Protein-Ligand Docking: A Critical Review of Molecular Dynamics, Robotics, and Rotamer Library Methods

-by Thomas Butler

Introduction

The hope of pharmaceutical companies for many years has been that computer-aided rational drug design would lead to explosive growth in the area of discovering novel drug molecules. While this goal has not been fully realized (for a number of reasons), there is much reason to remain hopeful. The current strategy for drug design focuses on using computational tools to find bound conformations of ligands to larger receptor proteins of a known structure. Ligands that bind specifically to a certain proteins can lead to enzyme inhibition or modulation of signal transduction, and thus can be used as drugs. (Rarey, 1996)

The main reason why the process of protein-ligand docking is so difficult is the tremendous complexity of the system; one must take into account hundreds of thousands of degrees of freedom in the two molecules, as well as the not-completely-known combination of energetic forces acting on them. (Singh, 1999) The dilemma becomes that while one needs to make simplifying assumptions to allow for reasonable computation time, these assumptions can also lead to unusable results. (Goodsell, 1990)

Most popular methods used today to predict protein-ligand docking are based on molecular dynamics, robotics, or rotamer libraries. Though not a fully comprehensive look at the field, this paper will describe AutoDock, a program that uses various methods including molecular dynamics, as well as robotics and rotamer library/side chain methods. An analytical discussion of their respective strengths and limitations will follow.

DOCK

This paper will deal primarily with current applications for protein-ligand docking, but it would be incomplete without mentioning DOCK, the pioneer program developed two decades ago by Kuntz at UCSF. (Kuntz, 1982) Very briefly, Kuntz's strategy for predicting ligand binding sites centered around placing small spheres on the receptor protein's surface and matching up the inter-sphere distances to the inter-atom distances of the ligand. Using this information in conjunction with energetic interactions between the rigid ligand and rigid receptor, he was able to determine a putative ligand docking position. Today there are more sophisticated methods available for predicting docking, but most have evolved from Kuntz's work.

AutoDock

The more recent AutoDock (current version 4.0) employs molecular dynamics to predict the docking of a flexible ligand to a binding site of a rigid protein, given the region of the protein containing the binding site and the substrate. (Goodsell, 1990) For maximum predictive accuracy as well as reasonable computational time, it employs the AMBER force field in conjunction with free energy scoring functions and a large set of protein-ligand complexes with known protein-ligand constants. (Website 2) The individual components of the program, whose basics will be explained below, are: AutoTors, AutoGrid, and AutoDock.

AutoTors is the simplest of the components – it defines which bonds in the ligand are rotatable, affecting the degrees of freedom (DOF) of the ligand, and thus the complexity of the computations. (Goodsell, 1996) Each rotatable torsional angle adds an extra DOF, so large ligands with many torsional angles quickly become too complex to compute.

AutoGrid pre-calculates a three-dimensional grid of interaction energies based on the macromolecular target using the AMBER force field. Since the structure of the receptor protein is rigid and known, interaction energies between the probe and surrounding amino acids can be calculated at each point in the grid and stored in a table. Additional tables are made for each atom type in the ligand, taking into account

dispersion/repulsion and hydrogen bonding energies. A second grid is made to allow for electrostatic effects, using a probe with a single positive charge.

After the grid has been completed, AutoDock can begin the simulation. First, the ligand moves randomly in any one of six degrees of freedom (either translation or rotation) and the energy of the new ligand "state" is calculated. If the energy of the new state is higher than the old state, the new one is automatically accepted as the next step in docking. However, if it is higher, then the step is accepted by the following probability function:

$$P(\Delta E) = \exp(-\Delta E/kb*T)$$

(kb = Boltzmann's constant, E = energy, T = temperature)

The system starts at a high temperature T in order to accept most initial steps. The steps are cycled, and at the beginning of each new cycle, the temperature is reduced, making it more it progressively more difficult for the docking to procede to a new step. A final low energy bound conformation is returned.

A weakness with this method is that the initial freedom of movement of the ligand makes it likely that the most energetically favored binding conformation will not be observed. (Goodsell, 1996) This problem is addressed by performing separate simulations starting from random initial states to find consistent conformations. However, finding unknown binding sites in a protein is not AutoDock's strength.

Rotamer Libraries

The folding of a polypedtide chain forms crevices on a protein surface, where sets of amino acid side chains interact so that they can only make non-covalent bonds with certain ligands. (Website 2) Steric interactions that cause the sidechain to twist away from neighboring atoms can inflict high dihedral energy on residues, causing certain rotameric states to be higher in energy than others. (Website 3) The regularities in residue side chain conformations are known as rotamers, and a rotamer library describes

the different conformations of each side chain seen among known protein structures. (Website 3) Side chain interactions are thought of as either "backbone-dependent," meaning that they are dependent on the conformation of the local backbone of the protein, or "backbone-independent," meaning that they are not. (Website 3) SCRWL is an example of a program that places sidechains on a backbone based on preferences found in backbone-dependent rotamer libraries and the evaluation of steric interactions.

Levitt and Koehl use rotamer libraries to help identify amino acid sequences that stabilize a given target protein conformation, and thus model the proteins conformations most likely to be observed. (Koehl, 2002) After rotamer conformations are taken from a rotamer library, side chain interaction energies are stored in a matrix, where van der Waals interactions, electrostatics, and an environment free energy (for the solvent) are taken into account. (Koehl, 2002) These interactions affect the final structure of the protein.

To determine docking, one can add the candidate ligand to the system, and then rapidly calculate the favorable side-chain conformations in the presence of the ligand. Unlike the other methods where the protein is considered completely rigid, the rotamer library/side chain method allows flexibility of the side chains in the protein. This allows us to see a docking that might be closer to what is actually happening, because the assumption of a rigid receptor protein is clearly false and arguably the weakest point of all previous methods. The modeling of course is much cruder than would be allowed in a continuous molecular dynamics simulation (and it only considers side chains), but given the realities of computational power, it is a significant improvement.

A major weakness of using rotamer libraries is that most of the time one does not know exactly where the ligand will bind. (Havranek, personal communication) This means that side-chain conformations must be determined for a large number of protein sequences in the presence of a ligand. Most of the time it is not possible to evaluate all sequences, so a heuristic search algorithm would be necessary to find the optimal sequence. (Havranek, personal communication)

Robotics

Whereas most techniques consider only static properties of binding, such as the energy of the final bound ligand-receptor complex, Singh, Latombe, and Brutlag suggest the novel approach of utilizing robotic motion planning to study the motion of the ligand during the process of binding. (Singh, 1999) The reasoning behind this method is that by sampling the space of all possible paths that the ligand can take to bind to its receptor, one can predict intermediate configurations of the ligand and get a distribution of energetically favorable paths to the binding site. (Singh, 1999)

The key to the robotics approach is that a flexible ligand can effectively be thought of as an "articulated robot" – with the arms and joints of the robot corresponding to the atoms and bonds of a ligand. With this model, 5 degrees of freedom are initially assigned: 3 for the coordinates, and 2 to specify the orientation of one band. Additional degrees of freedom are assigned as each additional non-terminal atom requires a torsional angle to define the ligand's orientation.

The next step taken is to adapt robotic motion planning to find the potential paths that the ligand may take through an energy distribution of the workspace. The path is calculated in the configuration space (C-space) of the robot, which includes one dimension for each DOF. Thus, it follows that larger molecules with many DOF's quickly become unmanageable to compute. Probabilistic Roadmap Planners (PRM's) (Kavraki, 1996), which efficiently handle large DOF's in robots, are used to generate "milestones" in C-space through random sampling, each of which is connected to several close neighbors by a physically realizable path. The undirected graph that results from this process provides the framework for the path chosen; and because each point in C-space corresponds to a particular configuration of the ligand, the total energy can be computed by summing the interactions between the atoms in the ligand. To compute the energy of interaction between the ligand and receptor, an electrostatic potential grid and a van der Waals potential grid are created, taking into account solvent and ionic effects. The energies of interaction can then be determined at each point by looking at the van der Waals and electrostatic potentials at grid points closest to the ligand atoms.

The application of PRM that Singh uses here differs from the one typically used in robotics in that instead of using a binary test to detect collisions of robot and obstacle, it computes a continuous energy value for all the points along the path. As a final result,

the robotics method assigns an average weight for all paths entering and leaving a potential binding site. A high weight indicates an energetically unfavorable path, and a low weight suggests a favorable path.

A surprising result observed in testing this method is that the average weights entering and leaving the true binding site is much higher than at non-binding sites. (Singh, 1999) However, since the weight for leaving the binding site is observed to be much higher than entering, it makes sense that once the ligand is bound the energy barrier for leaving is so high that it remains bound. Using the logic that higher overall weights along with higher weights associated with leaving than entering mean a likely true binding site, relative binding affinities can be determined.

Discussion

The methods described above all make certain limiting assumptions in order to be computationally reasonable, with each method having its own unique advantages and disadvantages. For example, it is completely unreasonable computationally to scan a large library of ligands using molecular dynamics to determine which is best for binding a receptor. On the other hand, molecular dynamics has no peer when it comes to detailed calculations once the potential ligand and binding site have been found. Conversely, the robotics approach is able to predict potential binding sites efficiently, yet cannot give relevant results when it comes to the level of detail that is required in later steps. One reason for this phenomenon is that molecular dynamics performs an incredibly large amount of calculations in a very short period of time to simulate the system as realistically as possible. Recalculations of energetic forces occur on the order of every few femtoseconds with molecular dynamics, whereas the robotics approach can offer relative binding strengths of a ligand with 1000-100,000 times fewer calculations. (Brutlag, personal communication) Even so, biological systems are immensely complex, full of processes that we do not understand, or are impossible to model. Even the best simulation is only that, a simulation.

Many assumptions are made in order for simulations to remain feasible, of which the following are a few: rigidity of receptor, a discrete model to describe conformational flexibility, and the model to predict geometry of protein-ligand interactions. (Rarey, 1996) Only recently has the effect of water molecules in solvent been considered in binding conformations. (Rarey, 1996) Of these limitations, Rarey argues that receptor rigidity is the most troublesome. However, it is nearly impossible to model true flexibility in the protein, as the number of DOF in just a small portion of the protein would overwhelm any computer. Methods that make use of rotamer libraries make a small step in the right direction in this regard, because the observed conformations in the rotamer library can make side chain variation reasonable to compute. This representation is of course much cruder than the continuous freedom of molecular dynamics, but because there are a finite number of conformational possibilities for each side chain, the combinatorial problem of determining the side-chain conformations is at least feasible. The reason this becomes important is that variation in side chains can allow for more energetically stable binding sites, and thus ones that would theoretically be observed more. Also, many proteins and ligands have experimentally been shown to change conformation upon binding, so modeling that is likely to be important. (Havranek, personal communication)

The robotics approach's strength is in computing kinetic properties of binding in a reasonable time. (Singh, 1999) It differs from the others in that it emphasizes the path of the ligand to potential binding sites, not only the final docked position. Therefore many properties of the binding process can be described by the robotics approach that would be impossible with other approaches. For example, Singh observed that an energy barrier seemed to be present around true binding sites, based on results which gave high difficulty weights for the ligand leaving and entering the binding site. It is also possible to localize transition states and various energy barriers that regulate the rate of ligand binding and dissociation (Singh, 1999) – allowing for a more complete picture of the process. Observation such as these would be impossible in a program like AutoDock and may lead to learning more about the mechanisms involved with binding.

Also important to consider is the maturity of each of the techniques. Molecular dynamics-based approaches have been available for decades, with thousands of scientists making improvements along the way based on experimental data. According to Alligner, most modern day force fields have converged, and although there are small differences between them, they are based on mostly arbitrary choices. (Alligner, 1996) Robotics and rotamer techniques have each been around for just a fraction of the time, and are experiencing much more dramatic improvements in applicability and accuracy.

It is important to realize that both "fast and inaccurate" and "slow and accurate" methods are needed for this process. In screening a large ligand library for the best drug candidates, for example, fast and inaccurate methods are most useful for initial testing, while slow and inaccurate methods are more appropriate to use after the list has been narrowed down to a select few. All told, even though we can predict docking with a surprising degree of accuracy, there is much room for improvement.

References

Allinger, N., Kuohsiang, C., Lii, J. *J. Comp. Chem.* 1996, 17, 642-668.

Goodsell, D., Olson, A. "Automated Docking of Substrates to proteins by Simulated Annealing." *Proteins.* 1990, 8, 195-202.

Goodsell, D., Morris, G., Olson, A. "Automated Docking of Flexible Ligands: Applications of AutoDock." *J. Mol. Rec.*, 1996, 9, 1-5.

Kavraki, L.E., Svestka, P., Latombe, J-C., Overmars, M.H. "Probabilistic roadmaps for path planning in high dimensional configuration spaces. *IEEE Tr Robotics and Automation.* 1996, 12(4:566-580)

Kuntz, I. "Structure-Based Strategies for Drug Design and Discovery." *Science.* 1992, 257, 1078-1082.

Koehl, P., Levitt, M. "Improved recognition of native-like protein structures using a familiy of designed sequences." *Proc. Natl. Acad. Sci. USA.* 2002, 99:2, 691-696.

Lengauer, T., Rarey, M. "Computational methods for biomolecular docking." *Curr. Opin. Struct. Biol.*, 1996, 6, 402-406.

Rarey, M., Kramer, B., Lengauer, T., Klebe, G. "A Fast Flexible Docking Method using an Incremental Construction Algorithm." *J. Mol. Biol.* 1996, 261, 470-489.

Rosenfeld, R., Vajda, S., DeLisi, C. "Flexible Docking and Design." *Annu. Rev. Biophys. Biomol. Struct.* 1995, 24, 677-700.

Samudrala, R., Huang, E., Koehl, P., Levitt, M. "Constructing side chains on near-native main chains for *ab initio* protein structure prediction." *Protein Engineering.* 2000, 13, 453-457.

Singh, A., Latombe, J., Brutlag, D. "A Motion Planning Approach to Flexible Ligand Binding." (http://cmgm.stanford.edu/~brutlag/Abstracts/singh98.html)

Website 1: http://www.scripps.edu/pub/olson-web/doc/autodock/

Website 2: http://www.accessexcellence.org/AB/GG/prot_Bindg.html

Website 3: http://www.fccc.edu/research/labs/dunbrack/sidechain.html