HIERARCHICAL STRUCTURE-BASED VIRTUAL SCREENING FOR DRUG DESIGN

M. Miteva
French National Institute of Health and Medical Research (INSERM), University Paris Descartes, Paris, France
Correspondence to: Maria Miteva
E-mail: maria.miteva@univ-paris5.fr

ABSTRACT
Virtual ligand screening (VLS) has become an established approach for identification of new hit compounds for drug discovery programs. Recently hierarchical structure-based VLS (SB-VLS) protocols consisting of different levels of filtering have been developed. It was demonstrated that such hierarchical procedures based on docking-scoring methodology significantly improve the speed and the quality of SB-VLS. Here we overview current SB-VLS methods and hierarchical VLS protocols. Finally we present recent success stories obtained with hierarchical VLS methods.

Keywords: virtual screening, structure-based drug design, docking, scoring, structural bioinformatics

Introduction
Recently the number of macromolecular targets potentially involved in drug discovery programs has drastically increased because of the progress of the human genomics projects. Today technologies like combinatorial chemistry and high-throughput screening (HTS) authorize biological assays of a large number of molecules against the desired targets. However the high costs of drug design projects demonstrate the need of developing novel approaches allowing to explore faster larger chemical diversity. In this respect, virtual ligand screening (VLS), or in silico screening, is established as an attractive approach to permit in silico experiments of large number of molecules and to improve the “hit-rate” of drug discovery programs (36, 39).

Among the various VLS methods used for hit identification, two large families of methods are available: ligand-based screening (2, 40) and structure-based screening. VLS based on the 3D structure of macromolecular targets (structure-based SB-VLS) is widely applied to identify chemical entities that have a high probability to bind to a target molecule to obtain desired biological responses (38, 51).

For SB-VLS methods (Fig. 1), it is assumed that the 3D structure of the target is known either by X-ray crystallography or NMR experiments, or predicted by homology modeling. The protocol here is to dock all the ligands present in a database into the binding pocket of the selected target and evaluate the receptor-ligand fit (scoring). It is well known that molecular complexes are stabilized by interactions such as: van der Waals, hydrophobic, ionic, hydrogen bonds, solvation and entropic effects (5). By now, accurate computation of free energy of ligand binding can be performed by free energy perturbation (FEP) and thermodynamic integration (TI) (17) but these methods are too time consuming and as such, they are not appropriate for screening large compound collections. More approximate models have been proposed to evaluate relative binding affinities so called scoring functions (5). Scoring functions are used in order to: 1) evaluate different bound poses for a single ligand proposed by the docking algorithm in order to select the energetically preferred pose; 2) rank different docked ligands in order to discriminate the active compounds. Scoring functions can be classified into three categories: forcefield-based, knowledge-based and empirical.

Several recent reviews about computer aided drug design/ VLS methods have been published (1, 8, 13, 20, 46). In the present review we focus on the main VLS techniques applied for structure-based drug design and molecular docking–scoring methodology, and particularly the hierarchical VLS.

Main Docking approaches
The complexity of molecular docking implies several approximations, from rigid body docking, to (pseudo)-flexible ligand docking or flexible docking (where both receptor and ligand flexibility are considered). Algorithms treating the flexibility employ three types of searches, namely systematic, stochastic and deterministic ones. Some VLS packages use more than one of these approaches.

BIOTECHNOL. & BIOTECHNOL. EQ. 22/2008/1
Rigid body docking

Among the numerous docking programs several rigid body docking programs have been reported to dock rigidly previously generated conformers by matching interaction points from the receptor site with ligand atoms. One of the first molecular docking program for protein–small molecule interaction involving rigid body docking was the DOCK developed by Kuntz and co-workers (26). The program DOCK generates a negative image of the receptor - spheres that fill the binding pocket represent potential interaction sites. The DOCK algorithm attempts to superimpose the ligand atoms onto the centers of the spheres. Another rigid-body docking program FRED (29) applies a Gaussian shape fitting function to optimize the contact surface between the ligand and the protein. Despite obvious limitations, rigid body docking methods are interesting because they are much faster than the flexible docking algorithms. Software such as FRED can dock up to 10 compounds per second on a standard mono-processor Linux workstation (29). The speed and the relative accuracy of multi-conformation rigid body docking methods make them attractive (34) especially as an initial filter of a hierarchical structure-based VLS project in order to remove from the docking library compounds that have low surface complementarity with the receptor (32, 47).

Flexible ligand docking

**Systematic search** algorithms try to explore all the degrees of freedom. Fragmentation/reconstruction algorithms (incremental construction methods) generally divide a ligand into small rigid cores and flexible linking parts. For instance, the general approach to treat the ligand flexible in the program DOCK (28) is an “anchor-and-grow method”. Firstly, a set of overlapping spheres in contact with the surface of the receptor site is determined. These spheres fill the molecular surface of the binding site and represent a negative image of the target site. Then, the center of these spheres is matched with the ligand atoms via the use of a graph-matching algorithm. Several approaches can be actually used in DOCK for the scoring step: contact score, chemical score, energy score (Lennard-Jones van der Waals potentials and Coulombic electrostatics with distance-dependent dielectric constant), as well as scores taking into account the contribution of the solvation energy for molecular recognition: the Generalized Born/Surface Area (GB/SA) score (19) and Poisson-Boltzmann/Surface Area (PB/SA) scoring function (14). The programs FlexX (35) and Surflex (21) widely applied for VLS studies are also based on incremental construction search.

**Stochastic** (or random) search algorithms involve random changes to modify the position of the ligand as well as torsion angles in order to generate different conformations. The main stochastic search methods are **Monte Carlo (MC)** and **Genetic Algorithms (GA)**. MC can generate an ensemble of conformations statistically consistent at a given temperature. Random perturbations are applied in order to explore the conformational space of the molecular system. An energy function evaluates whether the energy of the newly generated conformation is either lower than the one from the previous step or, if higher, is within an energy range defined by the so-called Boltzmann factor (Metropolis criteria). In the package LigandFit (44) a MC method is employed for the conformational search of the ligand. Multiple structural changes may thus occur at the same time during this step. Once a new conformation for the ligand is generated, the fitting of the compound in the binding pocket is carried out. If the shape is similar then the ligand is docked into the binding site and its binding energy is evaluated via an energy function called DockScore (44) involving a soft 9-6 van der Waals term and an electrostatic term with a distance-dependent dielectric constant and, eventually, the internal energy of the ligand. Several scoring functions are available including, for instance, Ludi (4), LigScore (24) or PLP (45). Other MC methods for VLS are MCDOCK (27), QXP (30), or QUASI (42).

In the GA different ligand conformations and positions are generated, forming a population of solutions. This population is submitted to crossing over and random or biased mutations in order to form the next population. These algorithms are used in many docking programs such as GOLD (22), AutoDock (33), DARWIN (41). For example, the program GOLD adds fitting points to hydrogen-bonding groups on protein and ligand, and maps acceptor points in the ligand on donor points in the protein and vice versa. The GA optimizes flexible ligand dihedrals, ligand ring geometries, dihedrals of protein OH and NH groups, and the mappings of the fitting points.

**Flexible Receptor in VLS**

One of the main challenges for the VLS methodology today is to take into account conformational changes of the receptor upon ligand binding. Often the receptor flexibility is neglected during VLS experiments. However, in many receptor-ligand interactions, significant conformational changes can occur upon binding, for instance the induced-fit of protein kinases upon inhibitor binding (6) or at protein-protein interfaces (43). Various strategies are proposed in the literature to take into consideration the receptor flexibility. The most relevant, but apparently the most time consuming way is to apply **deterministic methods**. In the deterministic search, the initial state determines the change that can be made to generate the next state, which generally has to be energetically preferred as compared with the initial state. Deterministic methods for VLS are energy minimization and molecular dynamics (MD) methods. MD as compared to MC simulations, cannot cross easily high-energy barriers within reasonable simulation time at room temperature (5). Some simulation methods that could be useful for VLS have been developed to overcome more rapidly the energy barriers, for example using simulated annealing molecular dynamics (SDOCKER (49)). Some authors propose to carry out time-consuming MD simulations at the final steps of hierarchical VLS processes on a smaller pre-selected compound library in order to sample different conformations of the protein-ligand complex (25, 47).
One promising manner to account for the receptor flexibility in VLS projects is to perform the docking simulations on ensemble of several modeled (by MD or MC (16)) or experimental protein structures. The normal-mode-based methodology can also be able to incorporate receptor flexibility in ligand docking (10) and virtual screening (7). In addition, multiple experimental structures can be used to screen one target protein (3).

Cavasotto and Abagyan (6) proposed the ICM-flexible receptor docking algorithm (IFREDA) to account for protein flexibility during virtual screening. The ICM program is based on MC simulation that relies on global optimization of the energy function of the flexible ligand in the receptor field. Each step of the algorithm consists of a random change of two types, torsional or positional, followed by a local minimization. A third type of movement changes can also be applied, torsion moves of amino-acid side-chains at the interface. A partially flexible protein can also be considered with the program SLIDE (50) for docking flexible ligands in a binding pocket. A multilevel hashing procedure exhaustively detects matches between triplets of interaction points and triplets of ligand atoms. Once the anchor fragment is determined and that no collision is observed, the rest of the ligand atoms are flexibly added and optimized by rotating all single bonds. This includes some protein side chain flexibility.

Hierarchical VLS Methods

Various concepts have proposed to speed up the time-consuming procedures docking-scoring of SB-VLS. Recently multistep VLS protocols with funnel strategy for docking, consisting of different levels of filtering have been developed (11, 12, 32, 47). It was demonstrated that such hierarchical procedures for docking-scoring methodology significantly improve the speed and the quality of SB-VLS procedures (9, 23, 37). Several of these approaches start with pharmacophoric constrains or a geometrical matching of the target and the ligands. The following more time consuming filtering steps usually involve flexible ligand docking and eventually partial receptor flexibility and/or different level of precision in the estimation of binding energies with final free energy calculations involving estimation of van der Waals, Coulombic interactions, and changes in solvation and entropy due to the ligand binding (31). Recently hierarchical database screenings using a pharmacophore model, rigid-body docking, solvation docking, and MM(molecular mechanics)/PBSA have been shown to be useful in order to predict more precisely the binding energies (31, 47).

HierVLS (11) is a fast hierarchical docking approach that starts with a coarse grain conformational search over a large number of configurations filtered with a fast but crude energy function, followed by a succession of finer grain levels, using more and more accurate but more expensive descriptions of the ligand-protein-solvent interactions. The method GLIDE (15) uses hierarchical filters to explore plausible docking poses for a given ligand within the receptor site. The search begins with a rough positioning and scoring phase that significantly narrows the search space and reduces the number of poses to be further considered. Then, the 10 lowest-energy poses go through a MC procedure in which nearby torsional minima are examined, and orientation of peripheral groups of the ligand is then refined.

Success Stories with Hierarchical VLS

Recently several hierarchical VLS methods have been reported to be successfully applied for identification of new potent inhibitors. Some of these protocols start with an initial shape complementarity filtering by rigid-body docking (see Table 1). For instance Segers et al. (37) have investigated the structures of several domains interacting with the membrane and found

<table>
<thead>
<tr>
<th>Drug Target</th>
<th>Disease Function</th>
<th>Structure</th>
<th>VLS method</th>
<th>Activity of the best found hits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein kinase CK2 (9)</td>
<td>cancer</td>
<td>X-ray 1jwh.pdb</td>
<td>Shape filter by rigid docking: FRED, MOE-Dock/ GLIDE/Gold</td>
<td>Ki=20 nM</td>
</tr>
<tr>
<td>Factor V membrane Interaction (37)</td>
<td>coagulation system</td>
<td>X ray 1ezt.pdb 1czs.pdb 1czv.pdb</td>
<td>Shape filter by rigid docking: FRED, Surflex, LigandFit</td>
<td>IC50=3.5 μM</td>
</tr>
<tr>
<td>ubiquitin C-terminal hydrolase-L3 (18)</td>
<td>ubiquitin-proteasome system, programmed cell death</td>
<td>X ray 1xd3.pdb</td>
<td>DOCK GOLD</td>
<td>IC50=100 μM</td>
</tr>
<tr>
<td>CCR5 receptor Agonists (23)</td>
<td>favorable for protection against HIV-1 infection</td>
<td>Homology Model</td>
<td>Pharmacophore filtering, Surflex and Gold</td>
<td>IC50 = 17 μM</td>
</tr>
<tr>
<td>Dipeptidyl Peptidase IV (48)</td>
<td>type 2 diabetes</td>
<td>Xray 1n1m.pdb</td>
<td>Pharmacophore filtering, GLIDE</td>
<td>82% inhibition at 30 μM</td>
</tr>
</tbody>
</table>
that in several cases, at the protein-membrane interface, a druggable pocket was present, suggesting that it should possible to use SB-VLS methods to identify molecules able to disrupt protein-membrane interactions (37). The second discoidin (C2) domain of coagulation factor V (FV) binds transiently to the appropriate membrane surface and was selected as a proof of concept. A multistep SB-VLS protocol was carried out, starting with a collection of over 300,000 drug-like molecules. These molecules were rigidly docked using FRED into two different crystal forms of the FV C2 domain. The top 60,000 molecules after FRED docking forms were selected for flexible docking with Surflex. The top 2000 docked poses were re-docked and investigated with LigandFit to search for consensus poses with Surflex. The final lists of about 1000 selected molecules were tested in vitro in functional assays as well as with the Biacore system. Finally, seven drug-like hits were identified, indicating that therapeutic targets that bind transiently to the membrane surface can be investigated cost-effectively and that inhibitors of protein-membrane interactions can be designed.

Whereas the majority of hit compounds identified by SB-VLS approaches are in the micromolar range of activity, Cozza et al. have identified a nanomolar inhibitor of CK2 (9). CK2 is implicated in a wide variety of important cell functions, and evidences have been accumulating that its catalytic subunits may behave as oncogenes, consistent with the observation that they display an anti-apoptotic effect in prostate cancer cell lines. Cozza et al. (9) used a multi-step protocol proposed by Miteva et al. (32) and combined four docking tools and five scoring functions. They screened the ATP binding site of CK2 with their in-house molecular database of 2000 natural-occurring compounds. The SB-VLS protocol included a rigid-body docking step with FRED, then 50% of the remaining collection was submitted to flexible docking by MOE-Dock, Glide, Gold and the top 5% was selected for consensus scoring. One compound had a Ki of 20 nM, ellagic acid presents a unique binding motif among all known CK2 inhibitors, the peculiarity of this small molecule is its binding mode since it is able to simultaneously bind the hinge region and the phosphate-binding region of the ATP-binding cleft.

Another possibility for hierarchical VLS of a large database is application of initial pharmacophoric filters. For instance in (23) a library of 1.6 million compounds have been screened in silico against the CCR5 model by sequential filters. Identifying GPCR agonists by structure-based in silico screening is a challenging task, because GPCRs are expected to undergo significant conformational changes upon activation. After drug-like and 2D-pharmacophore filters the initial banks has been reduced to 44000 compounds subjected lately to 3D flexible docking Surflex and Gold yielded a hit list of 59 compounds. Out of these 59 molecules, 10 compounds were found to display micromolar affinities to the CCR5 receptor. Six hits were found to be agonists of the target receptor. Ward et al. have also been using hierarchical SB-VLS approach with pharmacophoric filters to search for inhibitors of Dipeptidyl Peptidase IV (DPP-IV) (48). A subset of 800,000 compounds was used a starting point and simple ADME/tox filtering reduced the collection to around 500,000 molecules. This multi-conformational database was then run against the two designed DPP-IV pharmacophores. Twenty thousand hits from each of the pharmacophores were selected based on the RMSD overlap of the compound with the pharmacophore. A single-conformer database of the 40000 selected molecules was then docked into the DPP-IV crystal structure using GLIDE. Clustering and visual inspection were used to select a final list of 4000 compounds for experimental screening. Finally 51 active compounds were identified.

Conclusions
Virtual screening methods based on the 3D structure of the receptor and particularly the hierarchical SB-VLS approaches offer real opportunities to improve the speed and the effectiveness for identification of new hits for drug discovery programs. Further progress of hierarchical SB-VLS methods will be required to involve steps for more precise evaluation of binding energies and better treatment of conformational changes of the receptor upon ligand binding, which is still challenging during the real VLS experiments.

Acknowledgements
We thank the French Institute of Health and Medical Research (Inserm) and the Egide organization.

REFERENCES


