(To appear in eLS John Wikey & Sons Ltd.)

Protein-Ligand Interactions: Computational Docking

A24027

1st Author:

Dhanik, Ankur

Ankur Dhanik

Rice University

Houston, Texas

USA

2nd Author:

Kavraki, Lydia E

Lydia E Kavraki

Rice University

Houston, Texas

USA

Abstract

A pharmaceutical drug compound is usually a small organic molecule, also termed as ligand, that binds to the target protein and alters the natural activity of the protein, thus, leading to a therapeutic effect. Computational docking or computer-aided docking is an extremely useful tool to gain an understanding of protein-ligand interactions which is important for the drug discovery. Computational docking is the process of computationally predicting the placement and binding affinity of the ligand in the binding pocket of the protein. Docking methods rely on a search algorithm which computes the placement of the ligand in the binding pocket and a scoring function which estimates the binding affinity, i.e., how strongly the ligand interacts with the protein. A variety of methods have been developed to solve the computational docking problems that range from simple point-matching algorithms to explicit physical simulation methods.

Keywords

docking; drug design; protein–ligand docking; flexible docking; flexible receptor; scoring function

Key Concepts

- Computational docking methods play an important role in the drug discovery process.
- A docking method computes the placement of a ligand in the binding pocket of a protein and estimates the binding affinity.
- Rigid-body docking methods treat both the protein and ligand as rigid bodies.
- Flexible-ligand methods treat the ligand as a flexible molecule and flexible-receptor methods treat both the ligand and the protein as flexible molecules.
- Two main features of computational docking techniques are a conformation search algorithm and a scoring function that estimates binding affinity.
- Most of the computational docking programs treat the protein as a rigid molecule and the ligand as a flexible molecule.
- Protein flexibility is an important determinant of the accuracy of docking programs.
- Efforts have been made to account for protein flexibility in docking methods, but more needs to be done.

Introduction

The goal of structure-based pharmaceutical drug design is, for a given protein, to find a ligand, a small molecule that will bind to its active site with high affinity and specificity. Such binding may be attributed to geometric (Figure 1a) and chemical (Figures 1b and c) complementarity. Combinatorial chemistry has made possible the synthesis of literally millions of small molecular compounds. Before the advent of high throughput screening,

testing a potential ligand was an expensive and time-consuming exercise in laboratory biochemistry, and even with modern, robotics-aided high-throughput screening, each candidate molecule must be obtained or synthesized and physically tested. The purpose of computer-aided lead discovery is to select from a pool of hundreds of thousands to millions of candidate molecules, those that are most likely to bind tightly to the active site of the target protein. This dramatically reduces the number of compounds that must be tested in the laboratory, and therefore reduces both the time and expense of the initial screening of potential drugs. See also: DOI: 10.1002/9780470015902.a0001340.pub2, DOI: 10.1038/npg.els.0000056, and DOI: 10.1038/npg.els.0001343

[figure 1 here]

Protein–ligand docking methods may be broken down into three classes based on how much rearrangement they allow the protein and ligand to undergo (Halperin *et al.*, 2002):

- 'Rigid body docking' attempts to dock a fixed conformation of the ligand into a fixed receptor conformation. Receptor conformations may be experimentally determined X-ray crystallography or nuclear magnetic resonance (NMR) structures, or novel, computationally generated structures. The receptor is held in place and different locations and orientations of the ligand are tested. Rigid methods are still common in protein–protein docking, but are rarely, if ever, still used for protein–ligand docking, except as part of a flexible-ligand or fully flexible approach.
- In 'semi-flexible, or flexible-ligand docking', the ligand is allowed freedom to change its conformation, through internal bond rotations (see Figure 2), but the receptor is still held rigid. Flexible-ligand docking is the most commonly used type of docking method.
- Finally, in 'flexible receptor docking', the receptor is also allowed limited flexibility. Because simulating the full flexibility of the protein is usually an intractable problem, this often takes the form of alternative side-chain conformations or the designation of particular bonds to act as hinges. Most new docking programs incorporate at least minimal receptor flexibility.

The remainder of this article will take the following form: In the next section, general features common to all protein–ligand docking methods will be introduced. The subsequent sections on Rigid-body methods, Flexible ligand methods and Flexible receptor methods will introduce selected approaches to rigid, flexible ligand, and flexible receptor docking, respectively, along with examples of programs of each type. Finally, a brief overview of the problem of scoring the results of docking experiments, and some of the functions that are currently used are discussed.

General Features of Protein–Ligand Docking Methods

The search algorithm

Protein–ligand docking is, at a fundamental level, a conformational search problem. That is, regardless of how much flexibility the protein–ligand system is allowed, the problem is essentially one of considering a number of conformations of the biomolecular system and determining which, if any, are likely to be similar to the complex that would be formed in nature. The ligand is modelled either as a rigid object with six degrees of

freedom (that is, its state can be completely specified by six coordinates: three translational, along the X, Y and Z axes, and three rotational to determine its orientation in space) or as a flexible object with the six basic degrees of freedom plus internal rotations about rotatable bonds (see Figure 2). The lengths of bonds and angles between adjacent bonds are much more tightly constrained than rotations about bonds, so these degrees of freedom are usually ignored. If the molecule stably binds to the protein, then the bound conformation should correspond to a conformation of minimal free energy. Solving the docking problem amounts to finding, or at least approximating, that optimal conformation, if it exists.

[figure 2 here]

The difficulty of the search for this conformation increases dramatically (in theory exponentially) with the number of degrees of freedom of the system. An exhaustive search of the conformational space of a protein–ligand system, even if restricted to the relatively limited set of conformations with the ligand in the binding pocket of the protein, would take far too long to be practical for a typical 15- to 25-degree of freedom flexible-ligand/rigid-receptor system, much less the hundreds or thousands of degrees of freedom introduced if receptor flexibility is simulated.

The scoring function

The search algorithm of a docking method generates a set of candidate dockings for a particular ligand, but some mechanism is needed to decide which ligand placements are better than others, and thereby rank conformations against each other. This mechanism is called a scoring function. A scoring function calculates a numerical score for a conformation based on its coordinates. For most purposes, such as high-throughput screening, a scoring function must be simple enough to be computed hundreds to thousands of times per protein/ligand pair, while still corresponding roughly with the free energy of the complex. Unfortunately, it is essentially impossible for a computationally efficient scoring function to take into account the full physics determining the free energy of a protein–ligand complex (Bohm and Stahl, 2002). The effectiveness of any particular scoring function depends on the properties of the system being scored.

The scoring function is used at the end of the search to rank the candidate ligands in terms of their binding affinity for the receptor of interest. For drug discovery, the ligands with the greatest affinity (usually 100–2000, from a database of 100 000–500 000) will usually be selected for further experiments (Bohm and Stahl, 2002). A scoring function may also be used to guide the search as it progresses, and this need not be the same function used to rank the final docked conformations.

Rigid-Body Methods

The earliest docking methods treated both protein and ligand as rigid objects. In rigid-body docking, the ligand has six degrees of freedom: three translational along the cardinal axes, and three rotational. Only one conformation of the ligand is considered. No internal rotations about bonds are permitted in either molecule. Although no modern docking program relies exclusively on rigid-body methods, they remain an integral component of many techniques.

Rigid-body methods break down into three principle types: Clique search, geometric hashing and pose clustering, which is itself a variant of geometric hashing. All are based on finding complementarity between the geometry of the binding pocket and that of the ligand.

Clique detection

A graph is a mathematical object consisting of nodes, which represent some kind of points of interest, and edges, which connect pairs of nodes. In graph theory, a clique is a subset of the nodes of a graph such that each node shares an edge with each other node. Some early docking programs used clique detection to find docked positions of rigid ligands, and clique detection remains an element of several modern, flexible-ligand docking programs, including recent versions of DOCK, the first published docking program (Muegge and Rarey, 2001). In DOCK, to place a given conformation of the ligand, the unoccupied space in the binding pocket of the protein is broken down into spheres, and the docking problem is solved by matching the centres of heavy (nonhydrogen) atoms in the ligand to sphere centres in the binding pocket using clique detection.

Pose clustering and Geometric Hashing

The placement of three noncollinear points of a rigid, three-dimensional object completely defines a position and orientation (or pose) of that object. In pose clustering and geometric hashing, triplets (triangles) of feature points from one object are matched to triplets of points in another object to define a placement of the two objects relative to each other. In docking, the objects of interest are of course the protein-binding pocket and the ligand, and the feature points selected are probable interaction sites.

Like clique detection, pose clustering and geometric hashing are never used as the sole search techniques in modern docking methods. They are used in the process of flexible docking by some docking tools, however. For example, the first step of the FlexX docking program (Rarey *et al.*, 1996) is to place a small, rigid fragment of the ligand in the binding pocket of the receptor using pose clustering. Triangles of ligand atoms are matched to triangles of complementary interaction sites in the binding pocket. After all such matches are computed, the resulting ligand placements are clustered by root mean square distance (RMSD). In docking tools that employ geometric hashing (Fischer *et al.*, 1995; Jackson, 2002), the triplets of points in the ligand and the receptor are placed in bins that are accessed by keys defined by the geometry of the triplets. A receptor-triplet is then matched to the ligand-triplet(s) that is (are) contained in the bin accessed by the key associated with the receptor-triplet.

Flexible Ligand Methods

Rigid-body docking methods may fail to recognize a suitable ligand for a protein because they consider only one internal state of each candidate ligand and receptor. Most small molecules have one or more freely rotatable bonds, allowing them to assume a variety of different, stable conformations, so the next logical step in the evolution of docking methods was to allow for ligand flexibility. Considering multiple ligand conformations makes it less likely that a ligand will fail to dock simply because the conformation chosen was incompatible with the receptor.

Explicit physical simulation of ligand with rigid receptor

Molecular simulations (molecular dynamics and molecular Monte Carlo simulation) existed well before the first dedicated docking program was introduced, so docking by simulation methods is more an application of simulation than a docking technique in itself. Simulating relatively slow events like protein folding and ligand binding is very computationally intensive and time consuming, and requires multiple runs with different starting states for meaningful results, so, although a few groups have reported successful docking using variants of molecular dynamics (Muegge and Rarey, 2001), it is not widely considered a viable approach for large-scale drug discovery, but rather may be used for refinement of dockings discovered by other means. See also: DOI: 10.1002/9780470015902.a0003048.pub2, and DOI: 10.1002/9780470015902.a0001341.pub2

MolSoft LLC's Internal Coordinate Mechanics (ICM) program (Abagyan *et al.*, 1994) performs a Monte Carlo-like search of the conformation space of a flexible ligand in the force field generated by the protein. The program performs several types of random perturbations on the ligand, combined with gradient descent minimization to identify local energy minima. It maintains a history of local minima that have already been visited to bias the simulation towards unexplored areas (Perola *et al.*, 2004).

Ligand fragmentation methods

Several of the more popular of the current flexible docking programs approach ligand flexibility by breaking the ligand down into fragments that are treated as rigid bodies and then reassembling it in the binding pocket of the protein.

In 'incremental construction' approaches, such as FlexX (Rarey *et al.*, 1996), DOCK (Moustakas *et al.*, 2006), and MS-DOCK (Sauton *et al.*, 2008), an anchor fragment of the ligand is placed in the binding pocket using a rigid-body approach. The rest of the ligand is then added to this anchor fragment piece by piece. Backtracking of this incremental construction is allowed, generating many possible ligand conformations for a single placement of the anchor fragment.

In 'place-and-join' methods, the ligand is broken into overlapping fragments. Each fragment is docked to the binding pocket using rigid-body methods, then overlapping fragments whose linker segments are sufficiently close together are joined to reassemble the ligand (Halperin *et al.*, 2002).

SURFLEX (Jain, 2003, 2007) is another method that employs both incremental construction and place-and-join approaches. First different types of molecular fragments (CH_4, C=O, and N-H) are placed in the binding site to form an idealized ligand. Then the actual ligand to be docked is broken into fragments which are aligned to the idealized ligand using pose clustering algorithm. The aligned fragments are then joined using an incremental construction or place-and-join based approach to produce a docked conformation of the ligand.

Pose filtering

The package FRED by OpenEye Scientific Software uses a sequential filtering approach to docking. It first generates an extensive library of conformations of the ligand. Those that

have adequate shape complementarity to the binding site of the protein are kept in the first phase of filtering. Conformations in the binding site can then be tested against user-defined pharmacophore maps, and up to three scoring functions. Minor conformational rearrangements may be performed to optimize scores (Schulz-Gasch and Stahl, 2003).

The Glide (Grid-based ligand docking with energetics) program by Schrodinger, Inc. is also a similar hierarchical filtering approach. A large set of minimal energy ligand structures is generated and clustered. These clusters are then docked into a force field representing the protein-binding pocket. A few hundred candidates are selected at this stage, and are then subjected to minimization under Van der Waals and electrostatic forces. Finally, approximately 10 structures are selected for randomized optimization of peripheral torsional angles, and the resulting structures are scored and reported (Perola *et al.*, 2004).

Randomized search methods

Docking is an example of a common type of problem in computer science, in which finding the solution requires searching a prohibitively large set of candidate solutions. A great deal of theoretical computer science research has gone into the development of algorithms to efficiently search the important parts of such large state spaces. Two of these, simulated annealing and genetic algorithms, have been applied successfully to protein–ligand docking.

The term 'simulated annealing' comes from an analogy to the cooling and solidification of metals in the smelting process. A simulated annealing search begins at an arbitrary ligand placement. For each step of the search, the degrees of freedom of the ligand are randomly perturbed. The resulting structure is accepted with some probability that depends on (1) the relative scores (usually estimated free energies) of the original and new states and (2) the current 'temperature' of the search. The higher the temperature, the more likely a move is to be accepted, even if it results in a worse score. As the search progresses, the temperature gradually decreases, and the search favours better scoring conformations more and more. The final structure has a high probability of being at a locally optimal score. Theoretically, if the temperature decrease is infinitely gradual, the result will be the globally optimal conformation. All versions of the program Autodock since version 2 have the option to perform docking using simulated annealing (Morris *et al.*, 1998).

Genetic algorithms perform searches using a system loosely analogous to evolution by natural selection (see Figure 3). Each possible solution is encoded as a string of numbers, analogous to a chromosome. The assumption underlying the use of genetic algorithms is that two partial solutions, combined in the right way, may yield a better solution. An initial population of candidate solutions, usually randomly generated, is established, and then, through selection, recombination and point mutation, the population evolves with time.

[figure 3 here]

The docking program GOLD (Jones *et al.*, 1997) uses a genetic algorithm to find docked conformations of a flexible ligand. In GOLD, each chromosome consists of two strings: the first is a string of torsional angles, with one entry per rotatable bond in the ligand.

The other is a string of integers coding for hydrogen bond interactions. The probability of selection is based on an approximation of the free energy of the corresponding protein–ligand complex. Because the outcome of a single run of a genetic algorithm is random, GOLD's results for a given protein–ligand pair are based on many independent runs.

Autodock (Morris *et al.*, 1998, 2009) uses alternating genetic algorithm and minimization steps, in what its developers call a Lamarckian genetic algorithm, after the failed genetic theory of Jean-Baptiste Lamarck, which held that an individual may pass on changes acquired during its life to its offspring. Another related program called Vina (Trott and Olson, 2010) combines a stochastic global optimizer and a gradient-based local optimizer for exploring the conformation space of the flexible ligand.

Flexible Receptor Methods

Most of the protein–ligand docking approaches assume, for simplicity, that the protein is a rigid, immovable object. In reality, however, most proteins of pharmaceutical interest are flexible objects, constantly shifting from one stable conformation to another (see Figure 4). When they bind a ligand, many proteins undergo a conformational change in order to better accommodate the ligand and its physical properties in their binding pockets. This phenomenon is called 'induced fit'. See also:
DOI: 10.1038/npg.els.0003140, and DOI: 10.1038/npg.els.0003012

[figure 4 here]

A docking program that assumes the protein is rigid might fail to dock a molecule that forms a stable complex with the protein in a previously undocumented conformation. It has been demonstrated that rigid receptor methods have a high probability of failing to dock ligands to some flexible proteins when the wrong protein conformation is used (Österberg *et al.*, 2002). While some ligands readily bind to a variety of conformations, others are highly selective. Thus, in drug discovery, failure to allow induced fit and conformational change may translate into passing over viable leads. Docking methods allowing protein flexibility are therefore an active area of research (Teodoro and Kavraki, 2003; Kokh *et al.*, 2011).

Cross docking

Cross docking is an attempt to dock a known ligand of a protein to a known conformation of the protein other than the one to which it binds. Given a crystal structure of a ligand bound to some conformation of a protein, a flexible-receptor docking method should, ideally, be able to reconstruct that complex regardless of the starting conformation of the protein. Cross docking experiments are thus a rigorous test of the efficacy of any flexible-receptor docking method.

Cross docking is also the basis of the simplest flexible receptor docking approach. Such an approach consists simply of attempting to dock each candidate ligand to all known crystallographic and NMR structures of the protein using some established flexible ligand approach, such as those introduced in the section on Flexible ligand methods. The obvious drawback of this approach is that the docking attempt will fail if the structure to which the ligand binds has not yet been discovered. It is also inefficient in that it requires, for each ligand, as many runs of the chosen docking algorithm as there are crystal structures of the protein (Totrov and Abagyan, 2008).

Ensemble methods

Another approach to flexibility is to model the protein as an ensemble of structures. An ensemble is generally a compact representation of a variety of conformations of the protein, which may come from experimental data or be generated by some computational process. The docking target in an ensemble method may consist of a superimposition of multiple structures or a set of separate structures.

The program FlexE (Claussen *et al.*, 2001), which is an extension of FlexX, uses a 'united' protein description. All structures of the protein documented by X-ray crystallography and NMR are merged into a representation in which parts of the molecule that do not vary significantly are represented singly, and parts that do vary, either by position or due to a point mutation, are represented as alternative substructures. Docking is performed using every geometrically feasible combination of known states for the variable regions.

Similar to GOLD and Autodock, the program FITTED (Corbeil *et al.*, 2007) uses a genetic algorithm based approach. The flexibility of the main chain and side chains, modeled by a library of conformations, of the protein is also encoded in the chromosome alongwith the flexibility of the ligand. The chromosome is evolved over time using genetic operators to obtain a solution chromosome that represents the docked conformation of the ligand.

Another docking method called 4-dimensional (4D) method (Bottegoni *et al.*, 2009) treats the protein flexibility as a fourth dimension, in addition to the three dimensions in which the ligand resides. Multiple protein structures in the ensemble are indexed by a discrete variable which is included in the optimization procedure that computes the optimal conformation of the ligand bound to one of the protein structures.

Local flexibility methods

Some approaches allow flexibility of the protein at specific points. For example, many protein–ligand docking programs maintain the overall structure of the protein but allow the side chains rotational freedom about their rotatable bonds. For example, the SPECITOPE tool (Schnecke *et al.*, 1998), as its final step, attempts to resolve ligand–receptor overlap by rotating offending side chains out of the. Some of the other examples are GOLD, Autodock and Vina docking programs that allow user-specified side chains to rotate during the docking procedure. ROSETTALIGAND (Meiler *et al.*, 2006) is another program that uses side chain repacking algorithm to model the protein flexibility. An extension of this program (Davis and Baker, 2009) also accounts for the flexibility in the backbone of the protein by performing a gradient-based minimization of the ligand conformation and the backbone and side chain torsion angles of the protein.

Other approaches analyse the structure of the protein for regions of high-expected flexibility, and allow them to move while holding the rest of the internal degrees of freedom rigid. In principle, this allows the protein an approximation of realistic flexibility, while keeping the problem computationally tractable (Teodoro and Kavraki, 2003).

Scoring Functions

At the end of a docking search, and throughout the search for some methods, a docking program needs a way to rank the candidate docked conformations it has found relative to each other in order to report those most likely to represent the binding conformation and affinity of the complex in nature. This ranking system, called a scoring function, takes the position of each ligand atom (and protein atom, for flexible-protein methods) and returns a numerical score. Some docking programs use two different scoring functions: one to guide the search, and another to rank the final set of prospective dockings.

Scoring approaches generally fall into three categories: force-field-based potential functions, empirical functions and knowledge-based or statistical functions (Sousa *et al.*, 2006).

Force field/physical potential-based scoring functions

Force fields were originally developed for use in molecular modelling and simulations. A force field takes the atom coordinates of a molecular system and calculates an estimate of its potential energy by explicitly modelling physical forces such as Van der Waals interactions, the resistance of bonds to bending and stretching, steric interactions due to torsional rotation about rotatable bonds and electrostatic forces (Bohm and Stahl, 2002).

For docking, often only the intermolecular forces (Van der Waals and electrostatic forces) are considered for final scoring, while the full energy may be used to guide the search if the protein and/or ligand is allowed to change conformation. The intermolecular energy of the docked complex provides an estimate of the binding affinity of the protein–ligand combination.

Some force field-based scoring functions have been augmented with a solvation term. This term attempts to account for the free energy change due to the loss of solvent accessibility to the binding pocket and ligand surfaces. As an example, while grid-based scoring function in the DOCK (Moustakas *et al.*, 2006) program takes into account Van der Waals and electrostatic interactions, another scoring function (DOCK3.5) in the same program also includes a solvation term. Note that it is standard for docking programs to have a multitude of scoring functions.

Empirical scoring functions

Empirical scoring functions, such as ChemScore used in Glide and the scoring function in Autodock, estimate the binding affinity of a protein–ligand pair by counting standard types of interactions and assuming an average contribution for each to the free energy of the system. Typically, the interactions include hydrogen bonds, strong electrostatic interactions (salt bridges), hydrophobic contacts, solvent-accessible surface area and, often, an entropic term proportional to the number of rotatable bonds made rigid due to binding (Bohm and Stahl, 2002).

Hydrogen bonding and ionic terms are usually assessed simply by counting the number of such interactions in the system. Some scoring functions weight each hydrogen bond by an angular term that penalizes for deviations from ideal hydrogen bond geometry. Hydrophobic interaction scores are usually weighted based on the area of the protein–ligand contact surface.

The average free energy of each type of interaction is found empirically, from examples with experimentally determined binding affinities, either by least-squares fitting or by a machine learning approach such as neural networks.

Knowledge-based scoring functions

Knowledge-based scoring functions, such as DrugScore (Velec *et al.*, 2005), and ASP scoring function (Mooij and Verdonk, 2005) in GOLD docking program, are developed through a statistical analysis of known protein–ligand complex structures, for example, from the RCSB Protein Data Bank. Atom-type pairs that are found in contact with each other more often than expected assuming random spatial distribution are considered to contribute to the binding affinity, while atom types that are found to be in contact less often than expected by chance are considered to have a negative contribution to the binding affinity (Bohm and Stahl, 2002). Some docking programs, such as ROSETTALIGAND, use scoring functions that are comprised of knowledge-based as well as force-field based terms.

Consensus scoring

Combining the results of several established scoring functions can help overcome their individual weaknesses (Halperin *et al.*, 2002). Called consensus scoring, this approach generally has the effect of decreasing the number of false positives, but also may lead to failure to identify active ligands that are well identified by only one function (Bohm and Stahl, 2002). VoteDock (Plewczynski *et al.*, 2011) is a docking method that employs a consensus scoring function called MetaScore which combines scores from docking programs such as FlexX, Glide, GOLD, SURFLEX, and others.

Conclusion

The research area of computational docking has over the years seen and continues to see the development of a number of docking programs. Given limited computational resources a couple of decades back, initial docking programs focused on rigid-docking approaches, which eventually evolved to flexible-ligand based approaches. It is now well understood that flexibility of the receptor (such as protein) plays an important role in the binding of a ligand to the receptor. Increase in the computational resources has helped many new methods to incorporate the flexibility of the protein into the search algorithms, but the incorporation of the flexibility adds to the computational expense of docking. Most of the new methods therefore do not model protein flexibility in a realistic fashion. Ongoing challenges facing the research community, thus, include the addition of more realistic protein flexibility and the development of more accurate scoring functions without sacrificing the speed of their computation. In the case of large ligands which are important for the discovery of peptide-based drugs, an emerging drug discovery paradigm, the increased flexibility of the large ligands itself is a challenge for the research community.

Despite the scope for improvment in docking programs, they have proven successful as an initial tool for lead discovery for the purpose of screening large databases of molecules for activity against a protein of therapeutic interest (Schneider, 2010). Docking programs facilitate easy comparison of docked conformations of the ligand and a reference conformation of the ligand that is usually obtained from the RCSB Protein Data Bank. Many studies (such as Li *et al.*, 2010; Plewczynski *et al.*, 2011) have been

performed that demonstrate how robustly different docking programs can predict known ligand conformations. These studies reveal that the docking programs are still not sophisticated and accurate enough to be able to predict the binding interactions and binding affinity between a putative drug compound and its target receptor in a robust manner. It should be noted here that computing tightly docked conformations of a putative drug compound is just a piece of the puzzle that is computer-aided drug discovery. A more holistic computational approach that includes factors such as solubility, specificity, etc. is needed to make the drug discovery process more efficient. As advances in physical chemistry enable more accurate molecular models, and as docking algorithms become more sophisticated, the research community is constantly working on new methods to solve an important piece of the drug discovery puzzle.

Glossary

Conformation
A geometric state of a molecule. Generally, bond lengths and angles are assumed to be constant, so a conformation of a molecule may be completely specified by torsional angles. Alternatively, a conformation may be specified by three-dimensional Cartesian coordinates of each atom.

Degrees of freedom
The number of variables necessary to completely specify the geometric state of a system. For example, a ligand with one rotatable bond has seven degrees of freedom: three translation along the cardinal axes, three rotational about the cardinal axes and one internal rotation about its rotatable bond.

Induced fit
In the process of binding to a ligand, proteins often undergo conformational changes that enhance the binding affinity both sterically (geometrically) and energetically. This phenomenon is called induced fit.

RMSD
Root mean squared distance, a common measure of the difference between two alternative conformations of a molecule or molecular system. It is calculated as the square root of the sum of the squares of the displacements of each atom between two conformations.

Torsional angle
Given three consecutive, non-collinear bonds, the torsional angle is the angle formed between the first and third bonds in a Fischer projection across the second.

Flexible ligand methods
Flexible ligand methods allow changes in the internal state of the ligand. Thus, the conformation space of the ligand, which is explored to compute the docked conformations of the ligand, is composed of the rigid body degrees of freedom as well as internal degrees of freedom (most commomly due to rotations around bonds).

Flexible receptor methods
Flexible receptor methods account for the flexibility of the receptor as part of the docking procedure. Methods can range from those that dock a flexible ligand to the multiple experimentally determined structures of the receptor, to those methods that computationally model the flexibility of the receptor using different strategies.

Scoring functions

Scoring functions estimate the thermodynamic stability of the receptor-ligand complexes and thus provide a measure to rank and ascertain the quality of the different docked conformations of the ligands.

References

Abagyan R, Totrov M and Kuznetsov D (1994) ICM – a new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation. *Journal of Computational Chemistry* **15**: 488-506.

Bohm H-J and Stahl M (2002) The use of scoring functions in drug discovery applications. *Reviews in Computational Chemistry* **18**: 41–87.

Bottegoni G, Kufareva I, Totrov M and Abagyan R (2009) Four-dimensional docking: a fast and accurate account of discrete recetor flexibility in ligand docking. *Journal of Medicinal Chemistry* **52**(2): 397-406.

Claussen H, Buning C, Rarey M and Lengauer T (2001) FlexE: efficient molecular docking considering protein structure variations. *Journal of Molecular Biology* **308**(2): 377–395.

Corbeil CR, Englebienne P and Moitessier N (2007) Docking ligands into flexible and solvated macromolecules. 1. development and validation of FITTED 1.0. *Journal of Chemical Information and Modeling* **47**(2): 435-449.

Davis IW and Baker D (2009) ROSETTALIGAND docking with full ligand and receptor flexibility. *Journal of Molecular Biology* **385**(2): 381-392.

Fischer D, Lin SL, Wolfson HL and Nussinov R (1995) A geometry-based suite of molecular docking processes. *Journal of Molecular Biology* **248**(2): 459-477.

Halperin I, Ma B, Wolfson H and Nussinov R (2002) Principles of docking: an overview of search algorithms and a guide to scoring functions. *Proteins* **47**(4): 409–443.

Jackson RM (2002) Q-fit: a probabilistic method for docking molecular fragments by sampling low energy conformational space. *Journal of Computer-Aided Molecular Design* **16**(1): 43-57.

Jain AN (2003) Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *Journal of Medicinal Chemistry* **46**(4): 499-511.

Jain AN (2007) Surflex-Dock 2.1: robust peformance from ligand energetic modeling, ring flexibility, and knowledge-based search. *Journal of Computer Aided Molecular Design* **21**(5): 281-306.

Jones G, Willett P, Glen R, Leach A and Taylor R (1997) Development and validation of a genetic algorithm for flexible docking. *Journal of Molecular Biology* **267**(3): 727–748.

Kokh DB, Wade RC and Wenzel W (2011) Receptor flexibility in small-molecule docking calculations. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **1**(2): 298-314.

Li X, Li Y, Cheng T, Liu Z and Wang R (2010) Evaluation of the performance of four molecular docking programs on a diverse set of protein-ligand complexes. *Journal of Computational Chemistry* **31**(11): 2109-2125.

Meiler J and Baker D (2006) ROSETTALIGAND: protein-small molecule docking with full side-chain flexibility. *Proteins* **65**(3): 538-548.

Mooij WT and Verdonk ML (2005) General and targeted statistical potentials for protein-ligand interactions. *Proteins* **61**(2): 272-287.

Morris G, Goodsell D and Halliday R *et al.* (1998) Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *Journal of Computational Chemistry* **19**(14): 1639–1662.

Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS and Olson AJ (2009) AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. *Journal of Computational Chemistry* **30**(16): 2785-2791.

Moustakas DT, Lang PT, Pegg S, Pettersen E, Kuntz ID, Brooijmans N and Rizzo RC (2006) Development and validation of a modular, extensible docking program: DOCK 5. *Journal of Computer Aided Molecular Design* **20**(10-11): 601-619.

Muegge I and Rarey M (2001) Small molecule docking and scoring. *Reviews in Computational Chemistry* **17**: 1–60.

Österberg F, Morris GM, Sanner MF, Olson AJ and Goodsell DS (2002) Automated docking to multiple target structures: Incorporation of protein mobility and structural water heterogeneity in AutoDock. *Proteins* **46**(1): 34–40.

Perola E, Walters WP and Charifson PS (2004) A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins* **56**(2): 235–249.

Plewczynski D, Lazienwski M, Grotthuss MV, Rychlewski L and Ginalski K (2011) VoteDock: consensus docking method for prediction of protein-ligand interactions. *Journal of Computational Chemistry* **32**(4): 568-581.

Rarey M, Kramer B, Lengauer T and Klebe G (1996) A fast flexible docking method using an incremental construction algorithm. *Journal of Molecular Biology* **261**(3): 470–489.

Sauton N, Lagorce D, Villoutreix BO and Miteva MA (2008) MS-DOCK: accurate multiple conformation and rigid docking protocol for multi-step virtual ligand screening. *BMC Bioinformatics* **9**:184.

Schnecke V, Swanson CA, Getzoff ED, Tainer JA and Kuhn LA (1998) Screening a peptidyl database for potential ligands to proteins with side-chain flexibility. *Proteins* **33**(1): 74-87.

Schneider G (2010) Virtual screening: an endless staircase? *Nature Reviews Drug Discovery* **9**: 273-276.

Schulz-Gasch T and Stahl M (2003) Binding site characteristics in structure-based virtual screening: evaluation of current docking tools. *Journal of Molecular Modeling* **9**(4): 47–57.

Sousa SF, Fernandes PA and Ramos MJ (2006) Protein-ligand docking: current status and future challenges. *Proteins: Structure, Function, and Bioinformatics* **65**(1): 15-26.

Teodoro M and Kavraki L (2003) Conformational flexibility models for the receptor in structure based drug design. *Current Pharmaceutical Design* **9**(20): 1635–1648.

Totrov M and Abagyan R (2008) Flexible ligand docking to multiple receptor conformations: a practical alternative. *Current Opinion in Structural Biology* **18**(2): 178-184.

Trott O and Olson AJ (2010) AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry* **31**(2): 455-461.

Velec HFG, Gohlke H and Klebe G (2005) DrugScore^{CSD}-knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction. *Journal of Medicinal Chemistry* **48**(20): 6296-6303.

Further Reading

Teague SJ (2003) Implications of protein flexibility for drug discovery. *Nature Reviews Drug Discovery* **2**(7): 527-541.

Fradera X and Mestres J (2004) Guided docking approaches to structure-based design and screening. *Current Topics in Medicinal Chemistry* **4**(7): 687–700.

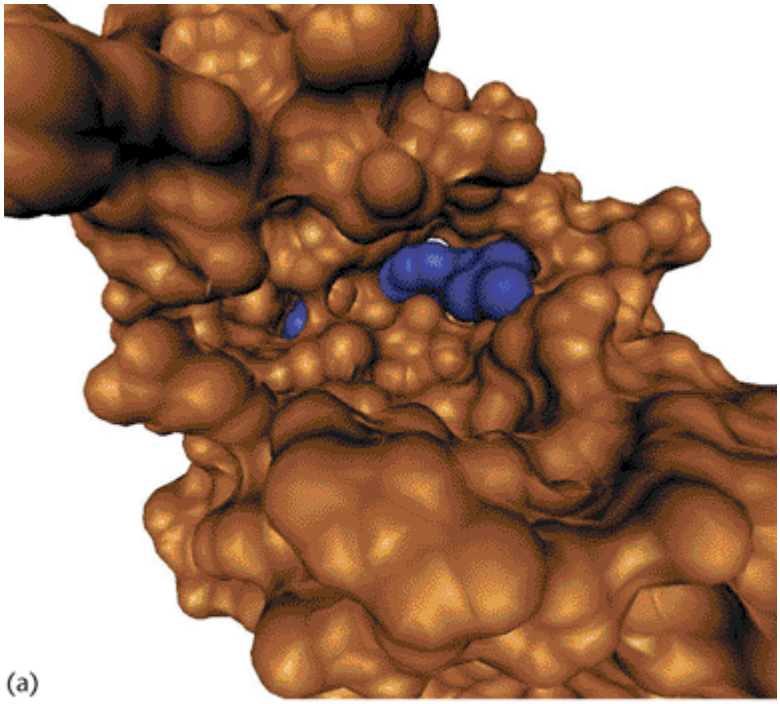
Mobley DL and Dill KA (2009) Binding of small-molecule ligands to proteins: "what you see" is not always "what you get". *Structure* **17**(4): 489-498.

Schlick T (2010) *Molecular Modeling and Simulation: An Interdisciplinary Guide*. New York: Springer ISSN 0939-6047.

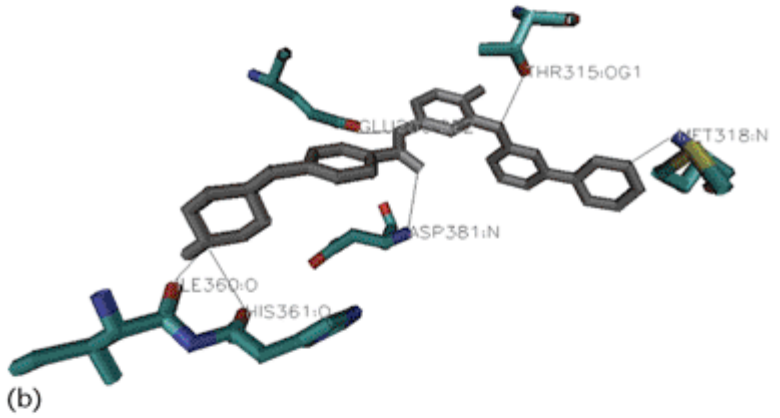
Plewczynski D, Lazniewski M, Augustyniak R and Ginalski K (2011) Can we trust docking results? Evaluation of seven commonly used programs on PDBbind database. *Journal of Computational Chemistry* **32**(4): 742-755.

Waszkowycz B, Clark DE and Gancia E (2011) Outstanding challenges in protein-ligand docking and structure-based virtual screening. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **1**(2): 229-259.

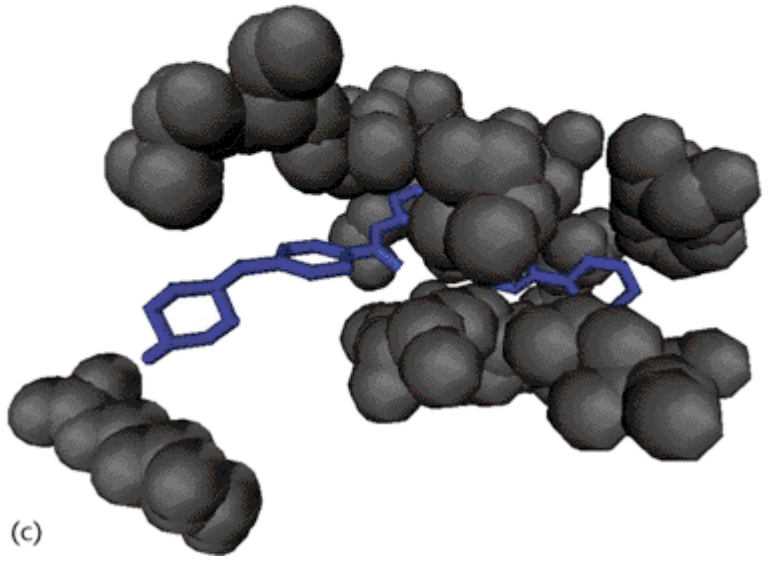
Figure 1. (a) The anticancer drug imatinib (blue) in the binding pocket of the Abelson kinase (orange), a proto-oncoprotein. A mutant version of this protein is involved in the development of chronic myelogenous leukaemia. Note the marked geometric complementarity of the two molecules: atoms of the drug occupy cavities in the surface of the binding pocket. Both molecules are rendered as Connolly surfaces. PDB structure 1IEP. (b) Several amino acid residues of the binding pocket of the Abelson kinase (coloured) form hydrogen bonds with partially charged atoms on a molecule of imatinib (grey). These interactions help define the chemical complementarity between protein and ligand that enables stable binding. (c) Hydrophobic amino acid residues of the Abelson kinase binding pocket help stabilize the hydrophobic rings of imatinib. Additionally, the presence of hydrophobic groups helps exclude water from the binding pocket, which might otherwise interfere with the hydrogen bonds illustrated in (b).



(a)



(b)



(c)

Figure 2. The effect rotation about the bond between atom groups B and C. This is the type of motion responsible for most large-scale rearrangements of organic molecules, including proteins and ligands.

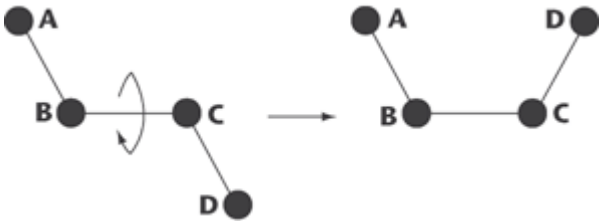


Figure 3. The basic operations in a genetic algorithm. In the initial population each gene is a string of numbers that represent a possible solution to the problem at hand (in this case, the docking of a ligand to a protein). In selection, several genes are randomly chosen from the initial population, with a bias based on a fitness function, perhaps the score of the docking each represents. In crossing over, pairs of genes exchange a part of their sequence. The genes resulting from crossing over are then copied in sufficient quantity to restore the original size of the population. Finally, in the mutation phase, some points of some genes are randomly changed. If the genes are represented as strings of bits, for instance, bits selected for mutation are flipped from 0 to 1 or vice versa. Thus, even genes sharing a common parent are likely to be slightly different, allowing a gradual evolution of the solution.

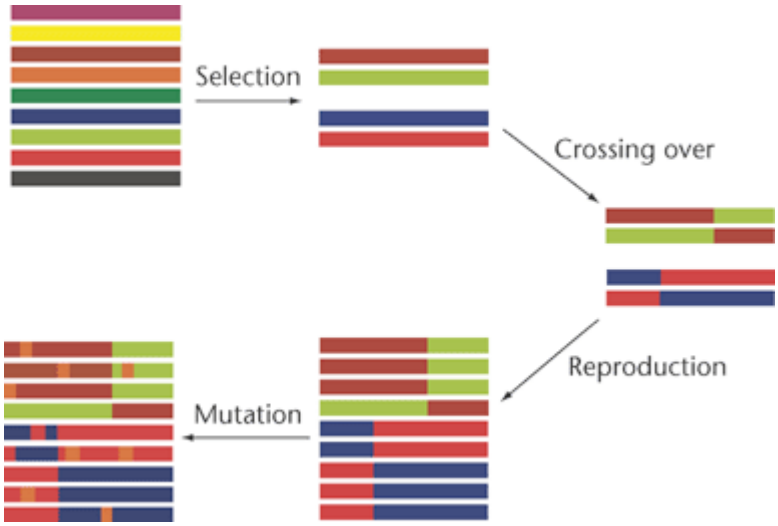
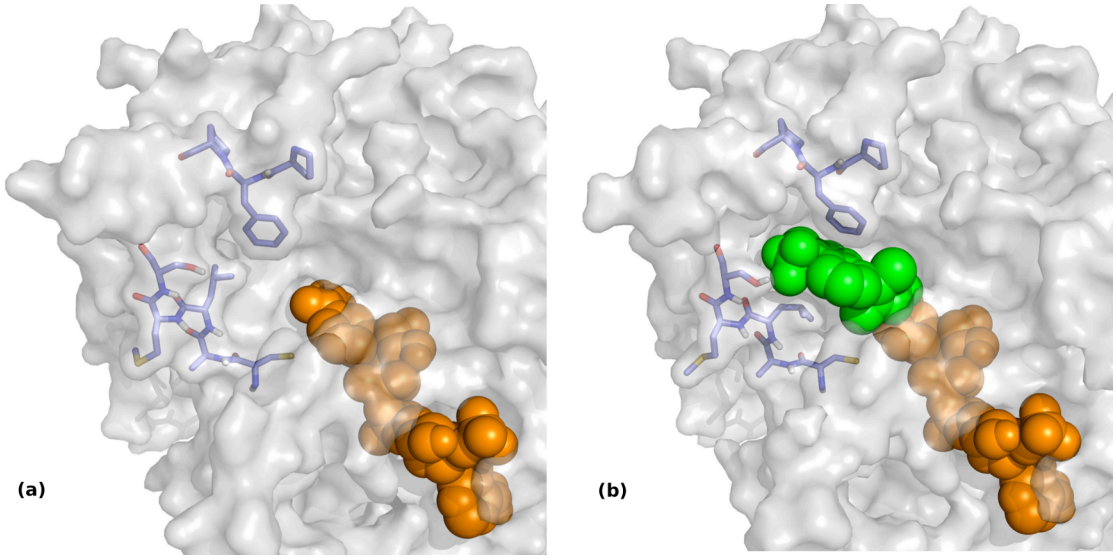


Figure 4. Protein flexibility is important for ligand binding in the aldose reductase enzyme (grey surface representation) that plays a role in diabetes-related complications. (a) PDB structure 1AH4, holo conformation of aldose reductase in complex with the coenzyme (shown as orange spheres), (b) PDB structure 1AH3, a few residues (shown as sticks) that surround the binding pocket of the aldose reductase change conformation to allow binding of pharmaceutical inhibitor tolrestat (shown as green spheres).


(a)
(b)