A Motion Planning Approach to Flexible Ligand Binding

Amit P. Singh[1], Jean-Claude Latombe[2], Douglas L. Brutlag[3]

[1]Section on Medical Informatics
[2]Department of Computer Science
[3]Department of Biochemistry and Section on Medical Informatics
Stanford University, Stanford CA 94305
[1]apsingh@cmgm.stanford.edu, [2]latombe@cs.stanford.edu, [3]brutlag@stanford.edu

Abstract

Most computational models of protein-ligand interactions consider only the energetics of the final bound state of the complex and do not examine the dynamics of the ligand as it enters the binding site. We have developed a novel technique for studying the dynamics of protein-ligand interactions based on motion planning algorithms from the field of robotics. Our algorithm uses electrostatic and van der Waals potentials to compute the most energetically favorable path between any given initial and goal ligand configurations. We use probabilistic motion planning to sample the distribution of possible paths to a given goal configuration and compute an energy-based "difficulty weight" for each path. By statistically averaging this weight over several randomly generated starting configurations, we compute the relative difficulty of entering and leaving a given binding configuration. This approach yields details of the energy contours around the binding site and can be used to characterize and predict good binding sites. Results from tests with three protein-ligand complexes indicate that our algorithm is able to detect energy barriers around the true binding site that distinguish this site from other predicted low-energy binding sites.

1. Introduction

The process of molecular recognition and binding is central to all biochemical processes and has been studied through a variety of computational techniques. Most of these techniques are based on discrete representations of the binding process and consider only static properties of binding, such as the energy of the final bound complex. In this paper, we present a novel approach to study the dynamics and kinetics of flexible ligand binding using robotic motion planning.

Several algorithms have been developed for studying and predicting protein-protein and protein-ligand interactions [Kuntz et al., 1982; Shoichet and Kuntz, 1993; Leach, 1994; Fisher et al., 1995; Oshiro et al., 1995; Lengauer and Rarey, 1996; Rarey et al., 1996; Sobolev et al., 1996; Gabb et al., 1997; Jones et al., 1997; Lenhof, 1997; Morris et al., 1998]. In the case of protein-protein binding, algorithms have been developed to predict the contact surfaces involved in the interaction and the conformation of the final bound complex. Due to complexity of the binding surfaces, these techniques are computationally intensive and often make the approximation that both molecules are rigid. Computational models of protein-ligand binding face a different but equally challenging set of problems due to the smaller size but greater flexibility of the ligand molecule. One objective of these systems is to scan a protein against a library of small ligands to find those that may bind to the protein and hence form a basis for potentially active drugs. Most computational models of receptor-ligand interactions are based only on the energetics of the ligand in its final bound conformation and do not take into account the dynamics of the ligand as it enters the binding site. These models attempt to compute, for any given receptor, the conformation of the ligand that maximizes a heuristic energy score. This score is usually based on an ad- model of the free energy of binding or the absolute energy of the substrate. Since these techniques only consider discrete instances of the receptor-ligand complex they cannot explicitly measure thermodynamic or kinetic properties of the binding process.

In order to study thermodynamic and kinetic properties of binding, researchers have relied mainly on computationally intensive simulation techniques. These methods attempt to use time averaging or ensemble averaging to calculate properties such as the free energy of binding or the rates of association and dissociation. The two principal simulation techniques used are molecular dynamics and Monte Carlo simulation. Molecular dynamics attempts to simulate the true dynamics of a system using Newton's equations of motion [Anderson, 1980; McCammon, 1987; Haile, 1992; Daggett and Levitt, 1993; Leach and Klein, 1995]. Monte Carlo methods use a randomized approach to generating successive ligand

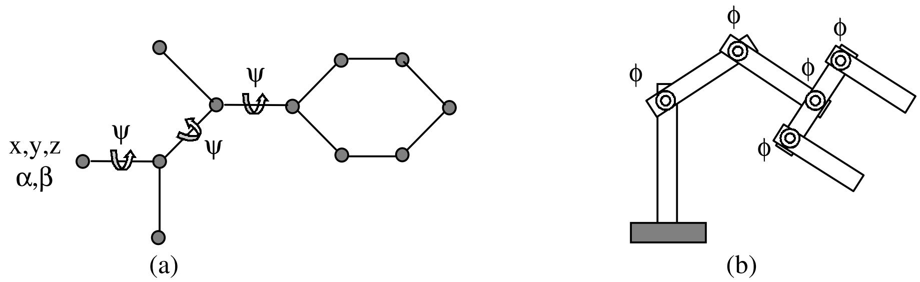


Figure 1. (a) A ligand with 8 degrees of freedom (3 coordinates (x,y,z) and 2 angles (α,β) for the root atom plus one torsional angle (ψ) for each non-terminal atom). (b) A 2-dimensional fixed base articulated robot with 5 degrees of freedom (one rotation angle (ϕ) for each joint).

configurations and compute thermodynamic properties by averaging over all samples that were generated [Rubinstein, 1981; Knegtel et al., 1994; Cummings et al., 1995]. While both molecular dynamics and Monte Carlo methods are theoretically capable of estimating the thermodynamic properties of a receptor-ligand interaction, they require very large amounts of computation time for ligands with many degrees of freedom.

We present a novel approach to studying the dynamics and kinetics of protein-ligand interactions by estimating the motion of the ligand during the process of binding. Our approach attempts to improve the speed and efficiency of simulation methods by using algorithms based on robot motion planning [Latombe, 1991]. In essence, we alleviate the time dependency of molecular dynamics methods and the Markov dependency of Monte Carlo methods by sampling from the space of all possible paths that a ligand may take as it binds to the receptor protein. Hence, instead of simulating the binding process, we effectively guess several possible intermediate configurations of the ligand and obtain a distribution of energetically favorable paths to the binding site (via these intermediate configurations). For each path, we generate a "difficulty weight" that represents the energy barriers that the ligand encounters along the path. For instance, paths that require crossing over large energy barriers are more difficult and are hence given higher weights than those following a descending energy field. By randomly generating several starting and ending configurations we can therefore estimate the average difficulty of entering or leaving different sites on the receptor protein. These numbers can be further used to estimate rates of binding and dissociation (K_{on} and K_{off}). Using motion planning to model the dynamics of the protein-ligand interactions can provide significant insights into the process of binding that would not be possible through traditional static models. For instance, tracking the intermediate states traversed by the ligand could indicate topological regions around the receptor that represent transition states or energy barriers.

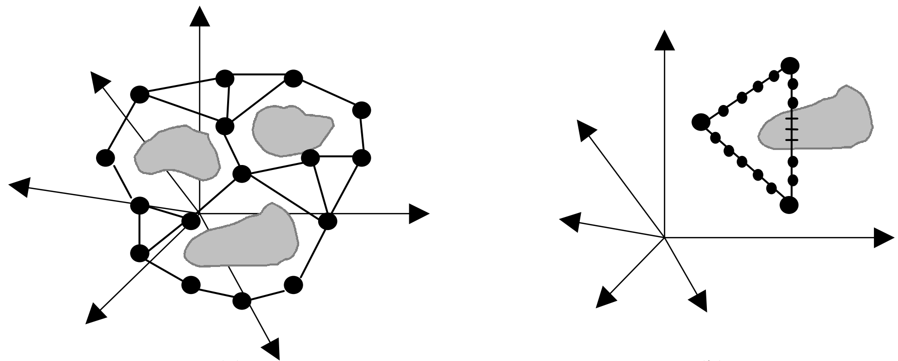
We have tested our algorithm on three protein-ligand complexes. For each complex, we use our technique to examine the average difficulty weight of all paths into the true binding site and compare the numbers obtained with the scores for various other potential binding sites. We also use our randomized motion planning algorithm to predict low-energy binding sites on the surface of the protein.

2. Ligand Modeling

Robot motion planning is applicable to the study of receptor-ligand interactions due to the fact that a flexible ligand can be naturally modeled as an 'articulated robot' (Figure 1). An articulated robot typically consists of several links that can rotate around or translate along one or two axes (joints). We model the ligand as an articulated robot with a free base. Each atomic bond of the ligand molecule maps to a joint of the robot with torsional freedom of motion. Bond angles and bond lengths are kept constant. The root atom, which represents the free base of the robot, is an arbitrarily chosen terminal atom from the ligand. It is given 5 degrees of freedom: 3 to specify its coordinates and 2 to specify the orientation of its only bond. Each additional non-terminal atom requires only a single torsional angle to define the orientation of the rest of the molecule. Bonds involved in a ring are modeled as being completely rigid (i.e., no torsional freedom), which is generally true of most organic molecules. Terminal hydrogen atoms are not explicitly modeled, but are accounted for by increasing the radius of the associated atoms.

3. Motion Planning using Probabilistic Roadmaps

The traditional framework of robot motion planning is based on manipulating a robot through a workspace while avoiding collisions with the obstacles in this space. Our application of motion planning, on the other hand, is aimed at determining potential paths that a robot (or ligand) may naturally take based on the energy distribution of its workspace. Hence, instead of inducing the motion of the robot through actuators, we examine the possible motions of the robot induced by the energy landscape of its immediate environment.



(a) (b)

Figure 2. (a) Schematic representation of a roadmap in a six dimensional configuration space. The irregular shaded regions represent obstacles. (b) Local path planning using discretized configurations along a straight-line path in configuration space. Paths with collisions ("--") are rejected.

Given a starting and ending configuration of a ligand, motion planning algorithms attempt to determine a collision-free path between these two configurations. Paths are usually computed in the configuration space (or C-space) of the robot, which contains one dimension for each degree of freedom of the robot. Motion planning in a high-dimensional configuration space is a difficult task and several techniques have been developed to address this problem [Faverjon and Tournassound, 1990; Ahuactzin et al., 1992; Barraquand and Latombe, 1991; Barraquand and Ferbach, 1994; Kavraki, 1995; Kavraki et al., 1996; Hsu et al, 1997].

Probabilistic Roadmap Planners (PRMs) [Kavraki et al., 1996] are particularly suited for our application since they can efficiently handle robots with many degrees of freedom and can be extended to represent energetic constraints in C-space. A roadmap, as the name implies, is a set of milestones or nodes, each of which is connected to several close neighbors by physically realizable paths. It is represented computationally as an undirected graph and is used to capture the connectivity of the configuration space in which the robot is maneuvering (Figure 2a). The roadmap is constructed by selecting a set of milestones from the C-space of the robot and then connecting each milestone to several of its nearest neighbors by a local path-planning algorithm. Complete representation of obstacles (in our case the receptor protein) in a high-dimensional C-space is computationally impractical. Hence, milestones in a high-dimensional C-space are not selected deterministically and are instead generated by sampling randomly from this space. The local path planner determines whether a path exists between any two nodes by connecting them with a straight line in C-space and testing this path for collisions (Figure 2b). Note that though the local paths are straight lines in C-space, they usually represent complex non-linear motions in physical 3D space. Once the roadmap is constructed, the robot can be maneuvered between two arbitrary initial and final configurations by simply finding, for these configurations, the two nearest milestones on the roadmap to which it can be connected by the local planner. A path between these two milestones can then be computed by performing a graph search on the roadmap.

The fundamental difference between our application of PRMs and prior applications of this technique is our use of energy values in C-space to estimate the naturally induced motion of the ligand. We are interested not only in finding whether a path exists but also whether the path is energetically favorable. Thus, while traditional PRM planners require only a binary test to determine whether a particular point in C-space corresponds to a collision between the robot and the obstacle, we need to compute a continuous energy value for all of these points. Since any given point in C-space corresponds to a particular configuration of the ligand, we compute the total energy at this point by summing the individual energy contributions of all ligand atoms in the corresponding configuration. We use a potential function consisting of electrostatic and van der Waals components to compute the energy of interaction of the ligand with the receptor. The electrostatic potentials are computed using the Poisson-Boltzmann equation, which models both solvent and ionic effects. We use the Delphi program [Sharp and Honig, 1990] to compute an electrostatic potential grid at a resolution of 0.5 Å. The van der Waals potentials are computed on the same grid by calculating, for each grid point, the potential contribution of all receptor atoms within a threshold distance of 10 Å. Different grids are computed to deal with the different radii of ligand atoms. Thus the energy of interaction of the ligand with the receptor is computed by simply looking up the electrostatic and van der Waals potentials at the grid points closest to each of the ligand atom. The internal energy of the ligand (i.e., energy of interaction of the ligand atoms with each other) is computed using standard Coulombic and van der Waals equations.

The following two sections (3.1 and 3.2) describe in detail the two fundamental steps of our energy-based motion planning algorithm, i.e., generating milestones and constructing the roadmap using the local path planner.

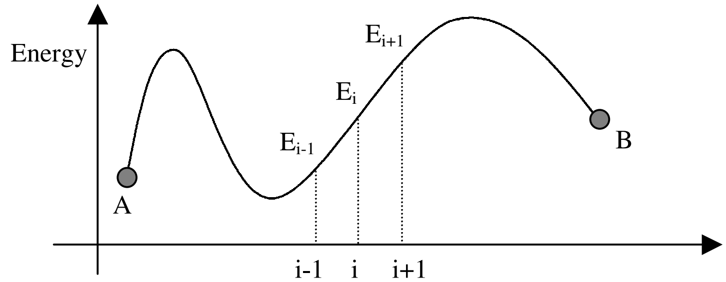


Figure 3. An energy contour for a path in a 1-dimensional configuration space.

Section 3.3 describes the methods used to search the roadmap and compare different potential binding sites. Finally, section 3.4 presents a simple algorithm that uses the previously generated milestones to predict potential binding sites.

3.1. Generating milestones

Our model of the ligand assigns each degree of freedom of the molecule to a separate dimension in C-space. Hence, ligand configurations are generated by simply assigning random values to each coordinate in this high-dimensional space. For each sample that is generated, its total energy is computed and used to determine whether or not the sample will be accepted as a milestone. As described in the previous section, the total energy of the ligand is computed based on electrostatic and van der Waals potentials. Since this term includes the exponentially high van der Waals energies at short inter-atomic distances, it completely models collisions between the ligand and the protein as well as self-collisions of the ligand atoms with each other. In addition to eliminating all points in C-space that involve collisions, we also bias our sampling process to generate more milestones in regions of low energy. Hence, a randomly generated ligand configuration is accepted as a milestone with the following probability:

$$P(\text{accepted}) = \begin{cases} 0 & \text{if } E_{config} > E_{max} \\ \dfrac{E_{max} - E_{config}}{E_{max} - E_{min}} & \text{if } E_{min} \le E_{config} \le E_{max} \\ 1 & \text{if } E_{config} < E_{min} \end{cases}$$

This method of probabilistic collision checking therefore results in a denser sampling of low-energy regions of C-space. For this study, we set E_{max} to 5 kcal/mol and E_{min} to –20 kcal/mol.

3.2 Constructing the roadmap

The probabilistic roadmap is constructed by first obtaining a set of S randomly generated milestones and then connecting pairs of milestones using the local path planner. The following algorithm is used to construct the roadmap from the set of randomly generated milestones (i.e. nodes):

1. For each node i (0 <= i < S)
   - 1.1. Sort all remaining nodes (i.e., with index > i) based on their distance from i
   - 1.2. While the number of edges at node i is less than N
     - 1.2.1. Use the local path planner to connect i to its first un-tested nearest neighbor (i.e., a node to which an edge has not yet been attempted with the local path planner)

The above algorithm results in an undirected graph with approximately N edges at each node. The distance function we used in step 1.1 above is the maximum distance between any two corresponding atoms, though other metrics are also possible. In our implementation we set N to the number of degrees of freedom of the ligand (#dof). In addition, if after M attempts with the local path planner a node does not form at least N edges, it is transferred to the connectivity enhancement step (described at the end of this section), after which the roadmap construction proceeds with the next milestone in the set S. This is done to place an upper bound on the number of times the local path planner is executed, since it is the most computationally intensive step of the algorithm. Setting M to about 3 to 5 times N usually results in acceptable execution times.

The local path planner we use is a modification of the one described in Kavraki et al., 1996. The planner connects two given milestones by a straight line in C-space and determines whether this straight-line path is energetically feasible (i.e., collision free). The path is tested by discretizing the line segment into a series of consecutive configurations, each separated by a maximum distance ε. The path is accepted only if all the configurations along the path have energy less than a maximum energy threshold (usually 5-10 kcal/mol). In our implementation, we discretize the straight line path such that the maximum distance between any two corresponding atoms in two adjacent configurations is less than 1 Å (i.e., ε = 1 Å).

For each accepted path, our local path planner also computes a dweight representing the energetic favorability of the path. This weight reflects the difficulty of traversing the path and is hence higher for paths that require crossing over large energy barriers. We use the energy of each discretized configuration along the path to determine the overall probability of traversing the path in a particular direction. Figure 3 is a simple energy contour for a straight-line path between two nodes in a 1-dimensional configuration space. The line segment between nodes A and B is discretized into consecutive configurations that are less than ε units apart. For any three successive discretized configurations along the straight line path (i-1, i, and i+1 with energies E_{i-1}, E_i, and E_{i+1}) we use the following equation to determine the probability of moving from configuration i to i+1:

$$P(\text{i to i}+1) = \frac{e^{-(E_{i+1} - E_i)/kT}}{e^{-(E_{i+1} - E_i)/kT} + e^{-(E_{i-1} - E_i)/kT}}$$

Using the above equation we compute the total weight of the path between A and B as:

$$\text{Weight of local path} = \sum_i -\log\left[P(\text{i to i}+1)\right]$$

Note that the path weight is not the same in both directions, though both weights can be computed simultaneously.

We use this paradigm of motion along a single dimension to compute difficulty weights for all local paths in our roadmap. Though each discretized configuration along the line segment is free to move in several dimensions along the path, we limit its motion along a single dimension (i.e.,. either forward or backward) along the path. This is analogous to placing infinitely high energy barriers on either side of the path thus forcing only linear motion along the bottom of this artificial energy valley. This approximation, though seemingly severe, is mitigated by the fact that all paths between nodes are short and of similar length. Since the approximation becomes worse as the length of the path increases, minimizing the path lengths by increasing the number of milestones and linking them only after all samples have been created, reduces the over-all error. In addition, since this model does not give adequately high weights to longer paths, we compensate for this difference by adding a weighted path length to the total weight of each edge.

Since the surface of the receptor protein is highly convoluted with many narrow cavities, milestones generated close to this surface are generally difficult to connect to the roadmap. Therefore, there are often several nodes in the graph that have less than N edges connecting them to other milestones in the roadmap (where N = #dof). To increase the connectivity of the roadmap in these regions, we added a connectivity enhancement phase to the roadmap construction algorithm. For each milestone j with less than N edges, this enhancement algorithm creates additional nodes close to this milestone by using the paths that failed due to the presence of a high-energy configuration along the path. These new nodes are created just before the high-energy configuration is encountered and are therefore guaranteed to connect to milestone j. The extra nodes that are generated are added to the set of milestones, thus allowing all remaining nodes to be connected to these newly created milestones.

3.3 Searching the roadmap

The roadmap constructed as above represents a distribution of plausible paths of the ligand through the space surrounding the receptor protein. The sum of the local path weights between any one node and all other remaining nodes of the roadmap reflect the kinetic and dynamic properties of the motion between these nodes.

For any randomly generated initial and goal configurations, we compute the minimum-weight path between them by first finding the milestones in the roadmap that can be most easily reached from these two arbitrary configurations (i.e., the milestones that are closest to and have the lowest weight to these configurations). A graph search is then performed to find the minimum-weight path between these two milestones in the roadmap. Since two weights are recorded for each edge representing the two opposite directions of motion, the minimum-weight path in either direction can be computed. The graph search is performed using Dijkstra's single-source shortest-path algorithm. This algorithm can either be terminated as soon as the end node is reached or it can be continued until all nodes have been discovered. The latter allows us to compute the distribution of all paths entering or leaving a given node. By computing the minimum weight for all of these paths, we can obtain an estimate for the average difficulty of all paths entering or leaving a given node. These two numbers can be correlated with the kinetic rates of binding (K_{on}) and dissociation (K_{off}) for this node. In this paper, we use these average weights to estimate the energy barriers around the binding site and to distinguish the true binding site from other predicted low-energy active sites.

3.4 Predicting binding sites

Though our energy based motion planning technique can be effectively used to examine the kinetics and dynamics of ligand binding, often the true binding site on the receptor is not known. In these cases, computational tools are required to predict potential binding sites. We have implemented a simple modification to the milestone generation technique to predict potential binding sites by over-sampling regions of low energy near the protein surface. The algorithm first sorts the initial set of randomly generated nodes in order of increasing energy. Next, for each of the P lowest energy nodes, Q extra nodes are created around it by sampling a region of configuration space close to this initial node. A new minimum energy node is then selected from among the Q extra samples and the process is iterated R times. The initial set of P low-energy nodes are selected so that their centers of mass are at least 5 Å apart. This process results in P distinct regions of C-space that are heavily over-sampled. The number of extra samples generated in each of the P regions is Q*R.

Table 1. Execution times and number of connected components for each test case.

	#dof	Sampling time	Linking time	Final nodes	#connected components
1ldm	7	9 sec	57 sec	6129	2
4ts1	9	27 sec	4 min 13 sec	6530	4
1stp	11	39 sec	4 min 43 sec	6635	5

The algorithm reports each of these P regions, as well as the lowest energy milestones contained within them, as potential active sites.

4. Results

We have tested our path planning system on the following three protein-ligand complexes:

1. PDB ID: 1ldm
   Receptor: Lactate Dehydrogenase (2386 atoms, 309 residues)
   Ligand: Oxamate (6 atoms, 7 degrees of freedom)

2. PDB ID: 4ts1
   Receptor: Mutant of tyrosyl-transfer-RNA synthetase (2423 atoms, 319 residues)
   Ligand: L-leucyl-hydroxylamine (13 atoms, 9 degrees of freedom)

3. PDB ID: 1stp
   Receptor: Streptavidin (901 atoms, 121 residues)
   Ligand: Biotin (16 atoms, 11 degrees of freedom)

For each of the above cases, we obtained the 3D coordinates of all atoms in the complex from the protein data bank (PDB) [Bernstein et al., 1977]. The PDB file contained the coordinates of the protein atoms as well as the ligand atoms in their bound state. Using these cases, we tested our algorithm on the following three fronts:

1. *Basic functionality:* Can the algorithm find a path to the true binding configuration? Do the number of connected components (isolated graphs that do not connect) found correspond to what can be reasonably expected?
2. *Characterizing the binding site:* Can the algorithm distinguish the true binding configuration from other potential binding configurations?
3. *Predicting the binding site:* Can the algorithm predict the true binding site using the iterative sampling technique described in section 3.4?

For each of the above tests we set the number of initial samples to 4000. From these 4000 nodes, 20 nodes with lowest energy were selected to seed the iterative sampling algorithm. For each of these 20 nodes, the iterative sampling step created 100 new samples, thus yielding a total of 2000 extra nodes (P = 20, Q = 10, R = 10 in section 3.4). The total number of nodes before the roadmap linking phase was therefore 6000. The number of nodes at the end of the linking phase was generally higher than 6000 because of the extra nodes generated by the connectivity enhancement algorithm.

Since each milestone is linked to several of its close neighbors and most of the energetic information is contained in the difficulty-weights associated with these links, the number of nodes required to construct a useful roadmap is not very large. We therefore use 4000 initial nodes to broadly cover the C-space of the ligand. These nodes are likely to reside within or close to a local energy minimum since they are probabilistically accepted as milestones based on their energy (the ratio of accepted to rejected samples is about 1:10). The low energy regions close to the surface of the protein are examined in more detail by the iterative sampling step which generates 2000 extra samples in these local regions of C-space. The extra samples increase the accuracy of the roadmap in these regions of greater energetic variability and also improve the likelihood of finding the true binding configuration (if it is located within one of these regions) or other potential binding configurations.

4.1 Basic functionality

For each of our three test cases, the algorithm was able to connect the configuration space using 2 to 5 connected components. Since voids or narrow cavities do occasionally occur within a protein structure or close to its surface, it is likely that some of the randomly generated nodes will lie within these regions thus yielding more than one connected component. In our tests, more than 98% of the total nodes in the graph were contained within a single connected component with only 2% of the nodes distributed among the remaining few connected components. For all three cases the true binding configuration was contained within the largest connected component thus enabling paths from virtually all points in C-space to this goal configuration. Table 1 lists the average execution times of our algorithm as well as the number of nodes and connected components generated. These tests were performed on a Silicon Graphics Octane with a 195 MHz MIPS R10000 processor. The number of final nodes is higher than the initial nodes due to the connectivity enhancement step described above.

4.2 Characterizing the binding site

We tested the biochemical validity of our algorithm by examining whether or not it was able to distinguish the true binding site from other low-energy sites on the protein

Table 2a. Average path weights for 1ldm.

Row number	RMSD from true binding configuration (Å)	Configuration energy (kcal/mol)	Avg path weight entering configuration	Avg path weight leaving configuration
0	**0.00**	**-11.79**	**112.98**	**134.54**
1	31.04	-13.65	85.07	109.94
2	27.49	-12.66	90.48	111.98
3	**1.73**	**-11.72**	**113.81**	**137.28**
4	28.99	-11.54	85.32	105.19
5	24.67	-11.31	86.26	103.95
6	29.84	-11.27	86.49	107.53
7	29.32	-11.04	85.24	104.64
8	27.07	-10.96	81.70	102.28
9	31.00	-10.13	87.69	104.50
10	28.24	-9.97	86.36	98.89

Table 2b. Average path weights for 4ts1.

Row number	RMSD from true binding configuration (Å)	Configuration energy (kcal/mol)	Avg path weight entering configuration	Avg path weight leaving configuration
0	**0.00**	**-19.44**	**130.73**	**173.76**
1	**1.91**	**-20.31**	**128.61**	**166.73**
2	21.59	-15.92	105.65	118.72
3	15.16	-14.53	109.82	129.15
4	23.55	-14.39	111.87	134.96
5	20.59	-14.30	114.13	133.87
6	22.19	-13.97	113.84	135.90
7	24.62	-12.89	118.82	138.15
8	19.13	-12.74	115.45	136.72
9	17.05	-12.31	120.24	142.72
10	36.81	-11.81	115.48	131.98

Table 2c. Average path weights for 1stp.

Row number	RMSD from true binding configuration (Å)	Configuration energy (kcal/mol)	Avg path weight entering configuration	Avg path weight leaving configuration
0	**0.00**	**-15.06**	**110.80**	**146.87**
1	21.76	-15.79	80.78	108.42
2	27.14	-12.83	96.29	117.67
3	18.59	-12.82	85.84	101.24
4	23.52	-11.45	96.45	122.01
5	13.67	-11.36	86.51	106.05
6	15.18	-10.79	88.22	96.89
7	13.93	-10.68	95.14	116.92
8	14.63	-10.42	85.61	105.16
9	24.64	-9.96	85.71	105.17
10	20.43	-9.87	83.81	102.54

surface. The two attributes we used to distinguish between true and predicted binding configurations were the absolute energy of the ligand and the average weight of all paths entering and leaving the configuration.

Tables 2a, 2b, and 2c list the results of our comparison of absolute energy and average path weights for the three test cases. The first row (row 0) in each table shows energy and average weights for the true binding configuration. The remaining 10 rows in each table show the average weights for the 10 lowest energy configurations found by our iterative sampling strategy. Note that each of these 10 configurations belongs to a

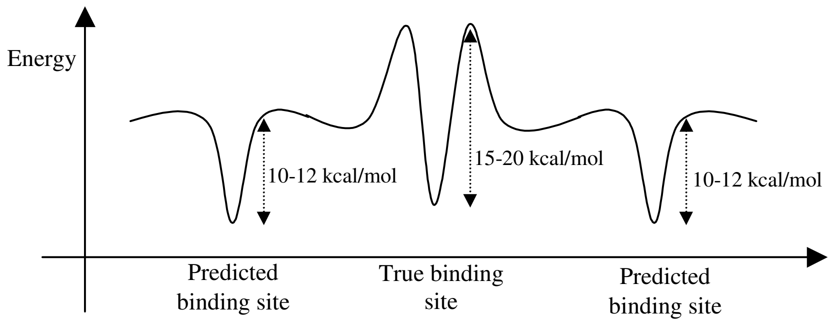


Figure 4. A schematic representation of the energy barrier around the true binding site.

different low energy cluster generated by the iterative sampling step. They were obtained by selecting the lowest energy configuration from each cluster and sorting them according to their energy. Each row therefore represents the lowest energy configuration within a cluster.

We observed that the absolute energy of the ligand was not a strong discriminating factor between the true binding site and other predicted low-energy sites. In two of our three test cases (1ldm and 1stp) the algorithm was able to find ligand configurations outside the true binding site with energies equal to or even slightly lower than the energy of the ligand in its true binding configuration. For instance, rows 1 and 2 of Table 2a and row 1 of Table 2c represent configurations found by the iterative sampling technique that are distant from the true binding site and have energies lower that the true binding configuration (row 3 in Table 2a and row 1 in Table 2b are not included in this list since they have low RMSD values and therefore lie in the same site as the true binding configuration). It is possible that some of these configurations yield energy values lower than the true binding configuration because of the approximations of our grid based energy model.

The second criterion we examined was the average difficulty weight of all paths entering and leaving a given configuration. Using this criterion, the algorithm was able to clearly distinguish between the true binding configuration and other predicted low-energy binding sites. We observed that the average weight of all paths entering and leaving the true binding configuration was significantly higher than the weights for all other low-energy configurations (including those with energy lower than the true binding configuration). Therefore, it was significantly more difficult for the ligand to *leave* the true binding site as compared to the predicted low-energy sites. Furthermore, it was correspondingly more difficult for the ligand to *enter* the true binding site as compared to other low-energy sites. While this latter result may seem counter-intuitive at first, we believe that it indicates the presence of a distinct energy barrier around the true binding configuration that traps the ligand within the site (see Discussion section below). Note that row 3 in Table 2a and row 1 in Table 2b also show path weights that are in the same range as the true binding configuration. This is due to the fact that the ligand configurations corresponding to these rows are in fact close to the true binding configuration (hence the low RMSD values) and lie within the same binding site. This indicates that configurations close to the true binding configuration (i.e., within the same binding pocket) also share the characteristic feature of having a high energy barrier around them.

4.3 Predicting the binding site

The final test of our algorithm examined whether the iterative sampling technique (section 3.4) is able to find the true binding configuration for each of the three protein-ligand complexes. The criterion used by the iterative sampling algorithm to search for the true binding site is the total energy of the configuration. As observed in the previous section, the total configuration energy is not the optimal criterion for identifying the true binding site. Hence, though the prediction algorithm was able to find a configuration close to the true binding configuration in two cases (1ldm and 4ts1), only one of these was correctly ranked at the top of the list of predictions (4ts1).

Rows 1-10 of Tables 2a, 2b, and 2c represent the top 10 predictions from distinct clusters, ranked according to their absolute energy. Note that since these clusters are chosen to be about 5 Å apart and only one configuration from each cluster is shown, each table can contain at most one correct prediction of the true binding configuration. For the case of 1ldm (Table 2a) the correct prediction has an RMSD of 1.73 Å and is ranked 3rd. For the case of 4ts1 (Table 2b) the correct prediction has an RMSD of 1.91 Å and is ranked 1st. Note that for both these cases, the correct predictions yield average path weights that are similar to the true binding configuration and hence significantly higher than all other predictions. Therefore, if the average path weights were used as the primary sorting criterion the algorithm would be able to correctly place the true binding site at the top of the list of predictions for both these cases.

The iterative sampling algorithm was not able to find the true binding site in the case of 1stp. This may be due to the fact that the ligand in this complex is bound in a very tight pocket with a narrow opening on the surface of the protein. Hence, none of the 20 lowest energy nodes from among the 4000 randomly generated nodes were close enough to the true binding site to allow the iterative sampling technique to converge on to this site.

5. Discussion

We have developed an algorithm based on robot motion planning to study the dynamics of a ligand as it binds to the receptor protein. We use motion planning to alleviate the Markov dependency of Monte Carlo simulation methods, thus allowing the computation of kinetic properties of binding in reasonable time. Our technique samples the space of all possible paths to a given configuration and computes a statistical measure of the relative difficulty of entering or leaving this configuration.

Tests of our algorithm on three different protein-ligand complexes yielded surprising results. We found that the average weight of all paths entering or leaving the true binding site was significantly greater than the weights for all other potential binding sites on the protein. Though the higher weights for *leaving* the true binding site were expected, our experiments also found that it was significantly more difficult for the ligand to *enter* the true binding site as compared to other low-energy sites. While these results may seem counter-intuitive at first, we believe that they indicate the presence of an energy barrier around the true binding site. Figure 4 shows a schematic of a possible energy contour that could yield results similar to those we obtained. As seen in this diagram, the true binding site is surrounded by a distinct energy barrier that significantly increases the difficulty of leaving this site. At the same time the energy barrier also makes it more difficult for the ligand to enter the true binding site as compared to other sites of equally low or lower energy. Hence, though it is relatively more difficult for the ligand to enter the true binding site, once it does enter it is very difficult for it to leave this site. The ligand is thus trapped in the binding site by the energy barrier. We believe that the high difficulty weight for leaving the binding site indicates a very low rate of dissociation (K_{off}), which dominates the standard free energy of the reaction. The results in Tables 2a, 2b, and 2c also show that the average weight of paths entering the true binding site is of the same order as the weight of paths leaving the predicted sites. Therefore, the difficulty of entering the true binding site is approximately equal to the difficulty of leaving the predicted sites.

Our algorithm for generating milestones for the roadmap biases the sampling towards regions of low energy and iteratively generates more samples around the lowest energy configurations. We use this process of iterative over-sampling around the current minimum-energy ligand configuration to predict potential binding sites on the protein surface. We found that in two of the three test cases this simple algorithm was able to correctly find the true binding site among the top 3 predictions. By sorting the predictions according to the path weights instead of energy, the algorithm was able to correctly rank the true binding site at the top of the list in both these cases.

We believe that computing the average weight of all paths into and out of a particular site provides a valid statistical measure for the relative difficulty of entering or leaving the site. Though the absolute value of these difficulty weights may not be biochemically significant, their relative values can be used to predict the relative rates of binding and dissociation for different binding sites. Our results also show that the energy barriers found by our algorithm are a unique feature of good binding sites and can be used as a criterion for distinguishing the true binding site from other predicted sites. There are several potential benefits of using our motion-planning model of ligand binding as compared to other static models. For instance, examining the distribution of paths into and out of the binding site can be used to help localize transition states and other energy barriers that regulate the rate of ligand binding and dissociation. We plan to improve the functionality of our technique by developing tools to partition the space around a given binding site into shells and compute the dynamics of ligand motion between each of these shells. This approach will allow us to study both distant electrostatic channeling effects [Tan, 1993] as well as detailed interactions within the binding site. Since receptor flexibility is an important factor in regulating ligand binding, we are also developing methods to include rotomer flexibility into the motion-planning model.

Acknowledgements

This work was supported by the NLM R01 LM05716-01 grant. Jean-Claude Latombe is partially supported by ARO-MURI grant DAAH04-96-1-007. The authors thank Lydia Kavraki, Pehr Harbury, and Dan Hershlag for discussions and suggestions during this project.

References

Ahuactzin, J.M., Talbi, E-G., Bessiere, P. and Mazer, E. (1992). Using genetic algorithms for robot motion planning. *10th Europ Conf Artificial Intelligence*, 671-675.

Anderson H. C. (1980) Molecular dynamics simulations at constant pressure and/or temperature. *J Chem Phys*, 72:2384-93.

Barraquand , J. and Ferbach, P. (1994). Path planning through variational dynamic programming. *Proc IEEE Int Conf Robotics and Automation*, 1839-1846.

Barraquand, J. and Latombe, J-C, (1991). Robot motion planning: a distributed representation approach. *Int J Robotics Research*, 10:623-649.

Bernstein, F.C., Koetzle, T.F., Williams G.J.B., Meyer E.F. Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., and Tasumi, M. 1977. The Protein Data Bank: A computer based archival file for macromolecular structure. *J. Mol. Biol.* 112:535-542.

Cummings, M. D., Hart, T. N. and Read, R. J. (1995). Monte Carlo docking with ubiquitin. *Prot. Sci*, 4(5), 885-99.

Daggett, V. and Levitt, M. (1993). Realistic simulations of native-protein dynamics in solution and beyond. *Annu Rev Biophys Biomol Struct,* 22:353-80.

Faverjon B. and Tournassoud P. (1990). A practical approach to motion planning for manipulators with many degrees of freedom. Robotics Research 5, H. Miura and S. Arimoto (Eds.), 65-73, MIT Press.

Fischer, D., Lin, S. L., Wolfson, H. L. and Nussinov, R. (1995). A geometry-based suite of molecular docking processes. *J Mol Biol*, 248(2), 459-77.

Gabb, H. A., Jackson, R. M. and Sternberg, M. J. (1997). Modeling protein docking using shape complementarity, electrostatics and biochemical information. *J Mol Biol*, 272(1), 106-20.

Haile J.M. (1992) Molecular dynamics simulation: elementary methods. New York, Wiley.

Hsu, D., Latombe, J-C., and Motwani, R. (1997). Path planning in expansive configuration spaces. *Proc. IEEE Int Conf Robotics and Automation*, 2719-2726.

Jones, G., Willett, P., Glen, R. C., Leach, A. R. and Taylor, R. (1997). Development and validation of a genetic algorithm for flexible docking. *J Mol Biol*, 267(3), 727-48.

Kavraki, L. (1995). Random networks in configuration space for fast path planning. Ph.D. Thesis, Comp. Sc. Dept., Stanford Univ., Stanford, CA.

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

Knegtel, R. M., Boelens, R. and Kaptein, R. (1994). Monte Carlo docking of protein-DNA complexes: incorporation of DNA flexibility and experimental data. *Protein Eng*, 7(6), 761-7.

Latombe, J-C. (1991). Robot Motion Planning. Kluwer Academic Publishers, Boston, MA.

Leach, A. R. (1994). Ligand docking to proteins with discrete side-chain flexibility. *J Mol Biol*, 235(1), 345-56.

Leach A. R. and Klein, T. E. (1995). A molecular dynamics study of the inhibitors of dihydrofolate reductase by a phynyl triazine. *J Comp Chem*, 16:1378-93.

Lengauer, T. and Rarey, M. (1996). Computational methods for biomolecular docking. *Cur Op Str Biol*, 6:402-406.

Lenhof, H. (1997). New contact measures for the protein docking problem. *Proceedings of RECOMB 97*: 182-191.

McCammon, J.A. amd Harvey, S.C. (1987). Dynamics of proteins and nucleic acids. Cambridge Univeristy Press.

Morris, G. M., Goodsell, D. S., Halliday, R.S., Huey, R., Hart, W. E., Belew, R. K. and Olson, A. J. (1998). Automated Docking Using a Lamarckian Genetic Algorithm and and Empirical Binding Free Energy Function. *J Comp Chem*, 19:1639-1662.

Oshiro, C. M., Kuntz, I. D. and Dixon, J. S. (1995). Flexible ligand docking using a genetic algorithm. *J Comput Aided Mol Des*, 9(2), 113-30.

Rarey, M., Kramer, B., Lengauer, T. and Klebe, G. (1996). A fast flexible docking method using an incremental construction algorithm. *J Mol Biol*, 261(3), 470-89.

Rubinstein, R. Y. (1981). Simulation and monte carlo methods. New York, Wiley.

Sharp, K. and Honig, B. (1990). Electrostatic interactios in macromolecules: theory and applications. *Ann Rev Biophys Chem*, 19:301-32.

Shoichet, B.K. and Kuntz, I.D. (1993). Matching chemistry and shape in molecular docking. *Prot Eng*, 6(7) 723-32.

Sobolev, V., Wade, R. C., Vriend, G. and Edelman, M. (1996). Molecular docking using surface complementarity. *Proteins*, 25(1), 120-9.

Tan, R.C., Thanh, N.T., and McCammon, A.J. (1993). Acetylcholineesterase: electrostatic steering increases the rate of ligand binding. *Biochemistry*, 32:401-3.