Bruno Villoutreix

Recherche de sondes pharmacologiques et candidats-médicaments dans le cyber-espace

Bruno Villoutreix est directeur de recherche à l'Institut national de la santé et de la recherche médicale (INSERM)[1].

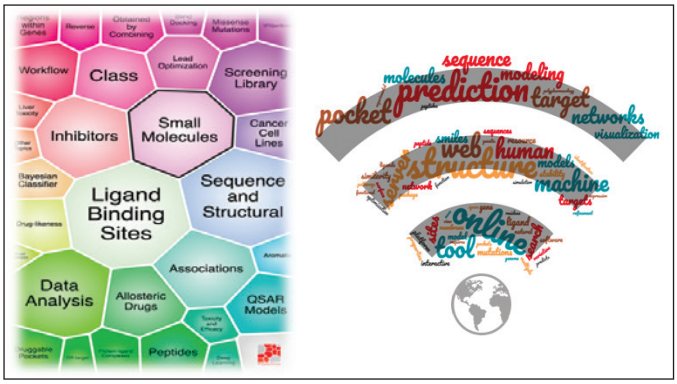
Le volume croissant de données biomédicales et chimiques en libre accès sur Internet ainsi que des logiciels et centres de calcul puissants permettant de les manipuler devraient aider à la découverte de sondes pharmacologiques et de candidats-médicaments.

Nous allons présenter les grands concepts du domaine, plusieurs bases de données et des outils logiciels en ligne qui facilitent l'étude des cibles thérapeutiques et des petites molécules chimiques, notamment le criblage virtuel, les prédictions ADME-Tox et le repositionnement *in silico* des médicaments (***Figure 1***).

1. www.inserm.fr

Figure 1

La croissance exponentielle des données récoltées dans le domaine de la pharmacologie et les multiples algorithmes publiés chaque semaine permettent de réaliser des calculs divers sur les macromolécules et les composés chimiques, et de générer de nouvelles connaissances.



1 Introduction à la pharmacologie informatique

Le processus de découverte d'un nouveau médicament s'étale sur plus de dix ans, comme schématisé sur la ***Figure 2***. La réalité est infiniment plus complexe que sur ce type de schéma : on commence en théorie par chercher des cibles (« *targets* » en anglais) potentiellement thérapeutiques ou des voies de signalisation dérégulées, puis on essaye de trouver des petites molécules capables de moduler ces cibles (« *hits* » en anglais). Ensuite, ces molécules doivent être optimisées à plusieurs niveaux, notamment sur les plans pharmacocinétique, pharmacodynamique et toxicologique (ADME-Tox[2]), avant d'être testées en phase clinique.

Ce processus est très long, très coûteux, et connaît beaucoup d'échecs. Les approches *in silico* devraient pouvoir aider et permettre de générer des hypothèses à chacune des étapes. Dans une revue

2. ADME-Tox : les propriétés d'Absorption, de Distribution, de Métabolisme, d'Excrétion et de Toxicité des petites molécules doivent être optimisées.

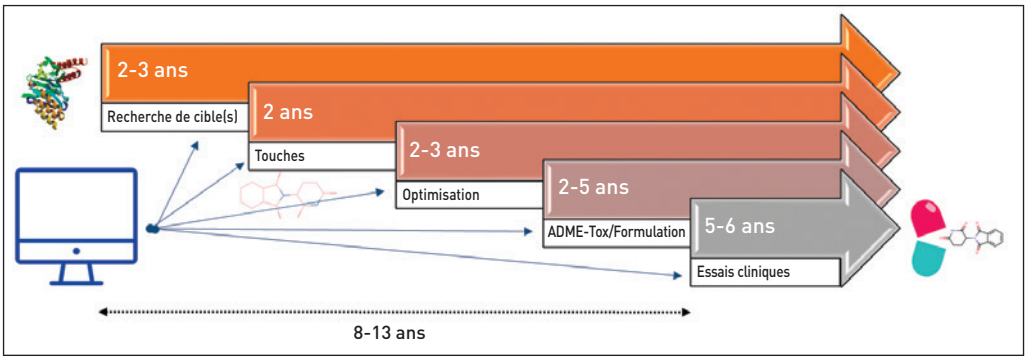


Figure 2

Le processus de développement d'un médicament n'est en réalité pas aussi linéaire que suggéré par le diagramme. Il comprend plusieurs étapes complexes qui s'étalent sur plus de dix ans (NB : la longueur des flèches n'est pas proportionnelle à la durée).

de Bayer publiée en 2015, on note que déjà 50 % des vingt nouvelles entités chimiques testées en phase clinique 1 avaient bénéficié d'approches bioinformatiques et chémoinformatiques.

Dans les années 2000, il existait environ 300 URL (adresse web qui permet d'identifier la ressource, son emplacement et le protocole Internet pour la récupérer) qui renvoyaient vers des bases de données et des logiciels dans le domaine du médicament au sens large. En 2019-2020 il y en a environ 3 500. Cette augmentation conséquente représente des millions d'heures de travail réalisées par des milliers de scientifiques dans le monde entier. Comme il est assez difficile d'identifier tous ces outils et services, une petite base de données a été créée, http://www.vls3d.com, pour les répertorier.

Certains de ces outils ne sont pas directement accessibles en ligne et doivent être installés localement. Cependant, nous nous focaliserons essentiellement ici sur les logiciels et bases de données facilement utilisables en ligne permettant de manipuler les cibles thérapeutiques comme certaines protéines et les petites molécules chimiques. Il est néanmoins important de souligner qu'il existe d'autres outils, pour par exemple faciliter le développement d'autres types de médicaments, comme les anticorps monoclonaux[3].

Les services web que nous allons aborder peuvent se regrouper en grandes catégories (***Figure 3***).

Parmi ces outils, 34 % concernent des approches pour identifier et analyser les cibles thérapeutiques et les protéines, 19 % prédisent les poches de fixation des candidats-médicaments, environ 20 % concernent le criblage[4] virtuel, 13 % les prévisions pharmacocinétiques et toxicité (« ADME-Tox ») et 5 % permettent de développer des modèles statistiques prédictifs. Une dizaine de pourcents des sites en ligne sont dédiés au repositionnement des médicaments, et on note qu'environ 8 % de ces sites sont des bases de données consacrées aux petites molécules chimiques.

Les principaux pays dans le monde qui offrent des services *in silico* dans le domaine du médicament, petites molécules et macromolécules, sont visibles sur la ***Figure 4***. Ces pays ont été identifiés *via* de multiples recherches, notamment avec PubMed, un moteur de recherche donnant accès à la base de données bibliographique MEDLINE rassemblant en 2020 environ 30 millions d'articles scientifiques spécialisés en biologie, chimie, approches *in silico* et médecine.

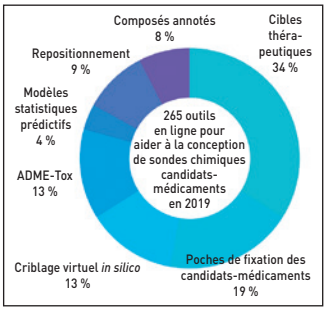


Figure 3

Répartition en 2019-2020 des 265 principaux outils en ligne et bases de données dans le domaine du médicament. Ces approches peuvent par exemple aider à la conception de molécules thérapeutiques, à la recherche de cibles impliquées dans une pathologie ou pour prédire la toxicité de certains composés chimiques.

3. Anticorps monoclonaux : anticorps produits par des lymphocytes clonés partir d'une unique cellule, ce qui résulte en une homogénéité des protéines obtenues.

4. Criblage : en pharmacologie, désigne généralement les techniques visant à identifier dans une chimiothèque des petites molécules chimiques biologiquement actives pouvant servir de base au développement de candidats-médicaments.

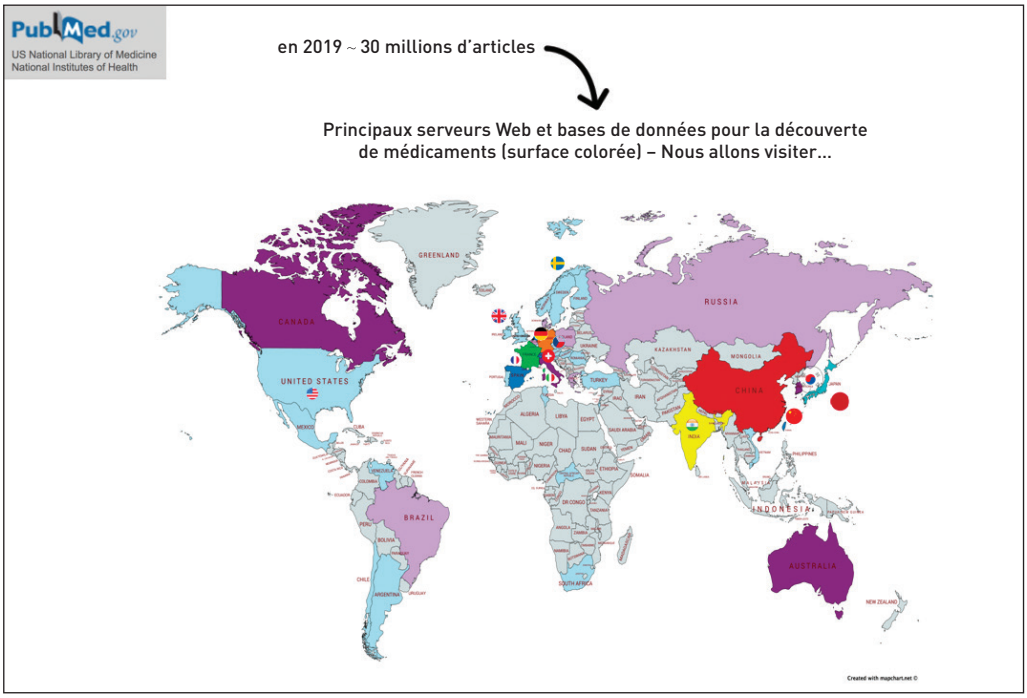


Figure 4

Les serveurs web dédiés à la conception de molécules thérapeutiques et aux études des petites molécules et macromolécules émanent de nombreux États, répartis sur les cinq continents.

2 Cibles thérapeutiques et poches de fixation des ligands

La recherche de cibles, souvent des protéines, est une étape critique dans la recherche de nouveaux agents thérapeutiques. Les connaissances sur ces cibles d'intérêt thérapeutique et des voies de signalisation émanent des travaux de multiples laboratoires de recherche dans le monde et de grands programmes internationaux. À titre d'exemples et en nous limitant à des projets américains et européens récents, nous pouvons mentionner plusieurs bases de données qui permettent de recueillir, structurer et organiser l'information sur les cibles potentiellement thérapeutiques, et la rendre accessible à tous (***Figure 5***).

Dans ces bases de données, les scientifiques vont notamment chercher si certaines cibles protéiques sont déjà connues pour être impliquées dans une pathologie, quelles sont les cibles ou les voies de signalisation qui ne sont pas encore modulées par des médicaments, et celles pour lesquelles on pourrait faire de la conception (« *design* ») de nouvelles molécules (PHAROS et Open Targets).

Ces bases de données peuvent être spécialisées sur les cibles pharmacologiques (IUPHAR), et certaines entièrement dédiées au vieillissement

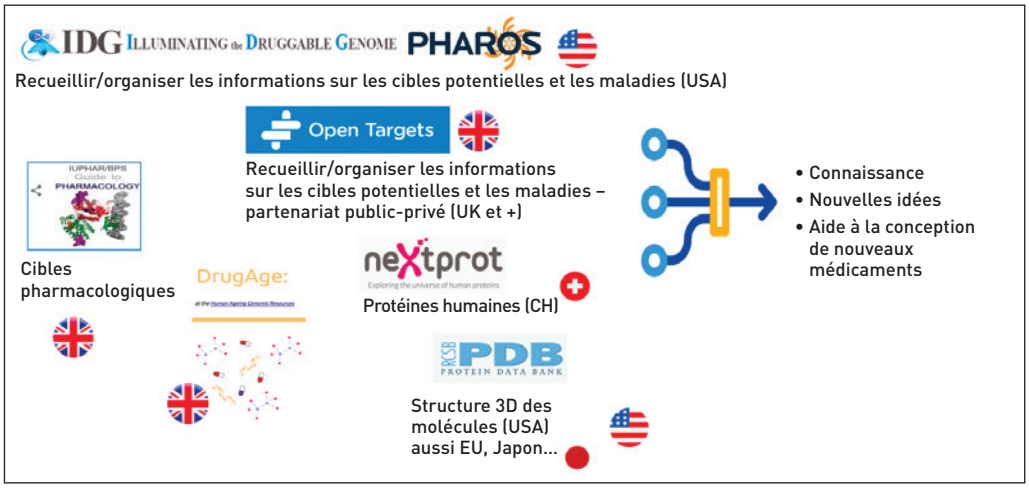


Figure 5

Des bases de données pour l'innovation visant à rassembler les connaissances sur les cibles potentiellement thérapeutiques émanent souvent de grands projets internationaux.

(DrugAge). En Suisse, neXtProt est une base de données spécialisée sur les protéines humaines. La Protein Data Bank (PDB) répertorie les structures 3D des molécules identifiées essentiellement par cristallographie des rayons X[5] ou RMN[6]. Ce type d'information va permettre de générer de nouvelles idées et de se focaliser sur certaines niches encore très peu étudiées.

Pour qu'une petite molécule chimique puisse moduler l'activité d'une protéine cible (par exemple une protéine surexprimée dans une pathologie), il faut en général que celle-ci présente une poche de fixation ; il existe des algorithmes qui permettent de prédire ces poches et les zones importantes pour les interactions moléculaires. Ces outils peuvent faire des prédictions avec juste la séquence d'acides aminés de la protéine, mais ils sont plus précis si l'on connaît la structure tridimensionnelle de la macromolécule en question (***Figure 6***).

Ces algorithmes prévisionnels sont basés soit sur la géométrie de la macromolécule (c'est le cas de services en ligne français comme Fpocket, ou indien comme PocketDepth), soit ils sont basés sur des calculs d'énergies. En effet, lorsqu'on bombarde la surface d'une macromolécule potentiellement impliquée dans une pathologie avec une petite liste d'atomes ou des fragments chimiques, il est possible de construire des cartes d'affinité d'interaction, et définir certaines zones appelées « hotspots », où des candidats-médicaments s'accrochent préférentiellement.

Ces zones sont donc utilisées pour faire de la conception de molécules capables de s'y accrocher.

5. Cristallographie des rayons X : technique d'analyse structurale fondée sur la diffraction des rayons X.
6. RMN : Résonance Magnétique Nucléaire, phénomène basé sur les propriétés magnétiques des spins nucléaires, à l'origine d'une technique de spectroscopie permettant la détermination de structures moléculaires.

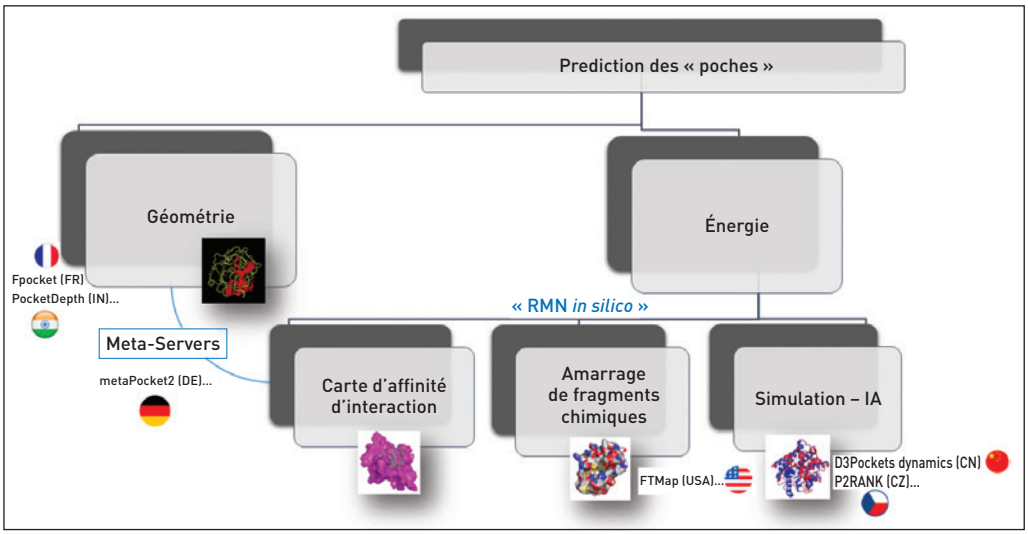


Figure 6

Deux grands types d'approches pour la prédiction des « poches » existent quand on connaît la structure 3D de la cible : les approches géométriques et les approches énergétiques. Ces méthodes permettent par exemple de définir des zones où une petite molécule pourrait se fixer et ainsi fournir des informations précieuses pour d'autres types de calculs, comme le criblage virtuel.

Par exemple dans une approche comme FTMap aux États-Unis, des petits fragments chimiques bombardent virtuellement la surface de la protéine cible et, à partir de calculs d'affinité, il est alors possible de prédire des zones où des petites molécules chimiques peuvent s'accrocher.

Dans ce type de prédiction, il faut cependant aussi prendre en compte la flexibilité des macromolécules. Il existe ainsi des serveurs dédiés à l'étude de la flexibilité des poches, notamment en Chine avec D3Pockets. De plus, certains outils utilisent des méthodes d'apprentissage automatique, comme P2RANK en Tchécoslovaquie pour prédire ces poches, et combinent alors des approches statistiques, géométriques et/ou énergétiques.

Pour illustrer la notion de flexibilité, prenons l'exemple du système interleukine et son récepteur impliqué dans un certain nombre de pathologies. Bloquer cette interaction protéine-protéine avec une petite molécule chimique pourrait être intéressant pour le développement de nouveaux médicaments, mais il faut trouver la bonne poche. Il existe plusieurs structures 3D de l'interleukine et en fonction de la structure sélectionnée, on distingue que la poche à la surface présente des cavités et des protubérances variables (***Figure 7***).

Par exemple, si l'on compare la protéine cristallisée seule (forme apo) et la protéine cristallisée en présence d'une petite molécule chimique (forme holo), il est possible de voir qu'une protubérance est présente dans la forme apo juste au milieu de la poche où se fixe le candidat-médicament. Cela correspond à un acide aminé qui bouge et se réoriente pour que le composé chimique puisse se fixer. Ainsi, si un scientifique utilise la structure apo pour rechercher une petite molécule modulatrice, les chances

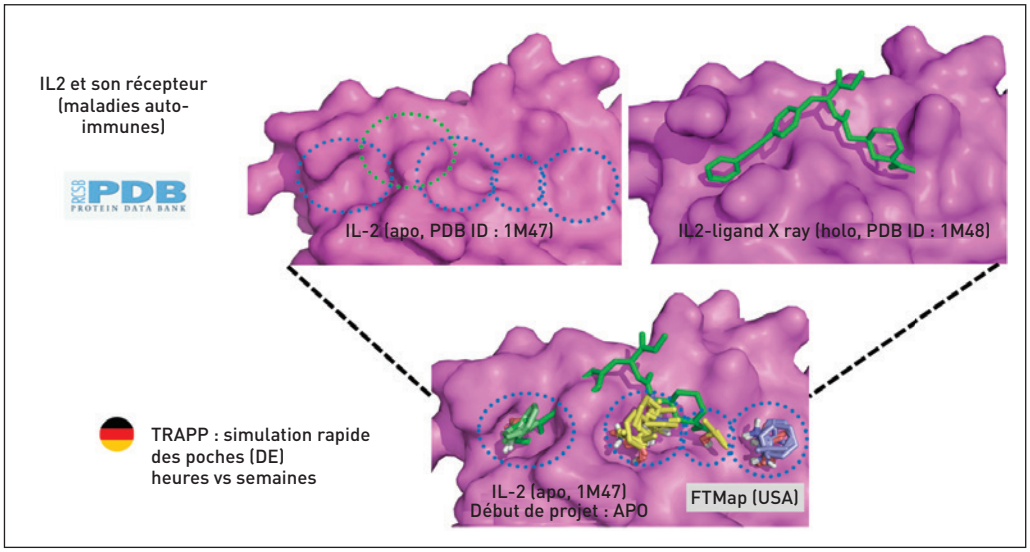
de succès sont pratiquement nulles. Pourtant, au début d'un projet, seule la structure apo d'une protéine est généralement connue (structure expérimentale ou prédite par des approches théoriques). Comment faire dans cette situation ? Il est possible d'utiliser un outil comme FTMap, qui va prédire des poches, même sur la forme apo de la protéine. Les petits fragments sont positionnés pratiquement dans toute la poche de fixation du ligand sauf au niveau de la protubérance (***Figure 7***). Il faudra ensuite utiliser un outil d'ouverture de poche, comme le serveur TRAPP en Allemagne, pour générer des conformations alternatives de la poche. Ainsi, en combinant les approches, même à partir d'une structure apo, il est possible de prédire plusieurs poches et de les cribler ensuite *in silico* dans l'espoir d'identifier des petits composés chimiques qui iront moduler l'activité biologique de la cible en question.

Il est aussi possible d'obtenir des informations sur l'importance d'une protéine et/ou d'une région de la protéine, ou encore de la poche de fixation, dans une pathologie en analysant les données du séquençage stockées dans de nombreuses bases de données.

Prenons l'exemple d'une protéine anticoagulante appelée antithrombine, qui fixe des médicaments comme le fondaparinux (un pentasaccharide anticoagulant qui se lie à l'antithrombine, et en se liant à cette protéine, le facteur Xa est inhibé). Nous avons travaillé avec des collaborateurs aux États-Unis et en Chine sur des familles de patients présentant des problèmes de coagulation ; ces personnes avaient une mutation ponctuelle dans certaines régions de l'antithrombine. Il existe de très nombreuses approches *in silico* pour analyser ce type de mutations. Un des objectifs est d'essayer de comprendre la relation entre le changement

Figure 7

À l'aide des données disponibles dans la PDB, il est possible de déterminer la zone de fixation d'une petite molécule chimique sur une protéine avec un outil comme FTMap et d'explorer la flexibilité possible de la poche identifiée avec par exemple le serveur TRAPP.
Source : Structures 3D des protéines utilisées dans cet exemple : Arkin *et coll.* (2003). *PNAS.*



d'acide aminé et le phénotype observé chez le malade. Nous avons dans un premier temps localisé les mutations ponctuelles identifiées chez les patients sur la structure 3D de la protéine avec des logiciels gratuits comme Chimera (https://www.cgl.ucsf.edu/chimera/) ou PyMol (https://pymol.org/) (***Figure 8A***). Dans ce cas précis, nous avions l'avantage de disposer de la protéine co-cristallisée avec un médicament. On pouvait donc voir qu'une mutation était directement dans la poche de fixation du médicament et pourrait donc altérer la liaison entre la petite molécule et la protéine. L'autre mutation était plus loin de cette poche, elle pourrait jouer plusieurs rôles, sur la stabilité de la protéine ou encore sur la dynamique du système avec perturbation à distance de la fixation du médicament, etc.

Pour aller plus loin dans l'analyse des mutations ou dans le cas où il n'y a pas de structure 3D protéine-médicament, il peut être intéressant de calculer, à partir du fichier PDB de la protéine étudiée, toutes interactions non covalentes entre tous les acides aminés de la protéine et créer ainsi un réseau d'interactions (***Figure 8B***). Cette analyse a été réalisée avec le serveur Italien RING. Chaque cercle représente un acide aminé (les nœuds du réseau) et, entre les cercles, les petits traits (les arêtes du réseau) représentent les interactions non covalentes. Il est alors possible de visualiser en 2D des informations tridimensionnelles telles que les liaisons hydrogène, les ponts salins… (***Figure 8B***). On peut ensuite transformer ces données d'interactions complexes en un autre type de visualisation avec le logiciel gratuit Cytoscape (https://cytoscape.org/) : ici tous les acides aminés sont projetés sur un cercle (petits points rouges) et ils sont classés selon le nombre d'interactions non covalentes (***Figure 8C***). Les résidus qui ont un grand nombre d'interactions non covalentes ont généralement un rôle majeur dans la structure et stabilité de la protéine. Une mutation dans ces régions peut entraîner un déficit de la protéine étudiée et une pathologie associée à la fonction de celle-ci. Une mutation d'un résidu faiblement connecté avec son environnement mais qui a un effet important sur le traitement peut donner des pistes sur l'importance de certaines régions de la poche de fixation du médicament et aider à la conception de nouvelles molécules si la mutation est fréquente dans la population. Dans le cas de l'antithrombine, les résultats des approches *in silico* suggéraient une perturbation directe et indirecte de la fixation du médicament induite par les mutations, en accord avec les travaux expérimentaux.

Plus généralement, en analysant les substitutions d'acides aminés chez les patients avec plusieurs outils informatiques et diverses approches expérimentales (biochimie, biologie moléculaire et biophysique), il devient possible dans certains cas de comprendre pourquoi les malades réagissent ou pas à un traitement, et ainsi de générer des nouvelles

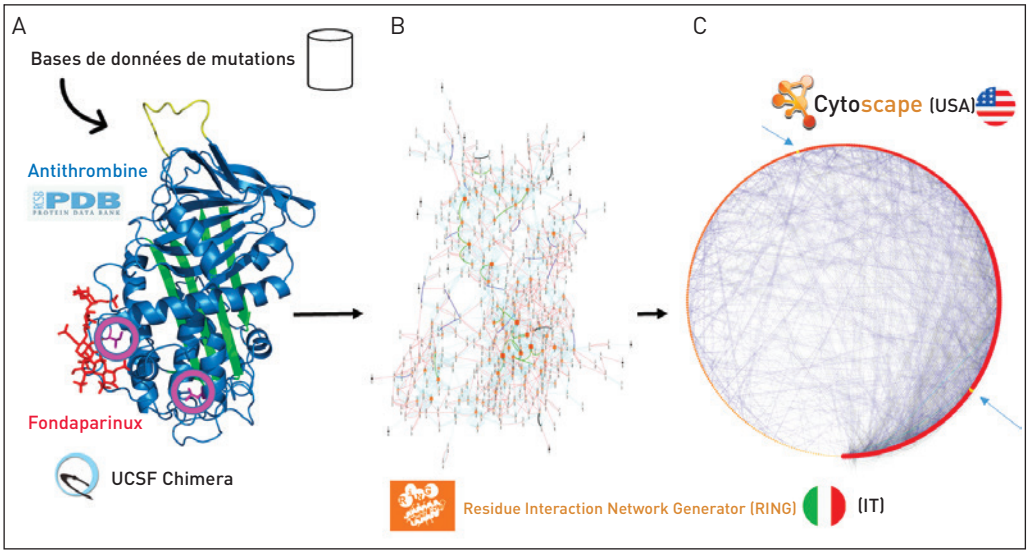


Figure 8

Les outils de visualisation permettent de gagner en connaissance. A) Structure 3D de la protéine antithrombine co-cristallisée avec un médicament anticoagulant, le fondaparinux, visualisée avec le logiciel Chimera. Des mutations ponctuelles identifiées dans des bases de données ou après séquençage de patients sont soulignées par des cercles ; B) visualisation du réseau d'interactions non covalentes présent dans cette protéine avec localisation des acides aminés mutés après traitement par le logiciel RING ; C) utilisation du logiciel Cytoscape pour analyser les données de RING afin de caractériser les interactions non covalentes de chaque acide aminé. Ce type d'analyse aide à comprendre l'impact des substitutions sur la structure et la fonction de la protéine, et dans certains cas facilite la conception de candidats-médicaments.

Source : Dinarvand *et coll.* (2018). *J. Thromb Haemost*.

hypothèses pour changer de traitement ou pour rechercher de nouveaux médicaments.

Toutes ces bases de données dédiées aux cibles et tous ces outils logiciels permettant de les analyser vont donc constituer une étape importante dans la recherche de candidats-médicaments ou de sondes pharmacologiques. L'étape suivante consiste à rechercher des petites molécules qui modulent ces cibles.

3 Les petites molécules et le criblage virtuel

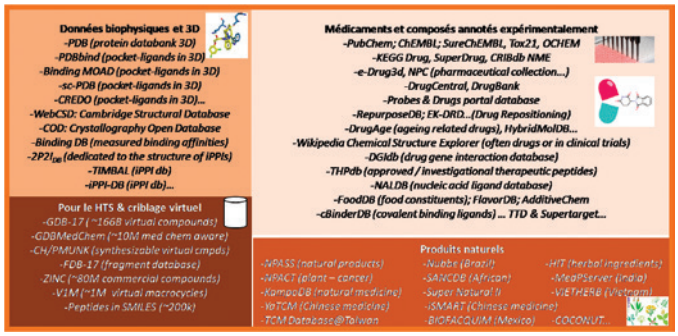
Les petites molécules chimiques sont aussi stockées dans un certain nombre de bases de données. On y trouve des petites molécules provenant des extraits de plantes : des plantes utilisées pour la médecine traditionnelle chinoise, des plantes utilisées pour la médecine ayurvédique[7], d'autres venant d'un certain nombre de régions du monde (***Figure 9***).

D'autres bases de données donnent accès aux propriétés physiques de ces petites molécules chimiques. Dans la « Protein Data Bank », on trouve des macromolécules et des petites molécules co-cristallisées avec une cible

7. Médecine ayurvédique : médecine traditionnelle originaire de l'Inde.

Figure 9

De très nombreuses bases de données spécialisées et contenant des informations au format adapté pour le traitement informatique sont disponibles sur Internet.



thérapeutique ou une macromolécule. Dans la Cambridge Structural Database ou dans la Crystallography Open Database, les structures 3D expérimentales de plus d'un million de petites molécules chimiques sont répertoriées. Certaines bases de données sont spécifiquement dédiées aux inhibiteurs d'interactions protéine-protéine. On trouve aussi des petites molécules virtuelles, c'est-à-dire qui n'ont pas encore été synthétisées. C'est le cas de la base de données Suisse (GDB-17), qui contient 166 milliards de molécules qui n'ont pas encore été synthétisées. La navigation dans cet espace chimique quasiment infini, avec des outils informatiques, va certainement permettre à terme de générer de nouvelles connaissances et des nouveaux médicaments.

Afin d'identifier rapidement des molécules chimiques qui modulent les cibles thérapeutiques potentielles, il est pertinent d'acheter les composés car on ne peut pas tous les synthétiser. En effet, pour synthétiser 1 000 composés en quantité suffisante, il faut souvent plusieurs mois de travail. Pour aider les scientifiques, plusieurs bases de données contiennent des catalogues de molécules que l'on peut acheter à des sociétés de chimie. Par exemple, à San Francisco, la base ZINC répertorie environ 80 millions de petites molécules existantes auprès des vendeurs et plus d'un milliard de composés virtuels facilement synthétisables. On peut aussi sélectionner pour certains projets des bases de données plus spécifiques comme par exemple celles qui contiennent uniquement des macrocycles, ou uniquement des petits peptides.

Il existe également des bases de données contenant des médicaments qui sont déjà sur le marché et des molécules en phase clinique comme DrugBank au Canada. Certaines bases répertorient des petites molécules qui ont été testées avec des approches de criblage expérimental haut débit, comme par exemple PubChem et ChEMBL. Ces bases de données annotées sont très intéressantes pour réaliser des études grande échelle *in silico* sur par exemple tout le protéome humain, ou pour développer des modèles statistiques prédictifs. D'autres bases de données concernent les extraits de produits alimentaires ou

des molécules annotées sur Wikipédia.

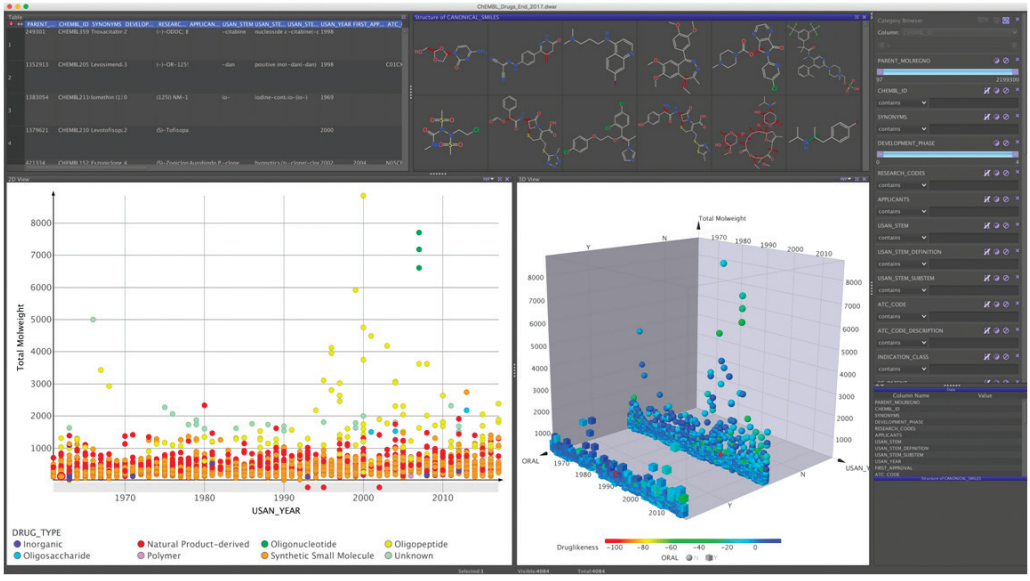
Une quantité d'informations considérable est donc maintenant disponible et augmente de jour en jour. Ces informations ne peuvent pas être traitées par le cerveau humain. Il faut donc des logiciels et des applications informatiques pour manipuler les données et en extraire de la connaissance.

Le logiciel DataWarrior (http://www.openmolecules.org/datawarrior/), gratuitement disponible, permet de manipuler assez facilement autour d'un million de molécules. Il fonctionne sur tous les systèmes informatiques. L'utilisateur peut ouvrir un fichier de molécules téléchargé d'une base de données ou récupérer directement des molécules sur ChEMBL ou Wikipédia (***Figure 10***). Des tutoriels en français sont disponibles sur Radar web création (https://www.radarweb.fr/), ils expliquent comment prendre en main ce logiciel un peu complexe. En ouvrant DataWarrior, plusieurs fenêtres sont visibles. Une sorte de tableau Excel décrit les composés, avec des représentations 2D et/ou 3D des molécules. Dans l'exemple sur la ***Figure 10***, une base de médicaments est ouverte et visualisée *via* des graphes en 2D et 3D. Ces molécules peuvent alors être classées par années, en fonction de leur origine, naturelle ou synthétique, de leur nature biochimique, de leur poids moléculaire, ou encore de leur mode d'administration.

Le criblage expérimental (ou « *high-throughput screening* », HTS, pour souligner le caractère haut débit de l'approche) est une technique de référence pour la recherche de petites molécules agissants sur une cible. Dans la pratique, il consiste à tester en parallèle l'action de dizaines de milliers de petites molécules chimiques sur une cible que

Figure 10

Le logiciel de visualisation et d'analyse de données DataWarrior (www.openmolecules.org ; tutoriel : www.radarweb.fr) est disponible gratuitement et marche sur tous les OS.



l'on estime importante pour une pathologie ou une fonction biologique au moyen d'un automate dans le but d'identifier des touches qui vont agir sur la cible.

Avec la miniaturisation et l'automatisation des tests biologiques, l'automate peut évaluer 100 000 molécules par jour environ (voire plus), mais les coûts sont énormes. Le criblage virtuel, dans sa version la plus simple, transpose *in silico* certaines idées du criblage expérimental. Le criblage virtuel a pour objectif principal de réduire le nombre de molécules à tester expérimentalement. Ainsi, une chimiothèque de plusieurs millions (ou milliards) de composés ou de milliers de médicaments peut être utilisée.

Ensuite, le criblage virtuel *per se* sera initié. On distingue deux grandes stratégies de criblage virtuel qui peuvent être couplées dans certaines circonstances : celles qui utilisent les propriétés structurales des petites molécules chimiques bioactives – on parle alors de criblage virtuel basé sur la structure des ligands (« *ligand-based virtual screening* », LBVS) – et des approches basées sur la structure tridimensionnelle (3D) de la cible (« *structure-based virtual screening* », SBVS).

Dans la pratique et à titre d'exemple, à partir d'une chimiothèque électronique initiale contenant un million de petites molécules, le criblage virtuel permettra, en quelques heures de calculs (ou en quelques minutes de calcul sur un cluster puissant), de générer une liste d'environ 500 composés potentiellement actifs à tester expérimentalement. Sur ces 500 molécules analysées en tubes à essai, il y aura immanquablement un nombre important de molécules inactives mais aussi une petite liste de molécules bioactives. Dès lors, au lieu de tester expérimentalement un million de composés, seules 500 molécules seront analysées dans le tube à essai.

4 Intelligence artificielle et apprentissage automatique

Les approches d'intelligence artificielle pour prédire l'importance d'une cible ou certaines propriétés des composés chimiques sont encore émergentes. Dans d'autres domaines que le médicament ou la chimie, on connaît par exemple le traitement du langage naturel, qui est un sous-domaine de l'intelligence artificielle. Ces approches regroupent par exemple les programmes de reconnaissance vocale et les autres applications liées aux mots dits ou écrits. Certaines de ces approches sont modifiées pour faciliter la recherche de molécules potentiellement thérapeutiques. L'intelligence artificielle est aussi connue dans le domaine du traitement et de la reconnaissance des images. Ces algorithmes peuvent être adaptés pour le traitement d'images de petites molécules chimiques et ainsi aider à par exemple prédire la toxicité d'un composé.

Toute une série d'autres algorithmes permet de classer ou de construire des

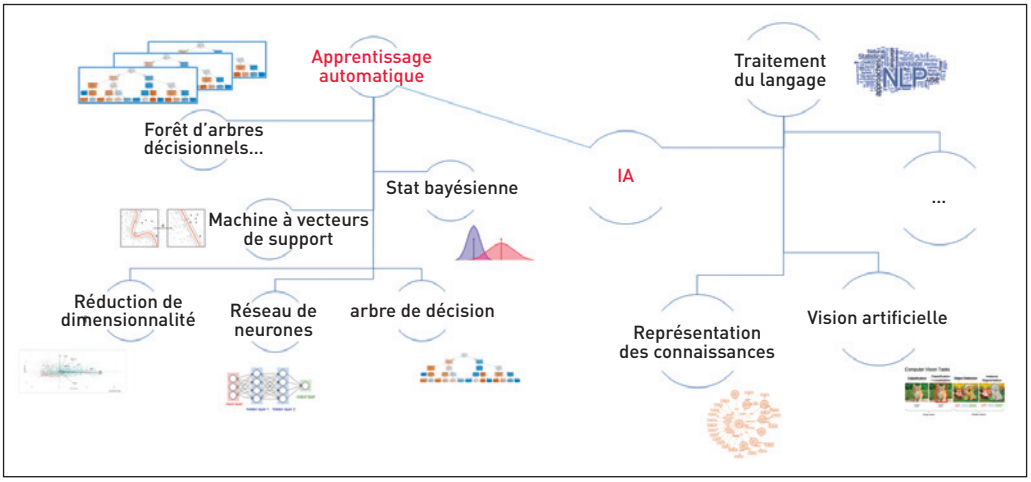
modèles statistiques, notamment de prédire si un composé peut être actif sur une cible ou toxique chez l'homme (***Figure 11***).

Dans le domaine de la chimie, certaines de ces approches sont plus connues sous le nom de méthodes QSAR (« *Quantitative Structure Activity Relationship* »). Ces techniques permettent de relier par une relation mathématique les descripteurs moléculaires, soit à l'activité biologique, soit à une propriété (physico-chimique ou pharmacocinétique). De tels modèles statistiques, basés sur des descripteurs moléculaires calculés, sont le point de départ de nombreux processus de sélection de molécules. Ces modèles sont souvent construits à partir d'un jeu de référence (par exemple des molécules annotées de ChEMBL), appelé jeu d'apprentissage, permettant de sélectionner le(s) descripteur(s) le(s) plus adapté(s) dans la construction d'un modèle. Les modes de construction de ces modèles peuvent être relativement simples (régression linéaire…) ou plus sophistiqués (algorithmes génétiques, réseaux de neurones, forêts aléatoires…). De nos jours, d'autres algorithmes peuvent être utilisés comme les réseaux de neurones convolutifs (exemple du « *deeplearning* »).

Le calcul des descripteurs et le développement des modèles statistiques sont complexes. Heureusement, certains services dédiés aux petites molécules sont disponibles en ligne, comme ChemSAR en Chine (***Figure 12***). Dans ce cas, l'utilisateur récupère les informations à partir d'une base de données comme ChEMBL, PubChem ou DrugBank. Les molécules sont insérées dans le système informatique qui nettoie les données et calcule toute une série de descripteurs moléculaires. Le système va ensuite utiliser différentes approches statistiques pour créer des modèles statistiques prédictifs. La qualité et la performance des modèles peuvent alors être visualisées. Le modèle statistique mathématique généré par le système

Figure 11

L'intelligence artificielle et les méthodes d'apprentissage automatique permettent d'apprendre des données et de développer des modèles mathématiques prédictifs. Certaines de ces prédictions visent à évaluer les propriétés ADME-Tox des petites molécules et ainsi réduire la nécessité de recourir à l'expérimentation animale.



de manière quasi automatique peut ensuite être utilisé et appliqué pour prédire l'activité de nouveaux composés.

5 Prédictions ADME et de la toxicité

Dans le cadre du développement thérapeutique, les composés chimiques doivent non seulement être actifs sur les bonnes cibles mais ils vont aussi devoir être absorbés, distribués, métabolisés, et excrétés (« ADEME »). De plus, ils doivent être peu ou pas toxiques. Le voyage d'un médicament dans l'organisme est représenté de façon très schématisée sur la ***Figure 13***. L'administration par voie orale d'un médicament fait intervenir le passage à travers toute une série de membranes biologiques et les petites molécules vont aussi interagir avec un grand nombre de macromolécules. Certains de ces processus et évènements peuvent être reproduit *in silico.*

Généralement, deux séries d'analyses sont effectuées sur les petites molécules. Une première étape implique le nettoyage des chimiothèques. Il s'agit de la standardisation de l'écriture des molécules et d'un filtrage ADME-Tox faisant appel à plusieurs règles empiriques s'appuyant souvent sur l'analyse de molécules connues et des médicaments existants. Nous avons développé un logiciel gratuit en ligne, FAF-Drugs, qui permet de préparer une chimiothèque avant un criblage ou de filtrer des molécules virtuelles avant la synthèse. Les fichiers sont soumis à toute une série de calculs et de filtres, à l'issue desquels les molécules sont soit rejetées car elles ne respectent pas des règles ADME simples ou parce qu'elles sont potentiellement toxiques, soit acceptées ou soit considérées comme de qualité intermédiaire (***Figure 14***). On dispose aussi de toute une série d'outils de visualisation qui permettent de mieux caractériser les produits. Des règles développées dans l'industrie pharmaceutique sont aussi

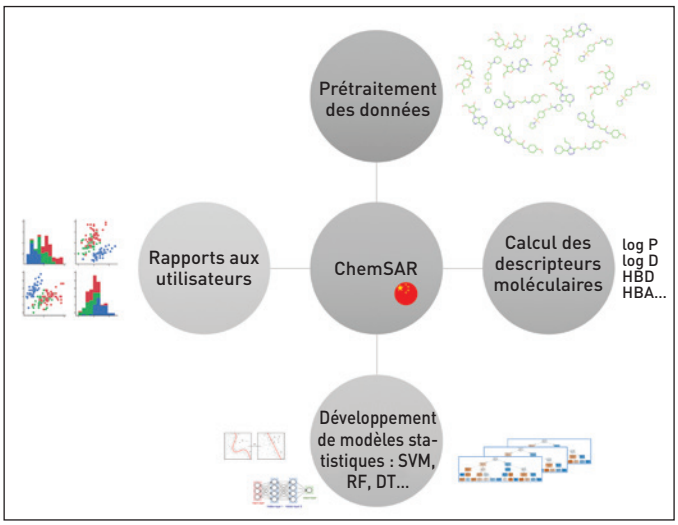


Figure 12

Des systèmes de gestion des données permettent un traitement automatisé des informations récoltées sur les bases de données afin de développer des modèles statistiques prédictifs.

Source: d'après Dong *et coll.* (2017). *J. Cheminform.*

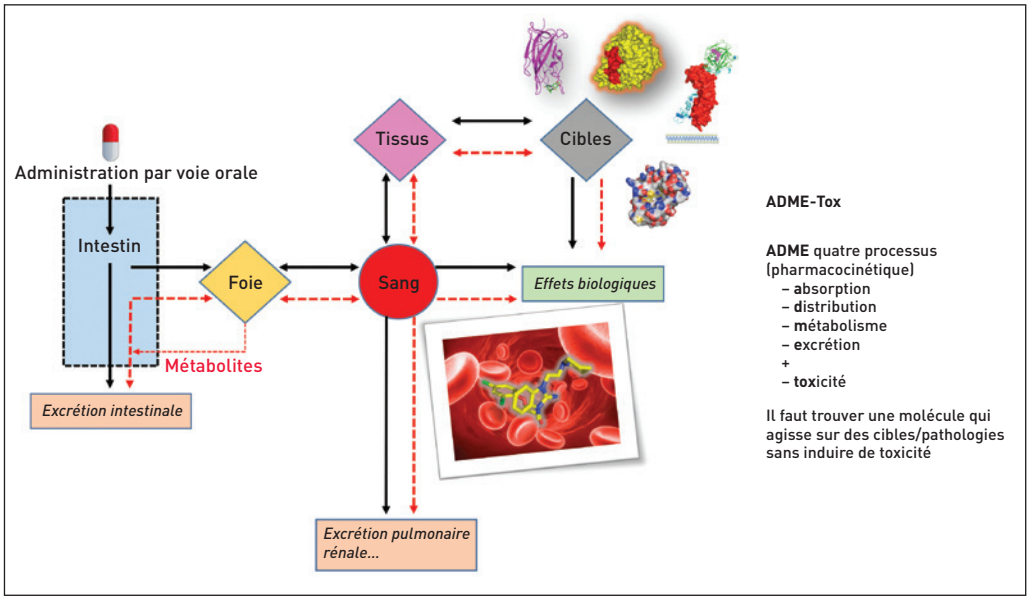


Figure 13

Le devenir d'un médicament administré oralement dans l'organisme : un voyage en plusieurs étapes. La pharmacocinétique, parfois désignée sous le nom de « ADME », étudie le devenir du médicament dans l'organisme après son administration. On distingue généralement plusieurs phases : absorption, distribution, métabolisme (transformation en produit actif ou inactif) et élimination (excrétion). Certaines de ces phases peuvent être prédites in silico. *De plus, il est important de prédire si un composé peut être toxique. Dans ce cas, certains modèles prédictifs sont utilisés en parallèle des approches expérimentales et peuvent même dans certaines circonstances remplacer l'expérimentation animale.*

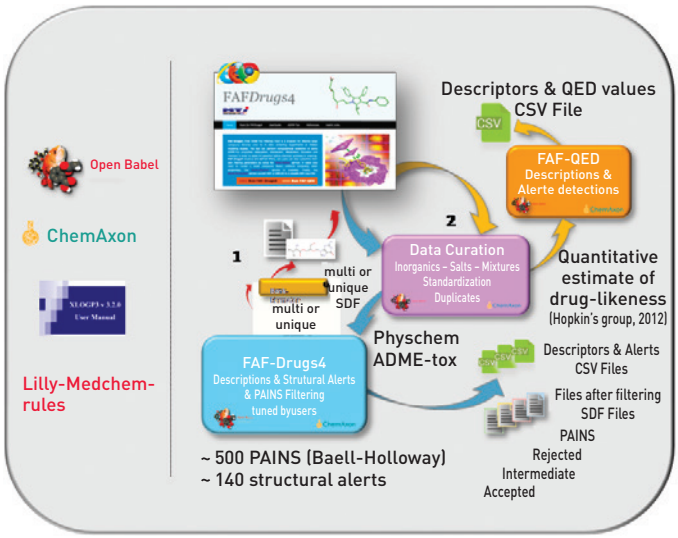


Figure 14

FAF-Drugs est un logiciel en ligne accessible gratuitement qui permet, grâce à un traitement de données, de prédire certaines propriétés ADME-Tox simples et filtrer une chimiothèque avant ou après un criblage expérimental ou virtuel.

implémentées. Si une molécule commence à allumer plusieurs feux rouges et se projette en outre assez mal dans un certain nombre de diagrammes ADME, alors le composé est rejeté.

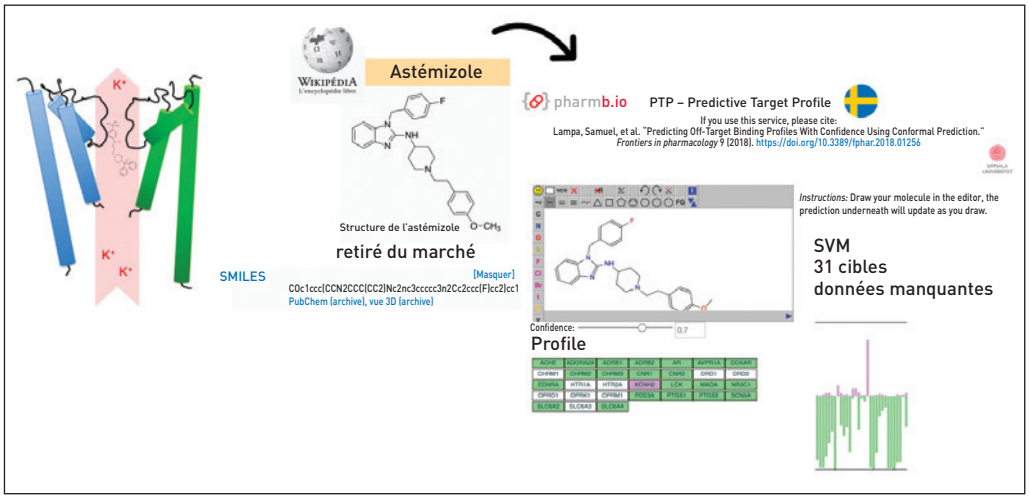
Ce premier niveau de filtrage est utilisé en amont du développement du candidat-médicament. Ensuite, pendant les phases d'optimisation des molécules (par exemple optimisation de l'affinité pour la cible, amélioration des paramètres ADME…), il est nécessaire d'affiner de plus en plus la recherche de toxicité potentielle, sachant qu'il y a des liens étroits et très complexes entre paramètres ADME et toxicité.

Après des années d'optimisation, une molécule peut par exemple atteindre le stade de la pharmacologie de sécurité préclinique. À ce stade, il est possible de coupler certaines approches expérimentales avec l'utilisation de modèles statistiques prédictifs beaucoup plus spécifiques que ceux utilisés dans un outil comme FAF-Drugs. L'objectif de cette étape est notamment d'essayer au maximum d'anticiper des interactions possibles entre une petite molécule et certaines cibles connues pour être dangereuses pour la santé si elles sont touchées par un composé. L'idée sous-jacente est de protéger les volontaires sains qui rentrent en phase 1, et ensuite les patients qui vont participer aux phases cliniques, et au final d'essayer de minimiser les échecs au cours du développement, ou même après la mise sur le marché.

Il est donc critique de détecter le plus tôt possible si une petite molécule chimique « touche » peut avoir des effets indésirables sur les systèmes physiologiques, et notamment sur les organes vitaux comme le cerveau, le cœur, le poumon. Il existe des listes de protéines cibles qui ont été publiées par l'industrie pharmaceutique et dans les laboratoires de recherche académiques qu'il faut éviter. Un premier jeu fait état d'une quarantaine de cibles à ne pas toucher ou à toucher de manière contrôlée. Dans la pratique, il y en a environ 150, ou 200 si on veut être plus exhaustif. Au niveau *in silico*, un objectif peut être de développer des modèles statistiques mathématiques prédictifs pour chacune de ces cibles et répondre à la question : est-ce que cette molécule qui inhibe une cible thérapeutique va aussi interagir avec une cible dangereuse pour la santé ?

Prenons l'exemple du canal potassique, hERG, qui est très important puisque son blocage par une petite molécule peut entraîner des arrêts cardiaques. Un certain nombre de médicaments ont été retirées du marché parce qu'ils bloquent ce canal hERG, notamment certains antihistaminiques, mais aussi d'autres molécules.

L'astémizole, retiré du marché (dont on voit la formule en 2D sur la ***Figure 15***), est un exemple type. Cette molécule peut aussi être représentée par un code beaucoup plus adapté à une manipulation informatique qu'une image, le code SMILES. Ce code peut être copié et collé dans un outil suédois, PTP, qui va évaluer (un modèle statistique pour



chaque cible a été construit à partir de molécules annotées extraites notamment de ChEMBL) si ce composé peut interagir avec 31 cibles dangereuses pour la santé. À ce jour, il n'y a pas encore assez de données libres pour construite un modèle mathématique pour toutes les cibles à éviter, mais c'est une étape.

6 Le repositionnement de molécules au service de l'innovation

Pour certains projets, comme dans le cas d'urgence sanitaire ou des maladies rares, il peut être pertinent d'utiliser des médicaments qui existent déjà pour essayer de traiter ou soulager les patients. Le repositionnement ou la réutilisation des médicaments signifie tester des molécules déjà connues pour une maladie différente de celle pour laquelle elles ont été développées. Cela peut se faire avec des approches expérimentales de criblage, mais les méthodes *in silico* peuvent être beaucoup plus rapides et/ou complémentaires. En comparant les étapes de découvertes classiques d'un médicament et les étapes de repositionnement, on note une réduction de temps de développement considérable (***Figure 16***). En effet, dans le cas du repositionnement, plusieurs étapes ne sont plus nécessaires et un certain nombre de connaissances ADME-Tox sont déjà établies. Plusieurs composés

Figure 15

Le service PTP en Suède (http://ptp.service.pharmb.io/) permet de tenter de prédire l'interaction entre une molécule chimique et 31 cibles à ne pas toucher pour un développement thérapeutique comme le canal potassique hERG.

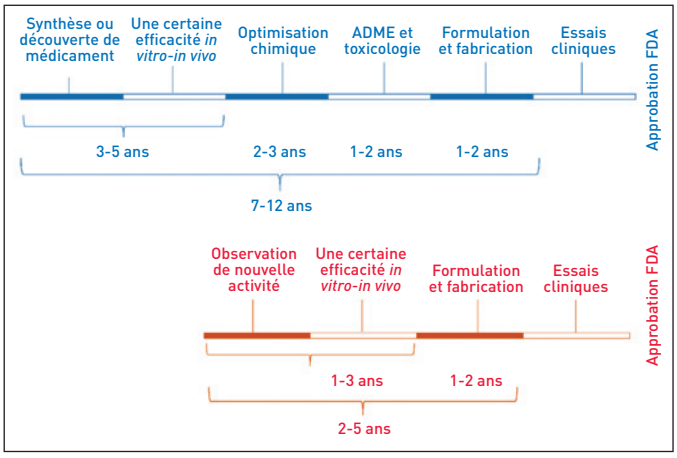


Figure 16

Les approches informatiques peuvent guider les projets de repositionnement de médicaments.

ont déjà été repositionnés sur d'autres pathologies ces dernières années comme par exemple le thalidomide.

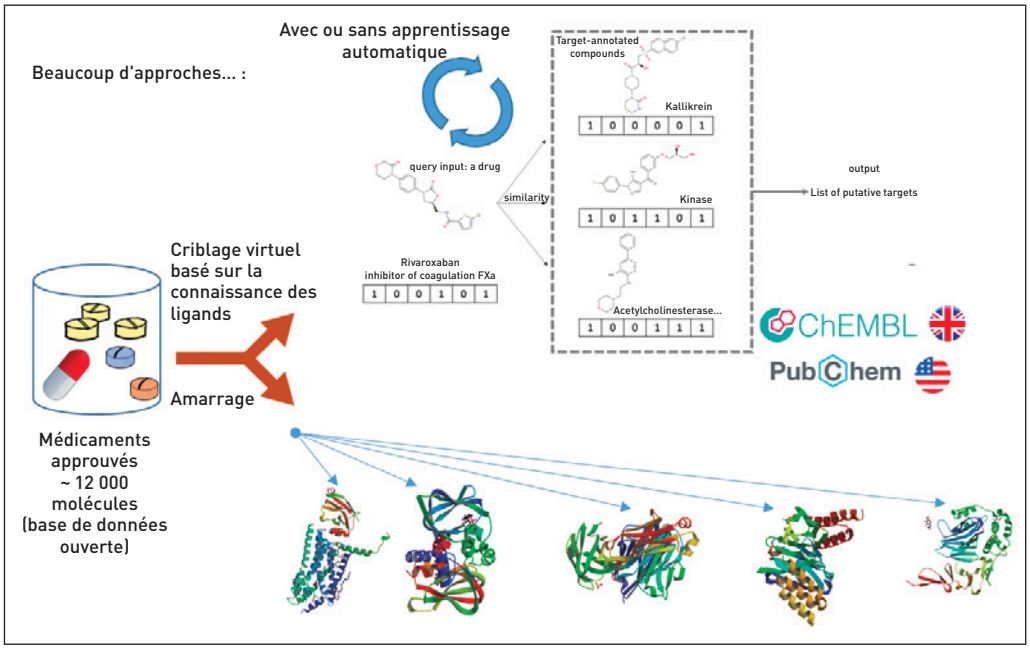
Plusieurs approches *in silico* peuvent être utilisées pour le repositionnement comme par exemple les méthodes basées sur les signatures transcriptomiques[8], sur les connaissances des ligands et sur la connaissance tridimensionnelle des cibles.

On dispose dans les bases de données ouvertes d'environ 12 000 molécules thérapeutiques qui sont sur le marché dans tous les pays ou seulement dans certains pays, et donc de composés chimiques qui sont déjà connus et administrés chez l'homme. Pour illustrer cette notion de repositionnement *in silico*, nous allons nous limiter à deux grands types de méthodes : les approches basées sur la connaissance des ligands et les approches basées sur la connaissance de la structure 3D des cibles (***Figure 17***).

Dans certains cas, nous savons que des cibles sont importantes dans une pathologie. Pour identifier les cibles qui pourraient être touchées par des médicaments connus, nous pouvons effectuer un criblage virtuel basé sur la connaissance des ligands. Sur l'exemple présenté sur la ***Figure 17***, la molécule à tester (« *query* ») est constituée de plusieurs groupes chimiques, elle est transformée en un vecteur de 1 et de 0 (notion d'empreintes moléculaires). Chaque fois qu'un fragment chimique (par exemple un cycle aromatique) du composé étudié se trouve dans une petite base de données sélectionnée par l'utilisateur, on donne la valeur 1, et quand le fragment chimique est absent, on lui met la valeur 0. Chaque médicament à tester est ainsi transformé en une suite de 0 et 1, ce qui permet de traiter rapidement les informations en informatique. Nous procédons de la même manière sur des millions de molécules annotées (nous connaissons au niveau expérimental certaines cibles touchées par ces composés) présentes dans des bases de données comme ChEMBL ou PubChem. Il est ensuite possible de calculer la similarité entre un médicament étudié et une petite molécule annotée. Si cette similarité est importante, nous pouvons suggérer que le médicament étudié se fixe aussi sur la cible de la petite molécule de ChEMBL ou PubChem.

Prenons l'exemple d'une molécule anticoagulante bloquant le facteur 10 de la coagulation. Il est possible de traduire sa formule chimique en une série de 0 et 1. Ce médicament peut être comparé, après un calcul de similarité, à toutes les molécules

8. Signature transcriptomique : le transcriptome est l'ensemble des ARN issus de la transcription du génome. L'analyse transcriptomique peut caractériser le transcriptome d'un tissu particulier, d'un type cellulaire, ou de comparer les transcriptomes entre différentes conditions cliniques. Ce type d'étude peut donner des pistes sur des cibles importantes pour une pathologie et sur des molécules qui pourraient rééquilibrer le système biologique perturbé.



qui se trouvent dans une base de données, donc actuellement à environ deux millions de composés. Le calcul suggère que la molécule anticoagulante peut aussi se fixer sur d'autres cibles, comme certaines kallikreines. Si ces kallikreines sont importantes pour la pathologie que l'on étudie, il est alors possible de tester expérimentalement cette hypothèse pour valider ou non la prédiction *in silico*.

L'autre approche, présentée sur la ***Figure 17***, est géométrique et énergétique, basée sur la structure de la cible et sur l'amarrage de la petite molécule. On connaît la structure 3D expérimentale de milliers de cibles (environ 150 000 en 2020), mais on a aussi environ 35 millions de modèles structuraux théoriques construits par homologie déposés dans des bases de données. On dispose ainsi d'une certaine couverture du protéome[9] humain, couverture qui devrait être pratiquement complète dans les dix années à venir.

Dans cette approche de repositionnement basée sur la connaissance de la structure 3D des cibles, un calcul d'amarrage des petites molécules-médicaments sur chacune de ces cibles est réalisé afin d'obtenir un score prédictif d'affinité entre toutes les petites molécules étudiées et toutes les cibles. Cela permet de tester rapidement un grand nombre de composés sur de multiples cibles. Même si les calculs d'affinités sont encore peu précis, on réalise bien le gain de temps et la réduction des coûts que peut constituer ce type d'approche. En effet, on n'a plus à tester tous les composés

Figure 17

Deux exemples d'approches in silico *pour repositionner des molécules chimiques. Des approches sont basées sur la structure chimique des ligands et sur la connaissance de ligands annotés présents dans des bases de données, d'autres approches sont basées sur des techniques d'amarrage moléculaire.*

9. Protéome : ensemble des protéines exprimées au sein d'un système biologique défini.

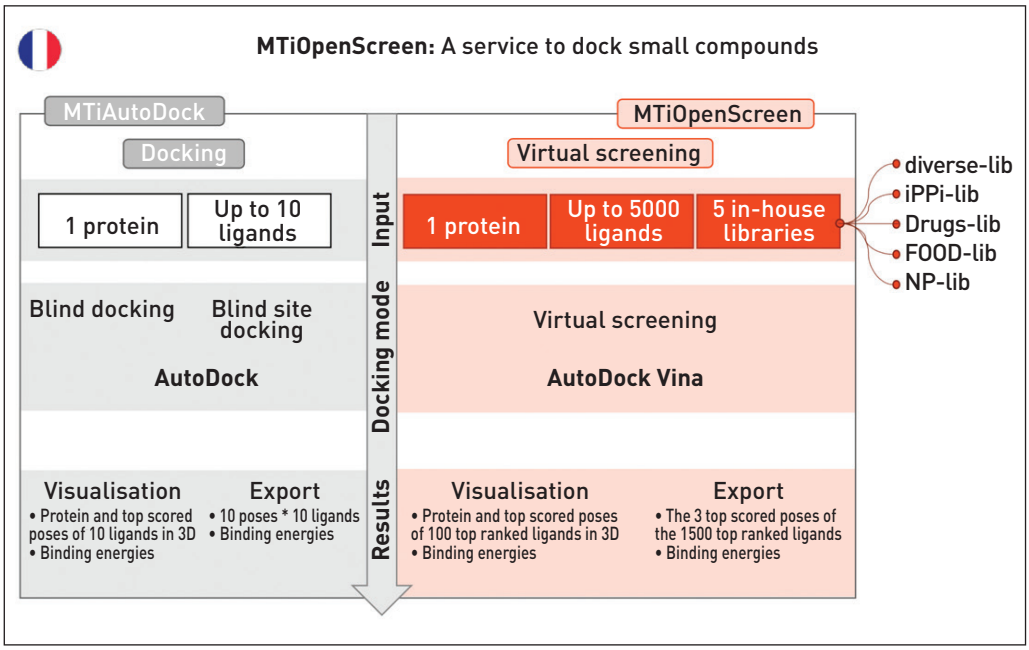


Figure 18

MTiOpenScreen (http://drugmod.rpbs.univ-paris-diderot.fr/index.ph) est un serveur pour le criblage virtuel basé sur la structure 3D des cibles.

expérimentalement, mais seulement ceux identifiés par les approches *in silico*.

Nous avons développé un serveur, MTiOpenScreen, dédié à l'amarrage des petites molécules et au criblage virtuel. Plusieurs banques de molécules sont déjà préparées pour les utilisateurs et notamment une banque de médicaments : Drugs-lib (***Figure 18***).

Pour développer ce serveur, il faut dans un premier temps préparer les banques de petites molécules. Pour cela, il est nécessaire d'extraire les données publiées dans différents pays, les agréger, éliminer les doublons et certains toxicophores, et effectuer toute une série de calculs pour aboutir à la base de données finale. Il faut aussi prédéfinir un certain nombre de règles pour ne sélectionner que les molécules qui sont adaptées à l'amarrage (***Figure 19***).

Ensuite, il faut un algorithme d'amarrage. Dans notre cas, nous n'avons pas redéveloppé une méthode mais implémenté deux outils libres publiés aux États-Unis, AutoDock et AutoDock Vina. La prochaine étape de développement du serveur est d'automatiser les processus pour faciliter le travail des utilisateurs, créer une interface intuitive et tester les outils afin de valider sur des exemples connus que les calculs sont corrects. Enfin, nous pouvons tester le serveur sur des nouvelles cibles et valider expérimentalement la pertinence de l'approche. Pour ce faire, nous avons par exemple étudié une protéine impliquée dans l'angiogenèse[10] et le cancer avec nos collaborateurs de Bordeaux (***Figure 20***). Cette protéine est cristallisée et on

10. Angiogenèse : processus de formation de nouveaux vaisseaux sanguins.

a pu, avec ce type d'approche et en moins d'une heure de calculs, identifier plusieurs antifongiques qui semblent bloquer l'activité catalytique de cette protéine. Ces composés antifongiques ont été validés expérimentalement et des brevets ont été déposés.

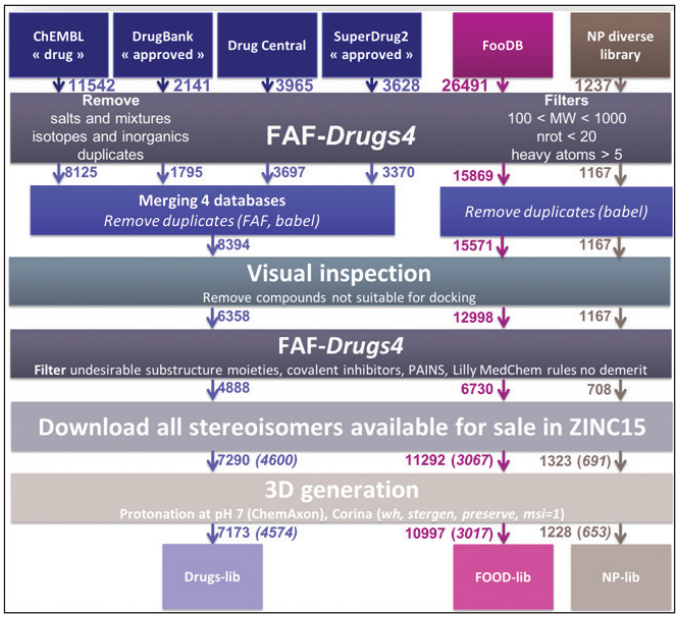


Figure 19

Un travail de nettoyage, de sélection et de fusion des bases de données est nécessaire afin de les rendre prêtes à l'emploi. Les critères ont été définis par les concepteurs du serveur.

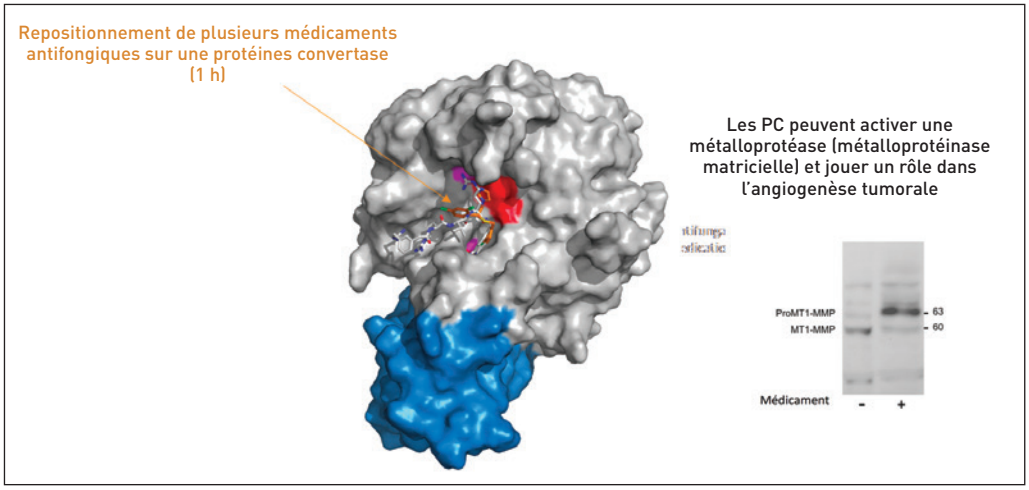


Figure 20

Utilisation du serveur MTiOpenScreen : il a été possible d'identifier des médicaments antifongiques qui ne sont pas connus pour se fixer sur les protéines convertases par notre approche informatique. Ces médicaments semblent bloquer l'activité de cette protéine impliquée dans l'angiogenèse et certains cancers. C'est une première étape dans notre processus de repositionnement.

Les perspectives de la pharmaceutique *in silico*

La conception de molécules candidats-médicaments en ligne est maintenant possible : les bases de données et les algorithmes peuvent être utilisés pour faire la conception de ces molécules (***Figure 21***).

Il reste encore des points complexes à résoudre :

– les données manquantes : par exemple dans le cadre du profilage pharmacologique, toute une série d'informations manquent sur un certain nombre de cibles ; il en est de même pour plusieurs types de prédictions de toxicité ;

- le manque d'uniformité de l'étiquetage des données : un travail de nettoyage important est à faire ;

- les options dans les services en ligne et l'identification des erreurs dans les prédictions : les algorithmes implémentés en ligne ne peuvent pas offrir toutes les options possibles aux utilisateurs car il est nécessaire de simplifier l'interface afin de permettre à des scientifiques qui ne travaillent pas dans le domaine *in silico* d'utiliser les services sans avoir besoin de coder ou sans se perdre dans des dizaines d'options complexes. Pour identifier les erreurs, il est possible dans certains cas de fournir un score de pertinence des calculs, mais le meilleur moyen reste encore d'étudier l'algorithme et de bien comprendre la force et la faiblesse des méthodes qui sont utilisées en arrière-plan.

Néanmoins le côté positif est qu'il est maintenant possible de concevoir des petites molécules-médicaments ou d'essayer d'optimiser des petites molécules « touches » avec ces approches en ligne. Ces observations sur les petites molécules chimiques sont aussi pertinentes pour les peptides et pour toute une

série de produits biologiques, protéines thérapeutiques et anticorps monoclonaux, pour lesquels l'utilisation de l'intelligence artificielle couplée à la bioinformatique structurale est en train d'émerger à très grande vitesse. Avec ces approches *in silico*, il est possible de tester de nombreuses hypothèses dans un délai très court, ce qui permet de réduire un certain nombre de travaux expérimentaux et de se focaliser sur ce qui est le plus important. De plus, ce type de recherche génère de nouvelles idées et, dans un certain nombre de cas, permet même de réduire de manière significative l'expérimentation animale.

Comme nous l'avons évoqué précédemment, ces ressources *in silico* sont non seulement importantes pour la recherche mais aussi pour l'enseignement. En effet, elles aident les étudiants à mieux comprendre certains sujets de recherche et certaines notions complexes. De plus, les services en ligne représentent une véritable vitrine de compétences, d'ailleurs on voit qu'un pays comme la Chine, qui était peu présent sur les outils en ligne dans le domaine du médicament il y a 5-6 ans, y est maintenant extrêmement actif avec pratiquement une publication scientifique dans le domaine toutes les semaines.

Ces approches *in silico* sont incontournables et pourtant restent insuffisamment utilisées en France. De plus, ces approches ne sont pas encore implémentées en France en cas d'urgence sanitaire alors qu'elles sont rapides et peuvent donner un éclairage nouveau sur de multiples questions scientifiques. Il va donc falloir combler au plus vite cette lacune. Le manque d'experts en bioinformatique structurale, chemoinformatique et autres disciplines récentes dans les cercles décisionnels et l'entre-soi expliquent en partie cette situation.