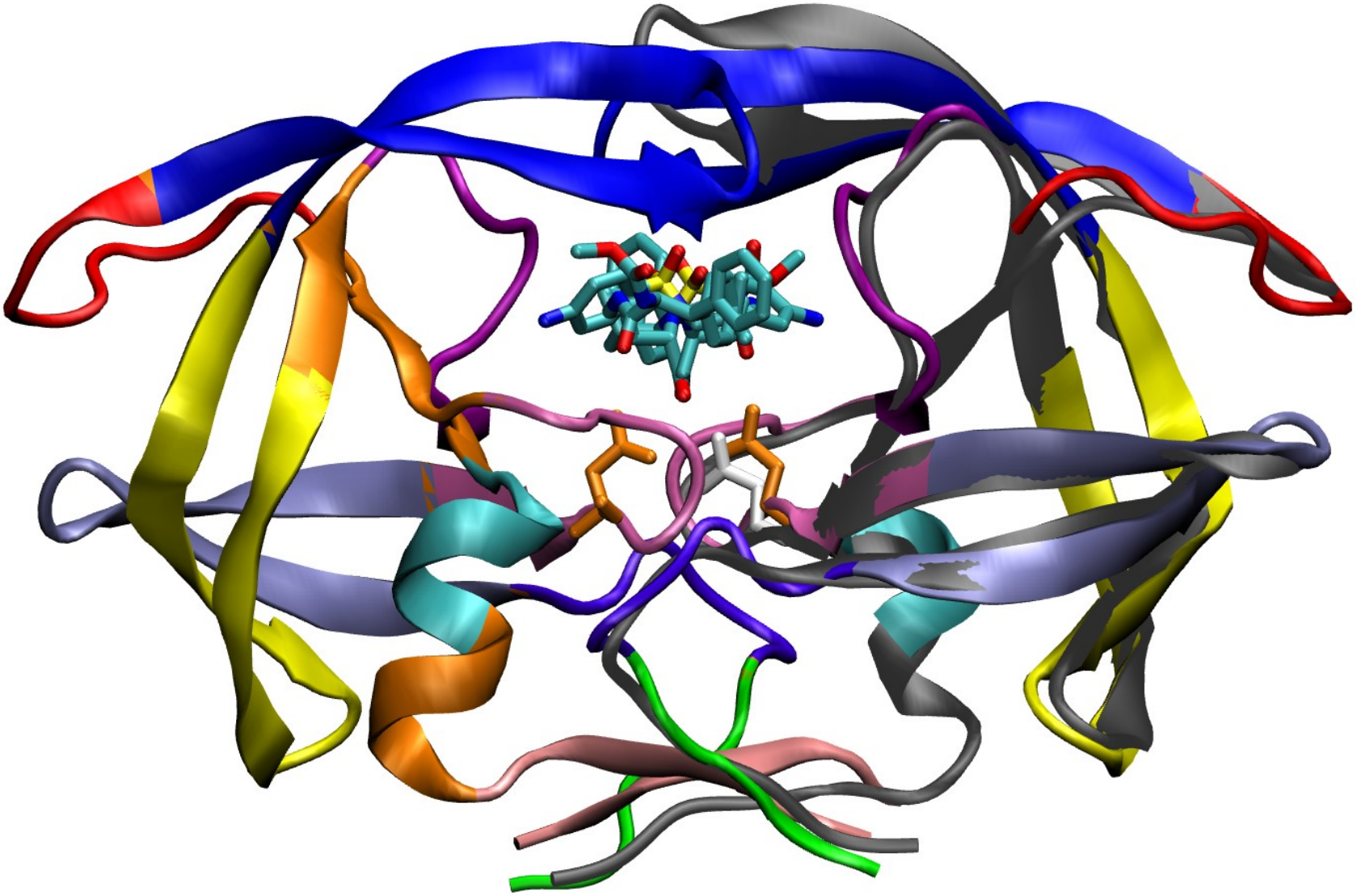
Grand Challenges in Global Health

Loria

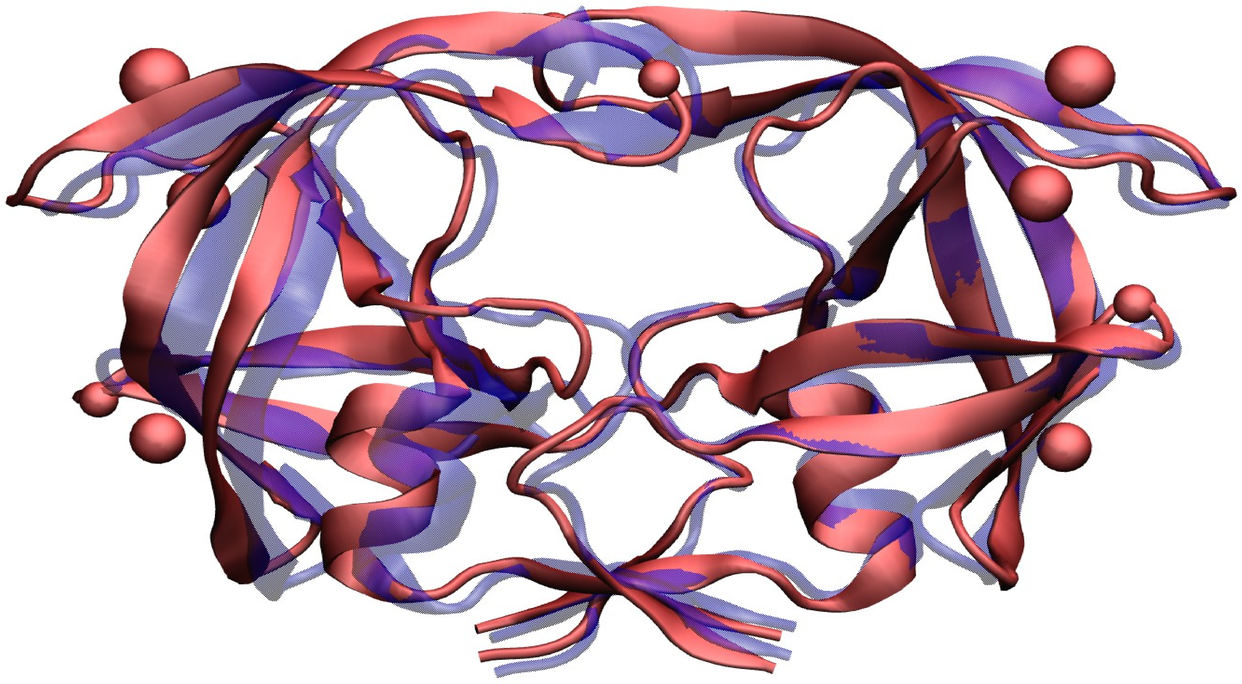
BILL & MELINDA GATES foundation


A
Results
HIV-PDI
To Overcome Drug Resistance
Ligand List
Results(5879)
NPWWPWBIYMJHSA
NPYBJTUGBJOYTF-U
NPZKQTSBTNUDPB-U
NQBIEJFZWHMMOA
NQDJWVGNAIXKLL-U
NQDJXKOVJZTUJA-UI
NQGLQOSLNZHQOB
NQHXCOAXSHGTIA-
NQLMBWKHBNXBKN
NQLVKCBKOFXVLN-U
NQMOSGJYMUFNHV
NQOMHRPWFKKMK
NQPKIDKDGLGDFI-U
NQUKYTXPMUKVOZ
NQZSQDCZUCSNJO-
NRAZTZTXJMWXIG-I
NRFCWJMWYNVFJT-
Display
B
Ligand
HIV-PDI
To Overcome Drug Resistance
HIV-PDI (Protein-Drug Interaction) Database
Ligand
PatentInChikey :
NQHXCOAXSHGTIA-GOXXAGPDBO
Name :
FDA-Viracept
Chemical Formul :
C32H45N3O4S.CH4O3S
Drug Function :
Inhibitor
Stage :
Commercial
Visualizations
1D
2D
3D
Resistance
Resistance ID
Treatment
Resistance ID
Patient
Patient ID
Mutation
Mutation ID
Interaction
Complex

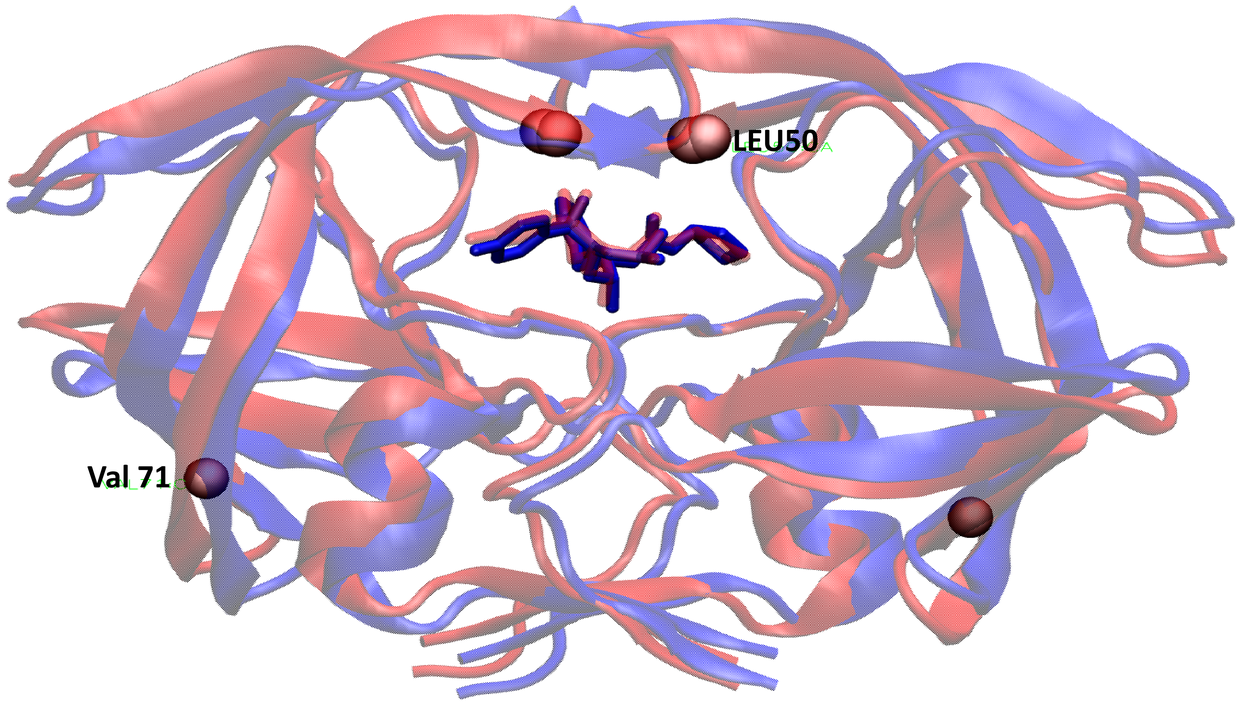

Flaps
Ear
Active Site
ASP 25
Cheek
Eye
Nose
Whiskers

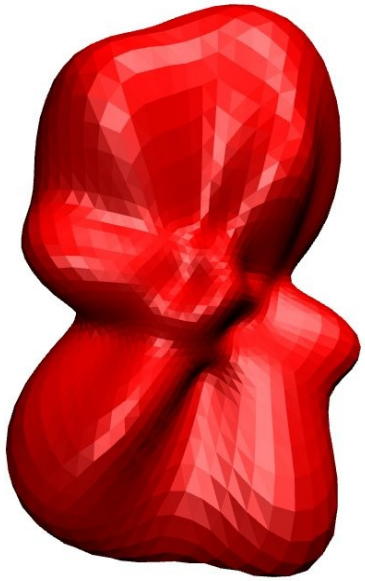


A

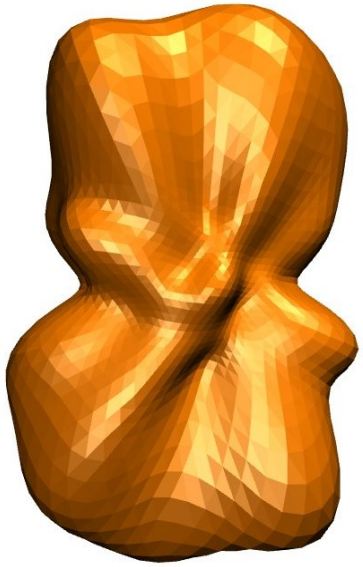


B

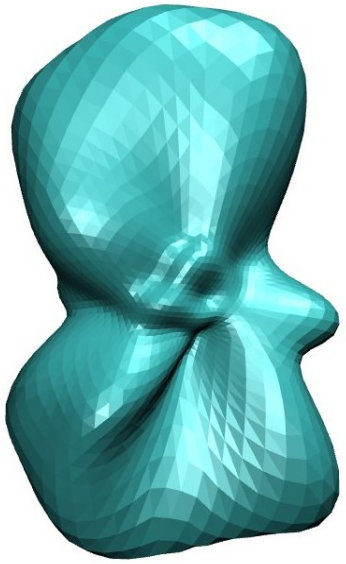




Surface: 770 A^2
Volume: 1171 A^3
3EKV pocket

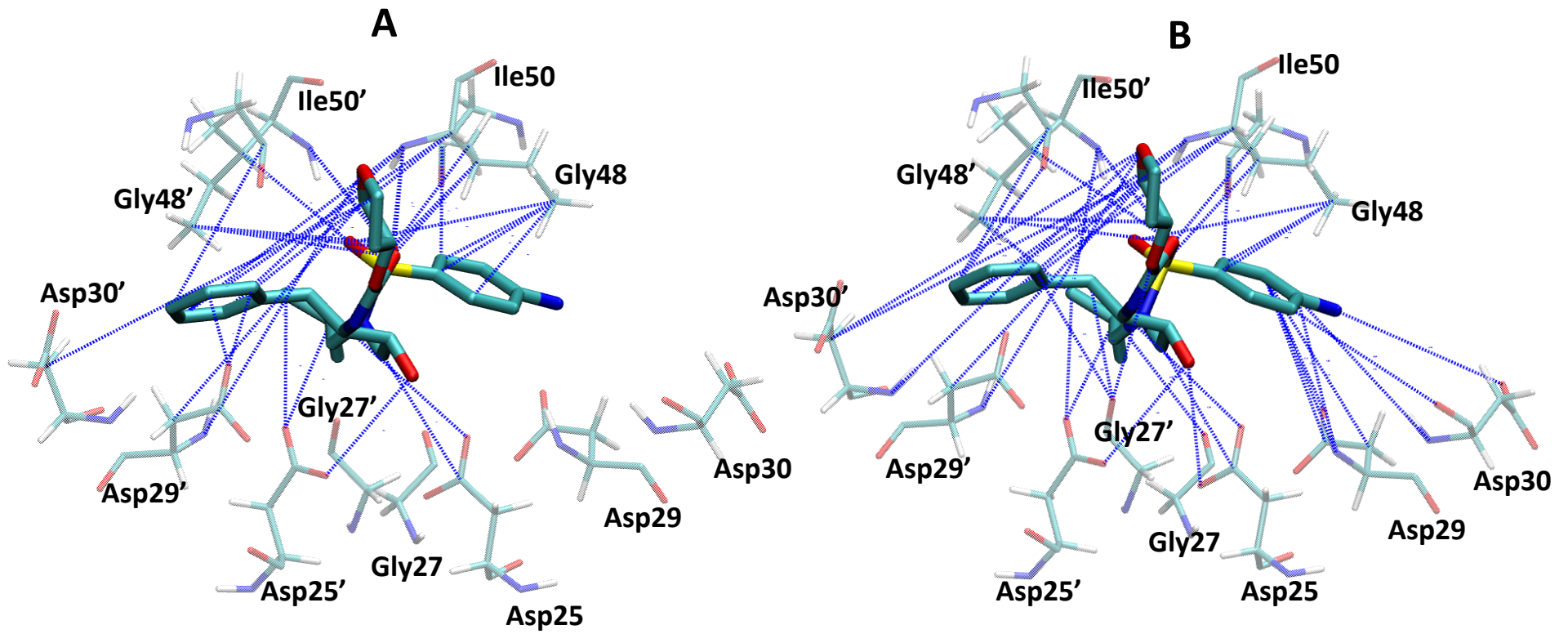


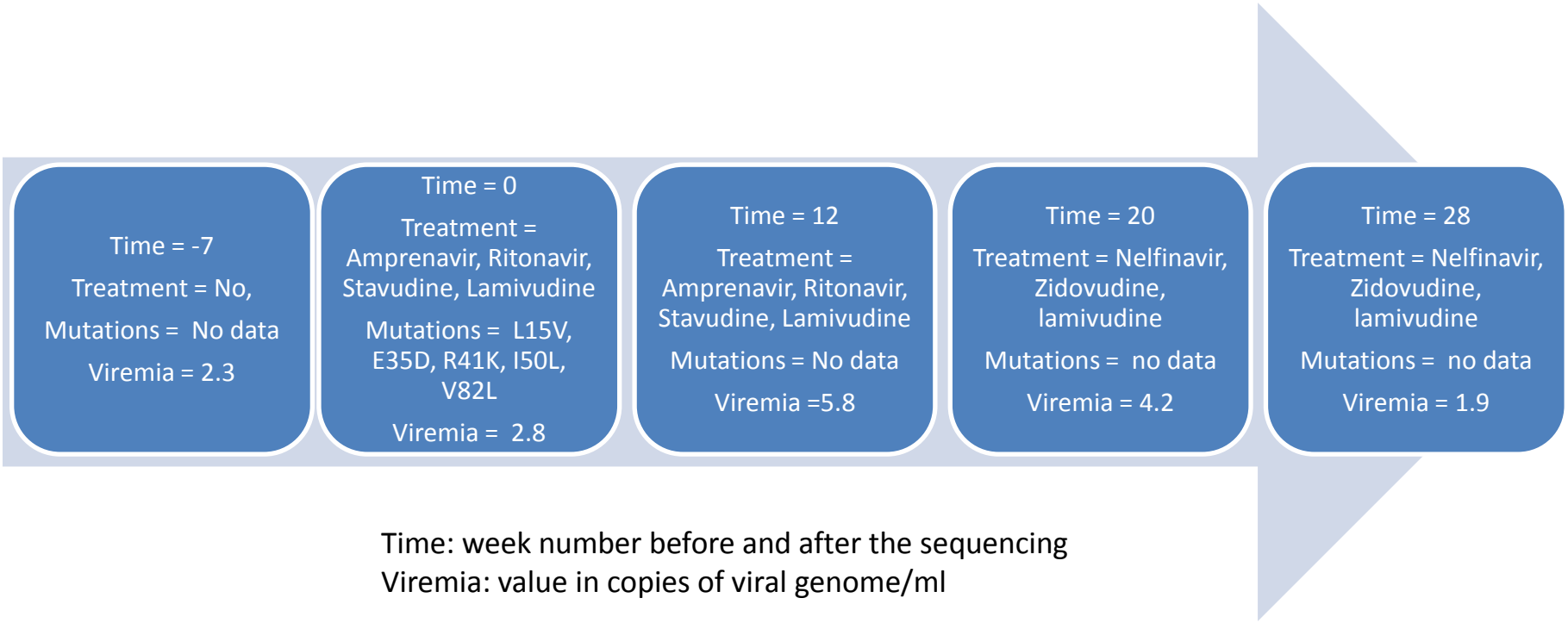
Surface: 783 A^2
Volume: 1202 A^3
3EM3 pocket



Surface: 663 A^2
Volume: 924 A^3
Patient 39546 Pocket

Combination of 3 pocket


A
Ile50
Ile50'
Gly48
Gly48'
Asp30'
Gly27'
Asp29'
Asp30
Asp29
Gly27
Asp25'
Asp25
B
Ile50
Ile50'
Gly48'
Gly48
Asp30'
Asp29'
Gly27'
Asp30
Asp29
Gly27
Asp25'
Asp25



Time: week number before and after the sequencing
Viremia: value in copies of viral genome/ml