# NO PICO

PAPER ID: PMC3823351

TITLE:
Docking, virtual high throughput screening and in silico fragment-based drug design

ABSTRACT:
The drug discovery process has been profoundly changed recently by the adoption of computational methods helping the design of new drug candidates more rapidly and at lower costs. In silico drug
design consists of a collection of tools helping to make rational decisions at the different steps of the drug discovery process, such as the identification of a biomolecular target of therapeutical interest, the selection or the design of new lead compounds and their modification to obtain better affinities, as well as pharmacokinetic and pharmacodynamic properties. Among the different tools available, a particular emphasis is placed in this review on molecular docking, virtual high-throughput screening and fragment-based ligand design.

Introduction:
Drug discovery is an interdisciplinary, complex, time consuming and expensive process. It is widely admitted that the pharmaceutical industry now spends far more on research and development but produces fewer new molecules than 20 years ago. The PriceWaterhouseCoopers Pharma report for 2005 stressed that the pharmaceutical industry needs to find means to improve the efficiency and effectiveness of drug discovery and development. It projected that in silico methods will become a dominant tool to address this issue, from drug discovery to marketing. Recently, advances in computational techniques and hardware have enabled in silico methods to speed up lead identification and optimization. Up till now, these techniques have contributed to the design of about 50 compounds that entered clinical trials, some of which are now FDA approved [1]. As of today, in silico drug design should not be seen as a 'voilà' technique able to suggest directly a small number of compounds with a high affinity and selectivity for the targeted macromolecule, along with favourable pharmacokinetic and pharmacodynamic properties, and using only the three dimensional (3D) structure of the target as a starting point. It rather consists of a systematic use of a wide range of different computational tools aiming, for instance, at improving the knowledge about the target-ligand interactions (molecular docking), increasing the yield of molecules screening by focusing the search on compounds more likely to bind the target (virtual high-throughput screening [vHTS]) or even suggesting new potential lead compounds (fragment- fragment-based ligand design [FBD]) [1]. Those methods are detailed below.

Protein and ligand flexibility ::: Docking:
During the physical binding, both the ligand and the protein adapt their conformations to each other. This phenomenon is called the induced fit. As a consequence, docking algorithms should handle the flexibility of both molecules. However, taking account of all these degrees of freedom (DOF) leads to a combinatorial explosion of the conformational space making the docking an even more challenging task. Therefore, almost all docking programs perform flexible ligand docking while the receptor is kept rigid. The main exceptions are GOLD, AutoDock, DOCK and EADock, which apply some flexibility to the protein during the docking through active site side chains rotations and more global minimizations, as well as FlexE and the IFREDA, which use a set of different pre-generated receptor conformations obtained experimentally or with in silico approaches.

Sampling algorithm ::: Docking:
Several approaches are used to sample the ligand-binding modes, and in some cases, to treat the flexibility of the protein. These sampling algorithms may be divided into three major categories: systematic search algorithms (FlexX or FlexE, DOCK, Glide [18], Hammerhead [19]), stochastic methods [AutoDock, GOLD, Quick Explore (QXP) [20], EADock] and simulation approaches.
The ideal systematic exploration of all DOF in a molecule to find its native binding mode is usually an impossible task due to the combinatorial explosion of the search space. Therefore, several methods that fall into the category of 'systematic search algorithms' use the technique of the incremental reconstruction of the ligand to compensate for this exponential dependence on the

molecular size. There are basically two ways to perform incremental reconstruction. In the first one (FlexX, FlexE), the molecule is divided into a single rigid fragment and several shells of flexible extensions. The rigid fragment, selected for its ability to make the highest number of interactions with the receptor, is docked first. The flexible moieties are then reconnected incrementally. After adding one flexible component, new interactions are searched for in compliance with the torsional database, and the scoring function is used to select the best partial solutions that are used for the next extension step. In the second variant of incremental reconstruction (Hammerhead and original version of DOCK), the molecule is decomposed into various fragments that are docked independently and subsequently fused into the active site using a hinge-bending algorithm. In addition to these reconstruction algorithms, other programs approximate a complete systematic search of the binding modes space of the ligand by narrowing the latter using several filters. For instance, Glide [18, 21, 22] performs an initial rough positioning and scoring phase to narrow the search space, followed by torsionally flexible energy optimization for a few hundred surviving candidate poses. The very best candidates are further refined via a Monte Carlo (MC) sampling of pose conformation to improve their accuracy.

In stochastic methods, the ligand is considered as a whole, and step-by-step changes are applied to a starting pose or a population of poses. Such methods subsequently score the new poses at each step trying to enhance the interactions with the protein, leading hopefully to the native binding mode. Evolutionary algorithms (EA) and MC simulations fall into this category.

EA mimic the process of the Darwinian evolution. The starting point is a collection of poses corresponding to plausible ligand-receptor complexes, also called the starting population or seeds. An objective function assigns a score to each binding mode, so that the less likely can be replaced by new ones to form a novel generation. These new poses are generated via computational procedures, called operators, that mimic biological mutations and crossovers. A mutation will introduce perturbations in the binding mode, like a rotation of one dihedral angle, while a crossover combines two poses. Operators are applied on the poses selected from the fittest elements of the population, with the hope that even fitter solutions will be generated. The algorithm ends after a given number of generations or energy evaluations, or if it has converged to a solution. The best-known programs in this category are GOLD and AutoDock, but several new promising EA-based algorithms are emerging, like EADock or MolDock [23]. These programs vary in the way they handle poses, in their operators and scoring functions. The reader is referred to relevant papers for a more detailed description of these methods.

MC-based methods start from a single randomly generated pose and apply subsequent random moves, like rotation of one dihedral angle and global translation or rotation of the whole ligand. After each modification, the new pose is scored, and the Metropolis criterion [24] is applied to choose whether the new pose is retained as a starting point for the next modification, or if the algorithm continues from the previous one. The algorithm ends similarly to EA-based approaches. As an example, the QXP [20] program belongs to this category.

Simulation methods group molecular dynamics and minimization methods. These approaches are often unable to cross high-energy barriers within feasible simulation time periods, and therefore might only accommodate ligands in local minima of the energy surface [5]. As a consequence, they are rarely employed as stand-alone search techniques. However, they can efficiently complement other search methods, by refining locally the poses that are suggested by one MC or EA-based step, like in AutoDock, DOCK or EADock.


Scoring functions ::: Docking:
The scoring functions typically implemented in protein-ligand docking can be divided into three major categories [5]: knowledge-based, empirical and force-field-based scoring functions. Knowledge-based scoring functions use inter-atomic interaction potentials obtained by a reverse-Boltzmann analysis of the occurrence of different atom–atom pair contacts in known experimental complex structures [25, 26]. Empirical scoring functions are based on the idea that binding free energies can be written as a weighted sum of uncorrelated terms, such as hydrogen bonds, non-polar and aromatic contacts or entropy penalties. The weighting factors of these terms are determined by regression analysis using protein-ligand complexes with known experimental binding free energy and 3D structure [13, 27, 28]. Although easy and fast, these methods suffer from a limited description of the physical aspects of the binding process and from a dependence on the experimental dataset used for their parameterization. On the contrary, the estimation of the binding free energy by force field-based methods use unfitted, universal and physically sound energy functions, such as van der Waals and electrostatic interaction energies, and intramolecular energies [10, 12]. Recently, implicit solvation models have been introduced into docking scores to

capture solvent effects upon association [2, 10, 29]. Docking programs generally approximate the exact force-field energy using a grid summation, in which the interaction energy between the protein and an atomic sample is calculated on different regularly spaced points. The binding energy of a ligand is then calculated by summing the contribution of the grid points occupied by the small molecule, taking account of the actual nature and charge of the ligand atoms. EADock is among the very few docking programs that make direct use of a universal and detailed force field such as CHARMM22 and an accurate solvation model such as Generalized Born using Molecular Volume (GB-MV2) [30, 31].

Performance ::: Docking:
The performance of docking programs is generally assessed through re-docking calculations. First, hundreds to few thousands of experimentally determined representative ligand–protein complexes are collected, like the Ligand–Protein Database [32], the Astex/Cambridge Crystallographic Data Centre (CCDC) [33] and Astex/Diverse [34] sets or the Mother of All Databases [35]. Ligands are then removed from their binding sites, and the ability of the programs to reproduce the native binding mode is assessed. Generally, a docking is considered successful if the root mean square deviation (RMSD) between the experimental and calculated binding modes is lower than 2 Å. Although it is the current standard, this definition is arguable since it has been shown that two binding modes within 2 Å RMSD can make very different interactions with the protein [36]. Several benchmarks of different docking algorithms are available [37–39], which show that the typical success rate for re-docking ranges from 70% to 80%, depending on the authors and the test sets. It is important to note that these figures overestimate the efficiency of these programs for typical drug design studies. Indeed, the re-docking process neglects the induced-fit issue, because the protein conformer that is used for the docking of a given ligand comes from the experimental structure of the complex and is thus adapted to fit that particular compound. This is not the case when the ligand is taken from a screening database or is designed by in silico methods. It has been recently confirmed that docking a ligand to a non-native protein conformer, i. e. performing what is called a cross-docking, is a more difficult task in which the success rate of docking programs is reduced by at least 20%[40]. However, progress might be expected from methods developed to handle the protein flexibility in a fast and efficient way. Several analyses have also shown that the performance of most docking software highly depends on the particular characteristics of the binding site and ligand, so that it is hardly possible to figure out a priori which method, or combination of search algorithm and scoring function, is the more suited for a particular study [37, 41–43].

Ligand-based vHTS ::: Virtual high throughput screening:
When the structure of the target is unknown, the measured activities for some known compounds can be used to construct a pharmacophore model. The latter summarizes the positioning of key features like hydrogen-bonding and hydrophobic groups to be matched by putative ligands. Such a model can be used as a template to select the most promising candidates from the library [47, 48]. This strategy can also be used as a filter before applying a structure-based vHTS, so that only 1–10% of the initial database has finally to be docked [46].

Structure-based vHTS ::: Virtual high throughput screening:
Structure-based vHTS is probably the most straightforward application of docking algorithms. It consists of using a molecular docking program to determine the binding mode on the protein target for an entire database of existing or virtual compounds [44, 46, 49]. The bound conformations are used to approximate the binding free energy or the related affinity of the compound. Then, the most promising compounds are retained for further experimental testing. The most widely used docking programs for vHTS are DOCK, FlexX, Glide, GOLD and AutoDock. The size of the libraries used in such an approach ranges from hundreds of thousands to a few million compounds, limiting the time available for each docking to a few minutes or less. The size of the database is a trade off between the number of molecules that can be treated in a reasonable amount of time, and the chemical space that is desirable to cover. Despite the steady improvement of computer hardware, the conformational sampling is, therefore, very limited and vHTS suffers from a lot of false negatives. Despite the vast amount of resources invested in HTS and vHTS, and several successful studies [50–55], the outcome in terms of new compounds reaching the clinics might be seen as rather disappointing [56, 57].

Theoretical advantages of FBD ::: In silico fragment-based drug design:

FBD has several theoretical advantages over vHTS. First, FBD samples a higher chemical diversity than HTS. Indeed, HTS chemical libraries typically contain 105–106 individual compounds. Although it is a huge effort to handle such an amount of molecules experimentally or even in silico, this only covers a tiny amount of the chemical space accessible to the small drug-like molecules. Several studies have estimated this number to be around 1060–10100[45, 61–64], far beyond what can be tested by vHTS. Even the largest possible effort that could be imagined nowadays, using the estimated 120 million compounds available worldwide [65], only scratches the surface of the chemical space. On the contrary, FBD allows sampling of a much larger amount of the chemical diversity using a much smaller number of starting molecules. As an illustration, a chemical space of 106 molecules can be obtained by connecting combinatorially three fragments belonging to a 100-fragment database. But, contrarily to HTS, it only requires one virtual or experimental assay per each of the 100 fragments themselves and the few molecules that can be constructed from the most promising ones. Also, it has been calculated that the number of stable and synthetically accessible molecular fragments is around $44 \times 106$[66]. This number is nearly of the same order of magnitude of what is tested with HTS, but covers a much vaster part of the chemical space.

Second, FBD leads to higher hit rates. This is illustrated by the fact that the probability of a bad ligand-protein interaction increases exponentially with the size and complexity of the molecule [67]. As a consequence, the probability that small and simple molecules bind to the protein, even with a low affinity, is much higher than for HTS-size compounds. This probability climbs up to 30% to 40% for simple fragments [67]. This supports the use of molecular fragments to anchor the drug design process rather than complex and large molecules.

Finally, FBD leads to molecules with a higher ligand efficiency. HTS chemical libraries are composed of complex molecules originally developed for other purposes than binding to the current target. As a consequence, even a HTS hit is expected to form sub-optimal binding interactions with the target. On the contrary, due to its size, a high proportion of the atoms in a fragment hit are directly involved in protein-binding interaction. Their optimization has thus a better probability to lead to more efficient and therefore smaller drugs (Fig. 2), with better chances of favourable pharmacokinetic properties [57].

Interestingly, the binding free energy of a molecule resulting from an optimal linking of two fragments is expected to be lower thus more favourable, than the sum of the free energies of binding of the two isolated fragments [68] (see Fig. 3). This results from the fact that the rigid body entropic loss upon binding of a molecule is large, whereas the entropic penalty associated with freezing the rotatable bonds is small in some circumstances. The rigid body entropic loss upon binding of one molecule is due to the freezing of 6 DOF: the three rigid translations and three rigid rotations of the small molecule. 12 DOF are frozen upon binding of the two separated fragments A and B. This leads to a higher entropic penalty than when freezing the 6 DOF of the A:B joined molecule. This favourable difference in rigid body entropy prevails the conformational entropic loss of the A:B molecule, which is due to the freezing of the rotatable bonds that do not exist in the A and B fragments.

Existing FBD methods ::: In silico fragment-based drug design:

The properties of 40 fragment hits identified experimentally against several targets indicated that they show, on average, properties consistent with a 'rule of three'[69], i. e. molecular weight < 300 g/mol, number of hydrogen-bond donors ≤ 3, number of hydrogen-bond acceptors ≤ 3, calculated LogP ≤ 3. In addition, it was found that the number of rotatable bonds and the polar surface area were usually lower or equal to 3 and 60 Å2, respectively. Fragments are usually obtained using a chemoinformatics approach by breaking down biologically active compounds into a limited number of fragments. Depending on the definition of molecular fragments that is used, the chemical space of drug-like molecules reduces to some hundreds [70, 71] to thousands of fragments [72]. Several approaches are available to automatically decompose molecules into rigid fragments [73, 74].

Several methods have been developed for in silico FBD (see Table 2), which differ in the building blocks used to construct the ligands (atoms or fragments), the target constraints applied (ligand- or receptor-based), the strategy used to sample the chemical space (depth first [59], breadth first [59], MC, EA), the structural sampling (mainly growing, linking and random structure mutations) and the scoring function used to rank the putative ligands. Among the most representative methods, one can find LUDI [75], Multicopy Simultaneous Search (MCSS) [76]/HOOK [77], PRO_LIGAND [78], Small Molecule Growth (SMOG) (DeWitte and Shakhnovich), LigBuilder [79],

LeapFrog (Tripos Inc., Tripos, St. Louis, MO, USA), CCLD [80] and Genetic Algorithm–based de Novo Design of Inhibitors (GANDI) [81].

In ligand-based FBD, new molecules are designed based on existing ligands. From the latter, different constraints and scoring functions can be derived, like pharmacophore models, molecular similarity or Quantitative Structure Activity Relationship (QSAR) scoring functions. On the contrary, receptor-based FBD uses the 3D structure of the protein binding site to design molecules that are expected to optimize ligand–protein interactions.

Several scoring functions, called the primary constraints, can be used to rank the suggested molecules and drive the search in the chemical space. They correspond mainly to those used by docking programs, i. e. force field-based, empirical and knowledge-based scoring functions. In addition, several other physico-chemical parameters related to the drug-likeness of the compounds, as well as terms accounting for molecular and spatial similarity to known ligands, can be used as filters or added to the scoring functions [81, 82]. The latter are called the secondary constraints.

The linking approach (Fig. 2B) starts with the placement of building blocks at key interaction sites of the receptor. This can be done by the fragment-based design software itself, or using a dedicated software like MCSS [76], Solvation Energy for Exhaustive Docking (SEED) [83] or EADock [2]. The latter is particularly suited for the fragment-based approach since, thanks to its cluster-based sampling algorithm and its universally applicable scoring function, it is able to both map fragments favourable positions and dock complete molecules [2]. The positioned fragments are then automatically connected to each other using linkers, resulting in several complete molecules that satisfy all key interaction sites. On the contrary, the growing procedure (Fig. 2C) starts from a single fragment located at one of the key interaction site of the target. This fragment can be chosen by the user or by the program. The structure is then grown from this first fragment iteratively, piece-by-piece. Each addition is made so as to yield favourable interactions between the target and the new fragments, while keeping those already shown by the starting molecule. Connection rules are derived from the existence of certain bonds in organic compounds, or from organic synthesis reactions. Both growing and linking strategies have strengths and weaknesses [59]. Growing might run into difficulties if the active site contains several distinct pockets separated by a large gap in which the interactions between a ligand and the protein are limited. When using a linking approach, slightly misplaced fragments or fragments with loosely defined spatial orientation (like a phenyl ring with no preferred orientation in a large lipophilic binding pocket) can lead to the construction of a suboptimal molecule.

We should not expect ab initio FBD to yield nanomolar compounds in the first instance. Rather, the methods will probably design new perspective lead compounds of medium affinity, which will be the starting point of further optimization [59]. However, FBD techniques already contributed to generate an impressive number of high affinity ligands [84–90] and drug leads for clinical trials, although they were only recently adopted in the drug discovery pipeline. FBD represents a very promising technique to address tomorrow's challenges of drug discovery.

Synthetic accessibility of molecules proposed by FBD ::: In silico fragment-based drug design: One critical aspect of in silico FBD is the synthetic accessibility of the proposed compounds. Obviously, experimental HTS hits are known to be synthesizable, since they have already been synthesized to be present in the tested molecules collection. It can also be expected that their derivatives are accessible using an approach similar to that used for the parental compound. On the contrary, all molecules assembled on the screen of a computer using in silico FBD are not insured to be easily synthesized. However, several strategies can be designed to optimize this aspect. First, drug design studies often aim at deriving new elements of a known class of drugs (the so-called 'me too' approach). In this case, the synthetic issue might be limited thanks to the knowledge already available for such families of molecules. Second, the fragments that are used in silico can be selected to involve organic reactions that are in the core competence of the in house pharmacochemist, or a set of other virtual organic reaction schemes, like in the Retrosynthetic Combinatorial Analysis Procedure (RECAP) [91] or Synthesize and Optimize System in Silico (SYNOPSIS) [88] approaches. Once a few fragments have been successfully assembled in the active site, another option is to screen databases like Zinc (http:// zinc.docking.org/) for compounds containing this motif. The results of this search will provide commercially available molecules, thus for which the synthesis has likely been described and optimized. It is also possible to assess the synthetic accessibility of the candidate compounds by an additional software attempting to define synthetic routes and select potential precursors from databases of available compounds [89, 92]. Similarly, scoring functions have been established

recently that try to mimic the intuition of the organic chemist and estimate the synthetic feasibility of molecules by examining their chemical structures, without suggesting any retro-synthesis [93].

Conclusion:
A brief outline on the most common types of in silico tools has been presented, emphasizing the great progress that in silico drug design has made great changes over the past years, making it a valuable and efficient tool for drug discovery.

Despite the numerous successful studies and the very positive picture that is often drawn, the docking problem is far from being solved [5]. Molecular docking still holds several limitations, like the lack of a universally applicable scoring function, able to efficiently combine accuracy and speed. Several directions of improvements are being investigated, like the use of implicit solvent models and entropic terms. In addition, although ligands are commonly handled with full flexibility, the protein flexibility is still only partially considered, at best. Further studies are still necessary to tackle this issue and address the induced-fit problem. Also, the dynamic inclusion of water molecules during the docking process, to take account of eventually important water-mediated hydrogen bond bridges between the ligand and the protein, could increase the efficiency of the approach. As of today, the results of a docking experiment should be taken with care, and be seen as a good starting point for more involved studies [5].

Several studies have illustrated the ability of vHTS to suggest putative lead compounds, and help its experimental counterpart by reducing drastically the number of molecules that will be effectively tested. However, despite the large efforts that have been deployed, the outcome in terms of new compounds reaching the clinical trials might be seen as rather disappointing [56, 57].

Structure-based FBD could also benefit from a better treatment of the flexibility of the target protein and improvement in binding free energy estimation methods. However, automated de novo design, and in particular FBD, has already proven its value for hit and lead-structure identification [59]. In silico designed molecules can provide the medicinal chemist with rational support to guide his ideas about valuable new chemical entities, and thus help the development of novel and patentable leads.