Le screening virtuel à haut débit (v-HTS)

Jacques R. Chrétien* *professeur*, **Marco Pintore*** *Ph. D*, **Frédéric Ros*** *docteur*

Summary : ***Virtual high throughput screening (v-HTS)***

Combinatorial chemistry and high-throughput screening (HTS) research programs are widely used in medicinal chemistry and agro-chemistry, in order to select new leads. But, due to the high cost associated to synthesizing and screening a very large number of compounds, there is an increasing need of efficient tools allowing to design and to classify large chemical libraries in order to get enhanced information content. This objective, defined as « data base mining » (DBM), can be achieved by analyzing molecular diversity in large databases with help of the most up-to-date methods based on Kohonen Self Organizing Maps, Fuzzy Logic techniques and Genetic Algorithm. In fact, the proposed methods allow to get a friendly representation of the compound distribution in the hyperspace derived from their molecular descriptors. These derived models of virtual high-throughput screening (v-HTS), linking the structures of the compounds with their biological properties, are suitable to predict activity values for new untested molecules.

Mots clés : ***Screening virtuel, data base mining, DBM, diversité moléculaire, chimie combinatoire***

La chimie combinatoire (CC) et le screening à haut débit ou « high throughput screening » (HTS) font l'objet de programmes de recherche ambitieux dans l'industrie pharmaceutique et l'agrochimie en vue de repérer de nouvelles têtes de séries ou « leads » [1-3]. Le but de ces travaux est d'accroître la diversité moléculaire des composés envisageables par la constitution de grandes collections de composés mettant simultanément à profit toutes les combinaisons de synthèses possibles. Ce sont ces nouvelles têtes de série, ainsi détectées, qui vont faire l'objet, dans un second temps, de pharmaco-modulations appropriées pour optimiser au mieux l'activité biologique recherchée. Ces optimisations résultent de modélisation moléculaire par docking, si toutefois la protéine, du récepteur ou de l'enzyme impliquée, a pu faire l'objet d'études 3D cristallographique préalable. Les modélisations moléculaires sont complétées par des études de « relations quantitatives structure activité » (QSAR), grâce aux analyses de champs de potentiels, avec par exemple la méthode « comparative molecular field analysis » (CoMFA) [4-6] qui reste la plus utilisée.

Mais, pour réduire le coût de la synthèse et des tests *in vitro* de grandes collections de composés, un besoin croissant d'une plus grande rationalité dans la conception et la gestion des plans d'expérience de chimie combinatoire se fait sentir. De plus, il y a, souvent encore, un grand chemin à parcourir entre une molécule active, dans un test *in vivo*, et une molécule candidate, en tant que principe actif d'un médicament, surtout lorsqu'elle résulte de synthèse peptidique, en particulier pour des raisons de stabilité ou de biodisponibilité. C'est là que les chimiothèques des industries concernées, ou les grandes bases de données de molécules accessibles sur le plan commercial ou technique, prennent tout leur sens, comme source de molécules pouvant rentrer dans le champ d'investigation ouvert par le nouveau « lead ». Le rapprochement CC/chimiothèques/grandes bases se situe, alors, en terme d'activité biologique visée, et non en terme de recherche sur une série structurale reposant sur une facilité de synthèse ou un savoir faire particulier en chimie organique classique. C'est ainsi que le concept de diversité moléculaire et sa maîtrise est devenu l'élément central et fédérateur des recherches actuelles de nouvelles molécules bio-actives.

La diversité moléculaire est impliquée à des titres divers dans les trois pôles de la créativité pour la recherche systématique de molécules bio-actives. Ces trois pôles, indiqués à la *figure 1*, sont les suivants :

- **1.** La chimie combinatoire/et le HTS ;
- **2.** Le data base mining/ou virtual-HTS (v-HTS) ;
- **3.** La modélisation moléculaire/ou docking et QSAR.

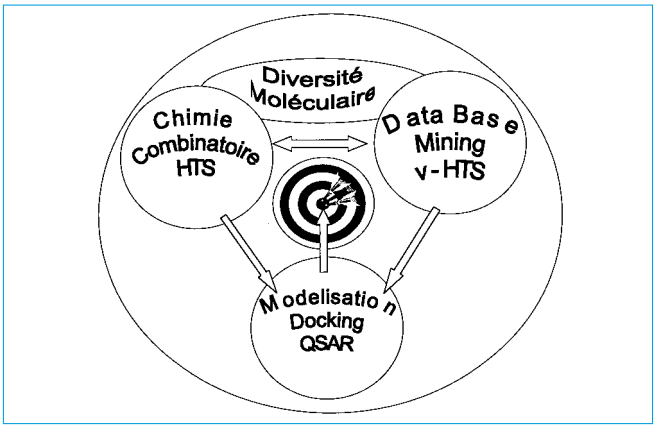

Figure 1 - Les trois pôles de la créativité pour la recherche de nouvelles molécules bio-actives.

* Laboratoire chimiométrie et bioInformatique, Faculté des sciences, Université d'Orléans, BP 6759, 45067 Orléans Cedex 2. Tél. : 02.38.41.70.76. Fax : 02.38.41.72.21.
E-Mail : Jacques.Chretien@univ-orleans.fr

Certes, très souvent, ces trois pôles sont dissociés et mis en œuvre indépendamment, pour des raisons de spécialités, technologiques ou conceptuelles. Et, pourtant, ils présentent de nombreux points de recoupements. En effet, une activité biologique donnée implique des interactions récepteur/ligand qui se ramènent elles-mêmes à des calculs d'interactions physico-chimiques. La maîtrise de ces interactions constitue le fondement de la modélisation moléculaire mais, aussi, des stratégies d'analyse de la diversité moléculaire qui sont l'essence du data base mining (DBM) de grandes bases qui est en plein développement *(figure 2)*.

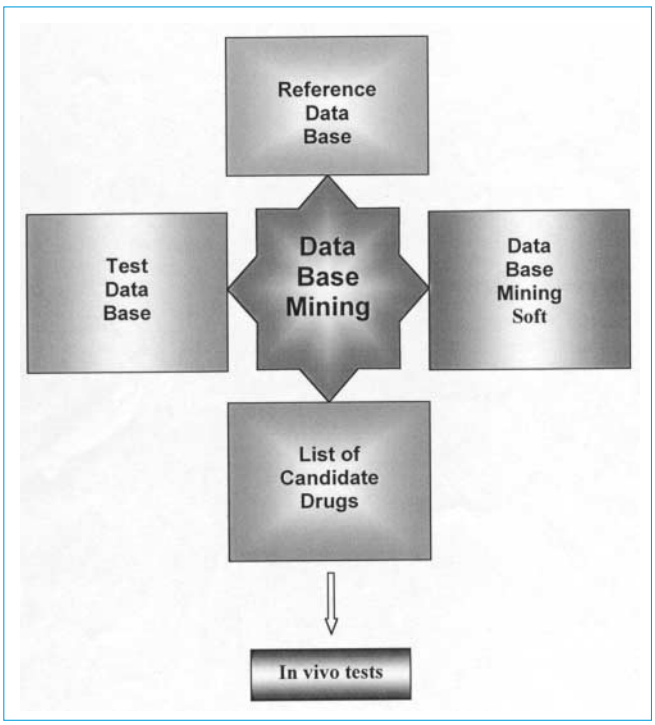

Figure 2 - Principes du data base mining (DBM).

L'explosion de la chimie combinatoire au cours des cinq dernières années a conduit à renforcer l'intérêt général pour la diversité moléculaire. C'est l'exploitation de la diversité moléculaire qui sert de trait d'union entre la chimie combinatoire et le DBM.

Si la chimie combinatoire s'est imposée dans les grands groupes industriels ou dans les sociétés spécialisées au milieu des années 90, la modélisation moléculaire a commencé à s'imposer il y a une quinzaine d'années. Par contre, les bases de données en chimie, documentaires ou factuelles, ont commencé à se développer il y a deux ou trois décennies. Elles correspondaient surtout à des besoins de stockage et de gestion de l'information. Cependant, certaines d'entre elles ont pu servir de bases de connaissances à des systèmes experts, par exemple dans le cas des psychotropes [7] et servir de précurseurs au data mining, en vue de pouvoir estimer, automatiquement, si une molécule nouvelle est susceptible de présenter l'activité neuroleptique ou non. Après l'engouement porté à la modélisation moléculaire, puis à la chimie combinatoire, c'est donc un juste retour des choses, qu'en terme d'épistémologie scientifique, le data base mining permette de restaurer l'intérêt porté aux « grandes bases maison », ou de mieux appréhender l'intérêt de grandes bases extérieures par la mise en œuvre de véritables stratégies d'exploitation innovantes. C'est là tout l'enjeu du DBM, avec la mise en œuvre de nouveaux concepts et de nouvelles procédures de traitement de l'information chimique, en vue d'aboutir au v-HTS.

Le data base mining (DBM)

Le screening virtuel : son objet.

La CC vise à générer, de façon la plus systématique possible, l'ensemble des combinaisons réalisables, cependant avec des limitations d'ordre cinétiques et thermodynamiques. La CC, dans sa phase de mise en œuvre, ne sait pas a priori si les molécules générées seront spécialement intéressantes pour atteindre la cible biologique visée. C'est le HTS qui permettra de conclure. En DBM, l'analyse de la diversité moléculaire vise à générer systématiquement de larges ensembles de descripteurs capables de traduire la complexité des interactions biologique possibles. Les moyens lourds en robotique de la CC sont remplacés, en DBM, par une nécessité de concepts nouveaux et d'algorithmes performants soigneusement validés. L'intérêt fondamental du screening virtuel est d'aboutir à une liste très réduite et la plus sélective possible de molécules présentant l'activité souhaitée, seules ces molécules seront soumises à des tests *in vitro* ultérieurs pour valider la prédiction. En terme d'effort, de coût et de gain de temps, l'enjeu est donc fondamental !

Générer la diversité moléculaire

Descripteurs moléculaires

La diversité moléculaire vise à une approche fragmentaire d'une information chimique globale. Cette information fragmentaire peut être établie sur des bases structurales comme dans le système DARC [8] ou les fingerprints. Ces approches fragmentaires structurales sont excellentes dans un but documentaire. Mais un nombre relativement élevés de fragments est généré, ce qui peut nuire aux besoins de généralité impliqué dans la maîtrise de la diversité moléculaire en vue de la prédiction d'activités biologiques, lorsque les composés sont en nombre très divers et présentent une diversité structurale très grande.

L'information fragmentaire peut être établie sur des bases comportementales, c'est l'objectif de l'utilisation des descripteurs moléculaires. Ils appartiennent en fait à trois classes : descripteurs topologiques, stériques et électroniques. Mais un descripteur moléculaire est par définition réducteur. Le fait de résumer le potentiel d'information biochimique qui se cache dans une structure 3D en un nombre est obligatoirement réducteur et ce, quelle que soit la nature de l'algorithme de calcul mis en jeu. Il engendre obligatoirement une perte d'information. On y remédie en testant de façon systématique un très grand nombre de descripteurs.

Une dégénérescence plus faible de l'information peut être obtenue en prenant mieux en compte la structure molécu-

laire 3D, mais il ne faut pas oublier que plus les descripteurs deviennent spécifiques, plus ils perdent de leur pouvoir généraliste dans un but de screening virtuel de grandes bases. Un bon compromis doit donc être trouvé dans le choix entre spécificité et généralité, c'est pourquoi les descripteurs de type 2D sont souvent retenus par rapport aux descripteurs 3D.

Métrique et organisation

Dans les bases de données de chimie, les composés sont représentés sous forme de graphes qui ne sont pas directement exploitables pour des comparaisons quantitatives. Pour pouvoir estimer a priori l'activité biologique de ces molécules, il faut être capable de les comparer les unes aux autres et de mettre en œuvre des raisonnements analogiques. Il faut passer de représentations sous forme de graphes, à une représentation numérisée, que l'on puisse comparer entre elles sous forme vectorielle, par exemple. Pour cela, on va passer d'une collection de graphes à une distribution dans un espace de descripteurs moléculaires, comme le montre la *figure 3*.

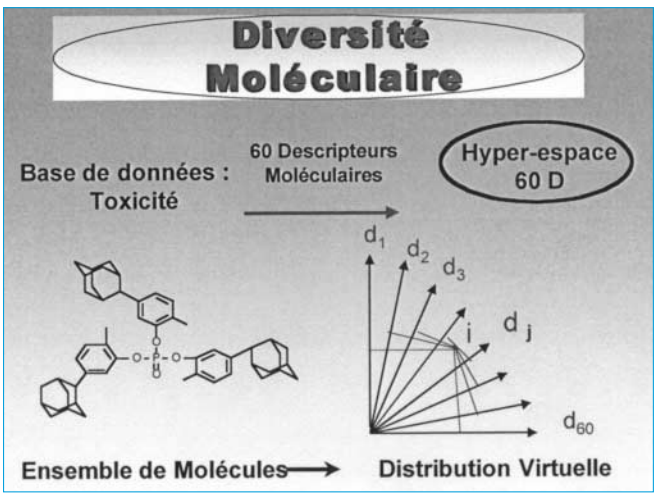

Figure 3 - L'objectif de la diversité moléculaire est de distribuer les composés dans l'hyperespace de leurs descripteurs.

Ainsi, chaque molécule correspond à une position dans l'hyperespace créé. Cet espace virtuel est muni d'une certaine organisation qui n'est accessible que par voie mathématique, en traitant les distances relatives entre les points représentatifs des milliers de molécules constitutives de cette base.

L'objectif de créer une métrique est d'induire dans la base de structure un ordre, une organisation qui reflète au mieux la diversité de la population moléculaire envisagée. Certes, cette organisation reste tributaire du pouvoir discriminant des descripteurs envisagés.

Visualiser ou modéliser la structure de grandes bases

Un hyperespace est un concept mathématique pour lequel on a du mal à imaginer une représentation intuitive pouvant conduire à une exploitation rationnelle. Une analyse en composante principale (ACP) nécessite facilement une dizaine de plans factoriels, même dans des bases à l'organisation peu complexe, pour prendre en compte 95 % de l'information. C'est ce qui fait l'intérêt de méthode de projection non linéaire par réseaux de neurones et notamment les « self organizing map » (SOM) de Kohonen [9-12]. La technique SOM conduit à une représentation de la distribution des molécules dans un hyperespace à une représentation dans un espace 2D. L'exemple de la *figure 4* illustre la possibilité d'offrir une méthodologie simple pour comparer visuellement l'originalité relative de deux bases de données [13]. La *figure 5* montre la possibilité de pouvoir cribler des bases de données en discriminant simultanément quatre mécanismes d'action de molécules anticancéreuses.

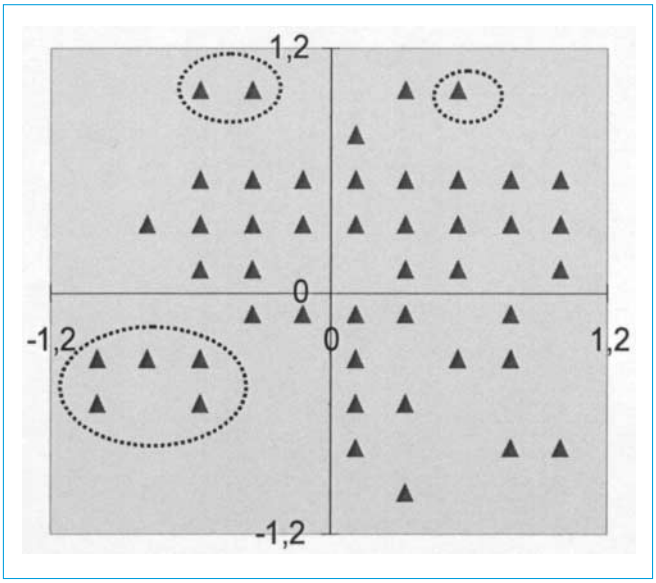

Figure 4 - Comparaison de deux bases de données sur les pesticides organophosphorés, l'une privée, l'autre correspondant aux pesticides commerciaux à l'aide d'une carte SOM de Kohonen. Le degré d'originalité plus grand de la base privée correspond aux composés qui se projettent dans les zones en pointillé de cette carte.

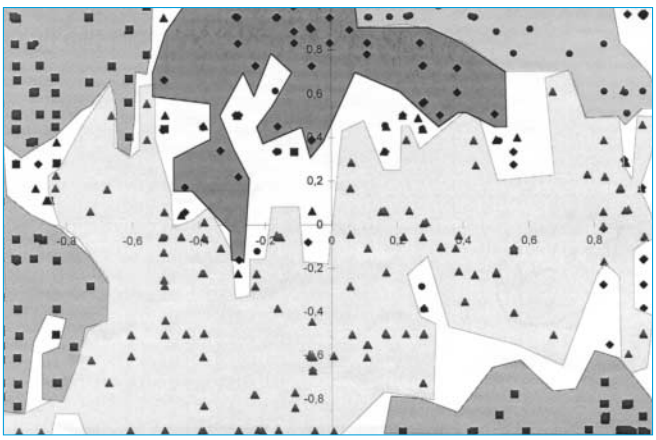

Figure 5 - Screening virtuel et prédiction d'activité pour des molécules anti-cancéreuses :
■ inhibiteurs de l'aromatase, ▲ inhibiteurs de la farnésyl transférase, ● inhibiteurs de la thymidylate synthase, ◆ inhibiteurs de la topoisomérase.

La technique SOM en tant que méthode de projection non linéaire par réseaux de neurones conduit à une certaine distorsion entre les distances relatives intermoléculaires et les distances propres dans l'hyperespace d'origine. C'est pour-

quoi des études directes dans l'hyperespace d'origine ont été entreprises par des méthodes déduites de la logique floue.

Les concepts de logique floue sont également très utiles pour traiter simultanément plusieurs propriétés. Ce cas a été abordé initialement pour les parfums [14-15] où une même molécule peut présenter simultanément plusieurs notes olfactives.

Conclusion

Les méthodes de travail en chimie organique dans la majorité des laboratoires de synthèse sont en train de subir une révolution sous l'influence de la chimie combinatoire. Elles se manifestent sous la forme de la miniaturisation des procédés de synthèse, des méthodes de séparation et d'analyse par le développement de synthèse sur support solide...

La chimie combinatoire comportent deux niveaux : (i) la constitution de « librairies de composés » et (ii) le screening in vitro (HTS). La chimie informatique y intervient surtout pour dépouiller les séquences de synthèse proprement dites et remonter à la combinaison efficace qui a donné naissance à une molécule active, donc à une possible tête de série.

Il existe toute une méthodologie applicable au DBM qui n'en est qu'à ses débuts. Mais, si nul chimiste organicien ne penserait se soustraire, aujourd'hui, de l'apport des méthodes séparatives (GC, HPLC) ou spectroscopiques (RMN ^{1}H, ^{13}C, SM...), nous pensons que le screening virtuel, indispensable à l'exploitation de grandes bases de données structurales, deviendra un outil complémentaire de la chimie combinatoire.

Au cours du temps, en chimie pharmaceutique et en agrochimie, l'intérêt s'est déplacé des bases de données vers la modélisation puis, plus récemment, vers la chimie combinatoire. Par un effet de balancier, le data base mining, avec la mise en œuvre de concepts nouveaux et puissants [16], redonne une nouvelle modernité à une compétence qui se banalisait : celle de la gestion des grandes bases de données. Gestion et exploitation poussée des grandes bases de sociétés, c'est l'objectif du data base mining qui est en train de conquérir une place de choix au cœur du processus d'innovation et de drug design. Le fond documentaire qui s'offre au DBM est immense : d'un côté, en terme rétrospectif, les *Chemical abstracts* ont déjà inventorié de l'ordre de 19 millions de molécules diverses, et une action concertée a permis de générer une base de molécules organiques plausibles du milliard de composés. Le DBM a donc encore de beaux jours en perspective au service du drug design !

Si le DBM est en plein développement et atteint sa phase de maturité avec des taux de molécules correctement prédites supérieurs à 80 %, il reste à mettre en place des méthodologies interactives totalement intégrées avec les robots de HTS.

La CC et le DBM procèdent d'une même démarche intellectuelle : générer une hypercomplexité dans la phase préliminaire, éventuellement avec un nombre presque infini de solutions, puis imaginer, dans la phase intermédiaire, des procédures de simplification, reposant soit sur des bases technologiques et robotiques coûteuses, ou des bases mathématiques et informatiques, elles mêmes très complexes, délicates à concevoir mais de coût infiniment moindre dans la mise en œuvre quotidienne. De toute façon, dans les deux cas, les objectifs se recoupent pour aboutir à des réponses, en nombre très limité, en l'occurrence à quelques molécules ou quelques dizaines de molécules.

Une intégration très poussée et des échanges bidirectionnels entre les trois pôles CC/DBM/modélisation sont une nécessité. Pour profiter au mieux de la synergie entre ces trois pôles qui ont leurs contraintes propres, une stratégie unitaire s'impose.

Références

[1] Dagani R., *C&EN*, **1999**, March 8, p. 51-60.
[2] Borman S., *C&EN*, **1999**, March 8, p. 33-48.
[3] Oldenburg K., *Annual Rep. in Med. Chem.*, **1998**, p. 301-311.
[4] Kubinyi H., Folkers G., Martin Y.C. *3D QSAR in Drug Design*, vol. 3, Part I « 3D QSAR Methodology. CoMFA and Related Approaches », p. 3-113.
[5] Cramer III R.D., Patterson J.D., Bunce J.D., *J. Amer. Chem. Soc.*, **1988**, *110*, p. 5959-5967.
[6] Golbraikh A., Bernard P., Chrétien J.R., *Eur. J. Med. Chem.*, **2000**, *35*, P. 1-14.
CoMFA / Tripos
[7] Chrétien J.R., Szymoniak J., Dubois J.E., *Eur. J. Med. Chem.*, **1985**, *20*, p. 315-325.
[8] Dubois J.E., in *Computer Representation and Manipulation of Chemical Information*, Wipke W.T., Heller S., Feldmann R., Hyde E., Editeurs, Wiley, New York, **1974**, p. 239.
[9] Kohonen T., *Self Organization and Associative Memory*, Springer-Verlag, Berlin, **1988**.
[10] van Osdol W.W.,, Myers T.G., Paull K.D., Kohn K.W., Weinstein J.N., *J. Natl. Cancer Inst.* **1994**, 86,
[11] Kireev D.B., Ros F., Bernard P., Chrétien J.R.,. Rozhkova N., in *Computer-Assisted Lead Finding and Optimisation*, van de Waterbeemd H., Testa B., Folkers G., Wiley-VCH, **1997**, p. 255-264.
[12] Kireev D.B., Chrétien J.R., Bernard P., Ros F., *SAR and QSAR in Envir. Res.* **1998**, *8*, 93-107.
[13] Bernard P., Golbraikh A., Kireev D., Chrétien J.R., Rozhkova N., *Analusis*, **1998**, p. 333-341.
[14] Ros F., Audouze K., Pintore M., Chrétien J.R., SAR and QSAR in Env. Res , **2000**, *11* (sous presse).
[15] K. Audouze, F. Ros, M. Pintore, J.R. Chrétien, *Eur. J. Anal. Chem*, 2000 (sous presse).
[16] Data Base Mining Soft. Logiciel développé au Laboratoire de Chimiométrie et de BioInformatique de l'Université d'Orléans.