Similarité moléculaire et criblage virtuel

Méthodes de recherche *in silico* d'analogues actifs pour la découverte de composés thérapeutiques

Dragos Horvath* *Ph.D, head of Molecular Modeling Department,* **Catherine Jeandenans**** *Ph.D, molecular modeler*

Summary : ***Molecular similarity and virtual screening. In silico methods to retrieve active analogs in the context of discovering therapeutic compounds***

The « similarity paradigm », currently used by the medicinal chemist to design analogs of a known active compound, inferring a property similarity on the basis of structural relatedness, is actually a statistical law : the density of actives in a set of molecules that are similar to an active lead will exceed the density of actives in a set of randomly chosen molecules - if the current definition of « molecular similarity » is a valid one. In the context of combinatorial chemistry and high throughput screening, *in silico* similarity-based queries rely on molecular descriptors and similarity metrics to select analogs of active compounds out of huge electronic molecular databases.

This paper is a brief review of the high throughput similarity search methodology, including a concrete comparative example of the ability of different descriptors and similarity metrics to retrieve active analogs from a library.

Mots clés : ***Similarité, descripteurs, métriques, échantillonnage conformationnel, librairies combinatoires.***
Key-words : *Similarity, 2D and 3D descriptors, conformational sampling, combinatorial libraries.*

Définir sans ambiguïté le concept de « similarité moléculaire » s'avère une tâche très délicate du fait de la coexistence de plusieurs niveaux d'interprétation d'une structure moléculaire, chacun permettant d'aborder le problème de la similarité sous un angle différent. En effet, on peut parler de la similarité « 2D (topologique) » faisant référence à la manière dont les atomes sont interconnectés dans la molécule, d'une similarité « 3D (stérique) » se rapportant à la ressemblance de forme, d'une similarité « pharmacophorique » se basant sur la comparaison des distributions dans l'espace des groupes fonctionnels pouvant interagir avec un récepteur macromoléculaire, enfin d'une similarité « quantique » des densités des nuages électroniques. Traditionnellement, le chimiste médicinal fait appel au concept heuristique de similarité moléculaire pour concevoir des analogues apparentés à un composé actif. En se fixant comme but l'optimisation de l'activité de ces analogues, il apprend, au fur et à mesure qu'il accumule de l'expérience, la « bonne » pondération de ces divers aspects définissant la similarité moléculaire. Néanmoins, son expertise sera biaisée envers des aspects intuitivement appréhendés par un humain comme la connectivité moléculaire.

La puissance de l'approche informatique de la similarité moléculaire réside en deux aspects :

• La puissance de calcul, permettant le criblage de vastes collections de structures chimiques.

• La possibilité de traiter des aspects clés de la structure moléculaire difficilement intelligibles par un chimiste, comme le motif pharmacophorique ou les « champs » moléculaires de différents types.

Pratiquement, l'exploitation des bases de données moléculaires par « criblage virtuel » exige une méthodologie mathématique et informatique formelle qui doit permettre :

• L'énumération des structures des composés contenus dans les bases de données ;

• La représentation de chaque structure par un descripteur moléculaire, soit un vecteur localisant chaque composé dans un point d'un « espace structural » ainsi défini.

• La définition d'une métrique de (dis)similarité entre deux composés, représentant la « distance » entre les vecteurs descripteurs de ces molécules dans l'espace structural.

Le contexte combinatoire, soit la gestion de millions de composés potentiels, entraîne certaines contraintes dans le choix des outils, en favorisant des approches empiriques et rapides, tout en conservant la faculté de mettre en évidence les aspects importants de la structure moléculaire.

Méthodologie du « criblage virtuel »

Modélisation des produits combinatoires

La gestion informatique de « l'explosion combinatoire » - la génération d'un grand nombre de composés égal au **produit** des nombres de syn-

*et** CEREP, 128, rue Danton, BP 50601, 92506 Rueil-Malmaison Cedex.
* Tél. : 01.55.94.84.00. Fax : 01.55.94.84.10
E-mail : d.horvath@cerep.fr
** Tél. : 01.55.94.84.27. Fax : 01.55.94.84.10.
E-mail : c.jeandenans@cerep.fr

thons impliqués dans la bibliothèque combinatoire - nécessite un algorithme de génération des structures chimiques à partir des motifs structuraux présents dans les réactifs de départ, qui simule *in silico* les ruptures et formations des nouvelles liaisons chimiques observées pendant la réaction de couplage.

Les approches de type « Analog Builder » [1] procèdent en déterminant une sous-structure commune à toute une librairie qui sera « décorée » des différents groupes fonctionnels apportés par chaque synthon. L'étape initiale consiste à extraire ces groupes des réactifs initiaux [2]. L'étape de construction des produits est une opération purement topologique qui consiste à mettre à jour une table de connectivité entre les atomes pour y inclure les nouvelles liaisons créées lors de l'accrochage de ces fragments à la sous-structure centrale. Les groupes interconnectés sont souvent des fragments bidimensionnels des dessins moléculaires et, de toute manière, leurs positions relatives après connexion ne sont pas optimisées afin d'éviter les mauvais contacts stériques. Si la connaissance des aspects stériques s'impose, il est impératif de soumettre les produits issus de la concaténation des synthons, à des algorithmes de conversion « 2D-3D » pour construire une géométrie correcte à partir de la table de connectivité et des valeurs standard des longueurs de liaison, angles de valence, etc. (module CatConf de Catalyst [2]).

Les méthodes orientées vers la définition chimique d'une réaction permettent de décrire n'importe quelle transformation chimique, même complexe, y compris les cyclisations [3]. L'environnement chimique commun à tous les réactifs doit alors être défini pour permettre au logiciel de détecter tous les groupes fonctionnels dans les réactifs initiaux. De plus, des règles de correspondance doivent être établies entre les groupes fonctionnels inclus dans les réactifs et dans les produits finals. L'énumération des produits est à nouveau une opération purement topologique.

Dans le cas particulier où la structure finale du composé combinatoire peut être formellement représentée comme résultant du couplage de deux sous-structures par la formation d'une **seule** nouvelle liaison ($A_I + B_j \rightarrow A_i—B_j$), une méthodologie développée récemment [4] permet l'accès direct à de multiples conformations raisonnables des produits A—B, construites quasi instantanément à partir des géométries prédéfinies des sous-structures A et B. Ces dernières sont issues d'une méthode spécifique d'échantillonnage conformationnel générant des géométries dans lesquelles les points d'ancrage des deux sous-structures A- et –B sont complètement dépourvus de gènes stériques afin que le couplage générant le produit A—B soit stériquement possible. Avec cet algorithme, l'utilisation des modèles multiconformationnels pour décrire des chimiothèques de dizaines de millions de composés combinatoires n'est plus prohibitive en terme d'effort de calcul. L'approche bénéficie donc de « l'avantage combinatoire » qui consiste à effectuer, autant que possible, toutes les étapes coûteuses en temps de calcul au niveau des $N_A + N_B$ synthons, afin de pouvoir rapidement extrapoler les propriétés désirées aux $N_A \times N_B$ produits de synthèse sur la base des résultats ainsi acquis.

Descripteurs moléculaires : 2D ou 3D ?

La discussion reste ouverte sur les performances comparées des descripteurs 2D relativement aux descripteurs 3D [5-6]. L'utilisation de descripteurs 3D peut en effet être source d'artefacts dus à la variance de ces descripteurs par rapport à l'ensemble des conformations d'un composé. Si on génère, par exemple, la géométrie d'un composé à partir de sa table de connectivité avec deux logiciels différents, on pourra obtenir deux conformations soit deux valeurs assez différentes pour un même descripteur 3D - jusqu'au point où ces deux conformères seront perçus comme des espèces complètement dissimilaires par une métrique de similarité 3D. La reproductibilité des descripteurs 3D augmente s'ils sont pris comme valeur moyenne par rapport à un ensemble de conformations représentatives du composé. Ceci exige de procéder à un échantillonnage conformationnel de chaque composé, d'où l'intérêt pour les approches du style de celle présentée dans la référence [4]. Actuellement, les logiciels dédiés au traitement d'un grand nombre de molécules se limitent plutôt au calcul de descripteurs physico-chimiques (logP, pKa) ou topologiques à partir de la structure 2D des composés [1, 7].

Dans le but d'accélérer encore plus le calcul des descripteurs moléculaires des chimiothèques combinatoires, l'exploitation de « l'avantage combinatoire » a été envisagée par plusieurs auteurs, notamment pour générer rapidement des descripteurs 2D pour les librairies combinatoires dans le contexte des représentations de Markush [8], ou pour définir des descripteurs pseudo-3D des produits comme une concaténation des vecteurs décrivant chaque synthon (sans néanmoins générer une vraie structure 3D du produit fini [9]). L'accès quasi instantané aux conformations des produits [4] ouvre la voie à l'exploitation des descripteurs 3D dans le contexte de chimiothèques de très grande taille [10]. La caractérisation complète de 80 millions de composés combinatoires en termes d'un descripteur pharmacophorique « FBPA - Fuzzy Bipolar Pharmacophore Autocorrelograms » - a pu se faire en un mois de temps CPU sur une station de travail Silicon Graphics.

Le descripteur FBPA [10] représente les densités de distribution des paires d'éléments pharmacophoriques (tels que « hydrophobe-cation », « hydrophobe-aromatique », « donneur-accepteur d'hydrogène »…) en terme des distances géométriques séparant ces éléments, moyennées sur l'ensemble des conformations. Ceci est un cas particulier de descripteur pharmacophorique, se basant sur la logique floue afin de minimiser l'impact des artefacts d'échantillonnage conformationnel. D'autres descripteurs pharmacophoriques « binaires » décrivent l'existence ou absence des motifs structuraux à 3, voire 4 éléments de pharmacophore [11]. Chaque motif, défini par les éléments qui le composent et les intervalles des distances entre ces éléments, est codifié par un bit dans une séquence binaire : un bit « allumé » (1) signalant la présence tandis qu'un bit « éteint » (0) signale l'absence du motif. Ce caractère binaire rend ces descripteurs assez sensibles aux artefacts géométriques, car des bits différents peuvent

être allumés par des conformères différents (un bit sera allumé dans le descripteur moléculaire final si le motif codé par celui-ci est au moins présent dans une conformation).

Métriques de similarité

En fonction de la nature des descripteurs moléculaires, il existe une grande variété de métriques qui sont des fonctions traduisant une « distance » entre les points occupés par deux molécules dans l'espace structurel défini par ces descripteurs. Généralement [12], ces métriques sont soit de type euclidien (mieux adaptés pour la comparaison des vecteurs de descripteurs 2D et 3D classiques), soit des coefficients de corrélation entre deux vecteurs (mieux adaptés pour les descripteurs binaires). Enfin, toujours dans le but de réduire l'impact des artefacts géométriques, une approche introduisant la logique floue est utilisée pour comparer les FBPA. Certaines approches peuvent inclure des paramètres ajustables dans le calcul de la métrique de similarité comme c'est le cas pour la métrique associée aux FBPA. Les métriques peuvent alors être calibrées afin d'optimiser des problèmes spécifiques dans une approche purement mathématique ou, au contraire, de manière à corréler l'intuition du chimiste qui examine les composés trouvés similaires par la métrique.

Validation

Une expérience test a été conçue pour déterminer :

• si l'utilisation de ces outils permettait de diminuer l'effort d'expérimentation à haut débit des chimiothèques dans le but d'y découvrir des composés actifs,

• la performance relative des différentes métriques de similarité dans la détection de composés analogues actifs.

Nous avons considéré deux familles d'inhibiteurs A et B de la protéine farnesyl transférase FT [13-14] (*figure 1*) pour traiter le problème suivant : supposons qu'un seul inhibiteur de la famille A soit connu et qu'une chimiothèque contienne plusieurs représentants de la classe B, les outils utilisés sont-ils capables, à partir de la

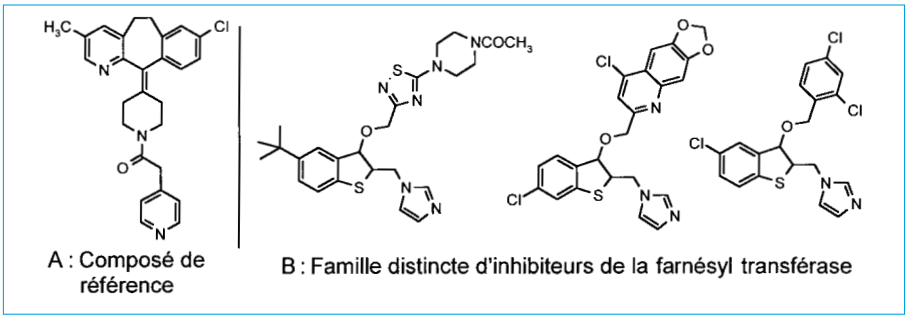

Figure 1 - Composé de référence A inhibiteur de la farnésyl transférase à droite utilisé pour retrouver les composés de type B (exemples des structures à droite) « cachées » dans une librairie combinatoire.

structure de A de classer les composés de types B comme plus proches voisins de A - prouvant ainsi que la stratégie du test sélectif de ces plus proches voisins est supérieure au test à l'aveugle de toute la chimiothèque. Cette chimiothèque est un sous-ensemble de 629 médicaments de la pharmacopée des États-Unis, ne contenant pas d'autres composés dont le rôle thérapeutique majeur soit inhibiteur de FT (bien qu'on ne puisse exclure que certaines de ces molécules aient une activité résiduelle sur cette cible). Parmi ces 629 composés, nous avons réparti 37 inhibiteurs de type B. Les familles d'inhibiteurs de type A et B, bien qu'agissant sur le même site de la protéine, sont structurellement très différentes, leur similarité supposée n'étant pas intuitivement reconnaissable par un chimiste médicinal.

Le degré de similarité de chaque molécule dans la chimiothèque par rapport à la référence A a été estimé par 4 méthodes différentes - dans les 3 derniers cas, les descripteurs ont été calculés à partir des ensembles d'au plus 20 conformations pour chaque molécule - :

1. descripteurs 2D, 3D (de forme) et topologiques « standard » calculés par le programme Cerius[2] [1], incluant des informations relatives aux hétéroatomes (états électrotopologiques [7]) ; métrique euclidienne,

2. descripteurs pharmacophoriques binaires à 3 éléments ; métrique de Dice [15],

3. descripteurs pharmacophoriques binaires à 4 éléments ; métrique de Dice,

4. descripteurs et métrique de Dice « floue » FBPA [10].

Les composés des chimiothèques ont été classés par ordre de similarité décroissante avec A et la fraction des composés B « cachés » (y %) trouvés parmi les composés classés comme les plus similaires vis-à-vis de A (x %) est représentée en fonction de x % sur la *figure 2*. Plus la pente de cette courbe est importante, moins on devra tester de

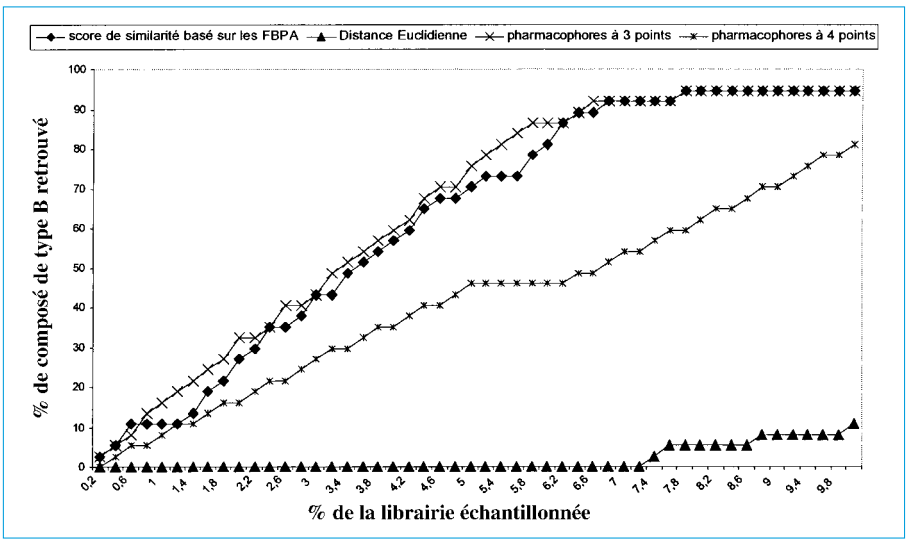

Figure 2 - Capacité de différentes métriques de similarité à retrouver les composés inhibiteurs de la farnésyl transférase de type B, disséminés parmi 629 composés de la pharmacopée US par sélection des plus proches voisins du composé de référence A (figure 1).

composés si on suit l'ordre indiqué par l'indice de similarité pour découvrir une fraction donnée des actifs « cachés » dans la chimiothèque.

En utilisant la métrique de Dice « floue » associée aux FBPA, nous aurions à tester seulement 10 % de la chimiothèque, dans l'ordre de similarité décroissante par rapport à A, pour trouver 90 % des composés inhibiteurs de type B. La stratégie utilisant le criblage virtuel basé sur l'approche FBPA est donc plus efficace que le test direct de l'ensemble de la chimiothèque.

En revanche, la métrique euclidienne associée aux descripteurs 2D classiques ne découvre pas la similarité existante entre la référence A et les composés de type B. Pire, les composés de type B apparaissent encore plus éloignés de A que les 629 composés de la librairie, en principe dépourvus d'activité sur FT. Bien que certains descripteurs (comme les clés électrotopologiques) contiennent des informations sur le type et le nombre des hétéroatomes, ils sont impuissants à traduire l'arrangement relatif des groupes fonctionnels des molécules. Tandis que l'analyse des motifs pharmacophoriques avec la technique FBPA arrive à mettre en évidence une similarité peu évidente entre A et B, l'approche basée uniquement sur topologie et forme moléculaire échoue.

Les empreintes pharmacophoriques basées sur la définition de pharmacophores à 3 points classent correctement les inhibiteurs de type B dans les composés similaires à A, avec une performance globale comparable à la méthode FBPA. Curieusement, les empreintes pharmacophoriques basées sur la définition de pharmacophores à 4 points sont **moins** performantes que les descripteurs 3D codant la position de paires ou triplets d'atomes. Si on examine les scores de similarité des composés identifiés comme plus proches voisins de A par les descripteurs à 4 points, on constate que le score de similarité est de 0,98 traduisant une similarité presque nulle (score de similarité maximum = 0, dissimilarité maximum = 1). On en déduit qu'aucun des composés de la librairie (y compris les inhibiteurs de type B) ne possède pratiquement de motifs pharmacophoriques à 4 points en commun avec la référence. Ils possèdent néanmoins des motifs **apparentés**, mais qui sont représentés par d'autres bits allumés dans les empreintes. On touche ici à la limitation de ces descripteurs 3D qui doivent contenir une certaine notion de « flou » pour compenser les problèmes dus à la comparaison de valeurs discrètes dans une métrique de similarité. Même si les descripteurs à 4 points contiennent potentiellement plus d'information sur l'arrangement 3D des éléments pharmacophoriques, et donc devraient permettre de classer plus précisément les composés en matière de similarité, les problèmes dus aux artefacts géométriques annulent complètement l'avantage de la définition précise des pharmacophores. L'augmentation de la taille de la famille des conformations utilisées pour la construction de cette empreinte peut partiellement atténuer ce problème.

Conclusion

L'apport des algorithmes de reconnaissance de la similarité moléculaire à la recherche d'analogues actifs peut, au-delà de la rapidité de l'outil informatique, se concrétiser par la découverte de composés qui présentent l'activité souhaitée, même si leur ressemblance avec le composé de référence n'est pas évidente. En effet, en utilisant des approches basées sur l'analyse des motifs pharmacophoriques 3D, le criblage virtuel met en évidence des similitudes significatives pour l'activité biologique, mais difficilement perceptibles par le chimiste. Bien que coûteuse en temps de calcul, cette analyse, nécessitant la construction de modèles conformationnels 3D, peut désormais être appliquée au traitement de grandes chimiothèques combinatoires. Les artefacts engendrés par l'utilisation des descripteurs 3D peuvent être minimisés par une définition appropriée de la métrique de similarité utilisant la logique floue par exemple.

Références

[1] *Cerius² v 4.0*, Molecular Simulations Inc., San Diego, California.

[2] *Catalyst v 3.5*, Molecular Simulations Inc., San Diego, California.

[3] *The Daylight Theory Manual*, http://www.daylight.com.

[4] Horvath D., Deprez B., Tartar A.T., High throughput molecular modeling using «" Fast 3D » descriptors, *Act. Chim. Ther.*, **1997**, *23*, p. 55-69.

[5] Matter H., Selecting optimally diverse compounds from structure databases : a validation study of two-dimensional and three-dimensional molecular descriptors, *J. Med. Chem.*, **1997**, *40*, p. 1219-1229.

[6] Brown R.D., Martin Y.C., The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding, *J. Chem. Inf. Comp. Sci*, **1996**, *36*, p. 572-584.

[7] Hall L.H., Kier L.B., Electrotopological state indices for atom types : a novel combination of electronic, topologic and valence state information, *J. Chem. Inf. Comp. Sci.*, **1995**, *35*, p. 1039-1045.

[8] Brown R.D., Downs G.M., Barnard J.M., Use of Markush structure-analysis techniques for rapid processing of large combinatorial libraries, Conference at the 218th National ACS Meeting, New Orleans, Louisiana, Aug. 22-26, **1999**.

[9] Sybyl 6.5, module Legion™, Tripos Inc., 1699 South Hanley Rd., St. Louis, Missouri, 63144, États-Unis.

[10] Horvath D., « High throughput conformational sampling and fuzzy similarity metrics : a novel approach to similarity searching and focused combinatorial library design and its role in the drug discovery laboratory », chapitre à paraître dans le livre de Ghose & Viswanadan édité par Marcel Dekker, N-Y.

[11] Mason J.S., Morize I., Menard P.R., Cheney D.L., Hulme C., Labaudiniere R.F., New 4-point pharmacophore method for molecular similarity and diversity applications : overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures, *J. Med. Chem.*, **1998**, *38*, p. 144-150.

[12] Menard P.R., Mason J.S., Morize I., Bauerschmidt S., Chemistry space metrics in diversity analysis, library design and compound selection, *J. Chem. Inf. Comp. Sci.*, **1998**, *38*, p. 1204-1213.

[13] Njoroge F.G., Vibulbhan B., Rane D.F., Bishop W.R., Petrin J., Patton R., Bryant M.S., Chen K-J., Nomeir A.A., Lin C-C., King I., Liu M., Chen J., Lee S., Yaremko B., Dell J., Lipari P., Malkowski M., Li Z., Catino J., Doll R.J., Girijavallabhan V., Ganguly A.K., Structure-activity relationship of 3-substituted N-(pyridinylacetyl)-4-(8-chloro-5,6-dihydro-11H-benzo[5,6]cyclohepta [1,2-b]pyridin-11-ylidene)-piperidine inhibitors of farnesyl-protein transferase : design and synthesis of in vivo active antitumor compounds, *J. Med. Chem.*, **1997**, *40*, p. 4290-4301.

[14] Kaminski J.J., Rane D.F., Snow M.E., Weber L., Rothofsky M.L., Anderson S.D., Lin S.L., Identification of novel farnesyl-protein transferase inhibitors using three-dimensional database searching methods, *J. Med. Chem.*, **1997**, *40*, p. 4103-4112.

[15] Willett P., Barnard J.M., Downs G.M., Chemical similarity searching, *J. Chem. Inf. Comp. Sci.*, **1998**, *38*, p. 983-996.