On Flexible Docking Using Expansive Search

—

Preliminary Results

Mark Moll
mmoll@cs.rice.edu

David Schwarz
dschwarz@cs.rice.edu

Allison Heath
aheath@cs.rice.edu

Lydia E. Kavraki
kavraki@cs.rice.edu

December 31, 2004

Abstract

The activity of most drugs is regulated by the binding of one molecule (the ligand) to a pocket of another, usually larger, molecule, which is commonly a protein. This report describes a new approach to creating low-energy structures of flexible proteins to which ligands can be docked. The flexibility of molecules is encoded with thousands of parameters making the search for valid complexes a formidable problem. Our method takes into account the flexibility of the protein as this can be encoded by its major modes of motion. The output of the program consists of low-energy protein conformations that can then be docked with a ligand using a traditional docking program. We employ a robotics-based approach for exploring the conformational space of the protein. Our long term goal is to develop an efficient, accurate, and automated algorithm that will be used to screen large databases of molecules for novel therapeutics.

Contents

1 Introduction

The ability to predict the bound conformations and interaction energy between small organic molecules and biomacromolecules such as proteins and DNA is of extreme physiological and pharmacological importance. For example, the activity of most drugs is regulated by the binding of one molecule (the ligand) to the pocket of another, usually larger, molecule, which is commonly a protein. Figure 1 shows HIV-1 protease, a molecule which is heavily involved in the replication of the HIV virus. If this molecule is blocked by a ligand, as shown in the figure, the virus can not create mature copies of itself. As another example consider chemotherapy drugs which are designed to selectively bind to specific proteins involved in cell duplication, and stop the proliferation of tumor cells by inhibiting their function. Estrogen proteins trigger the proliferation of breast cells upon binding to a ligand (estrogen). Most of the existing breast cancer drugs consist of small molecules binding to the estrogen protein and blocking its signaling function.

There has been a considerable effort from both academia and industry to develop computational methods that can be used to determine the affinity with which a ligand will bind a target protein. These methods are typically referred to as docking methods and their output is the three dimensional structure of a protein-
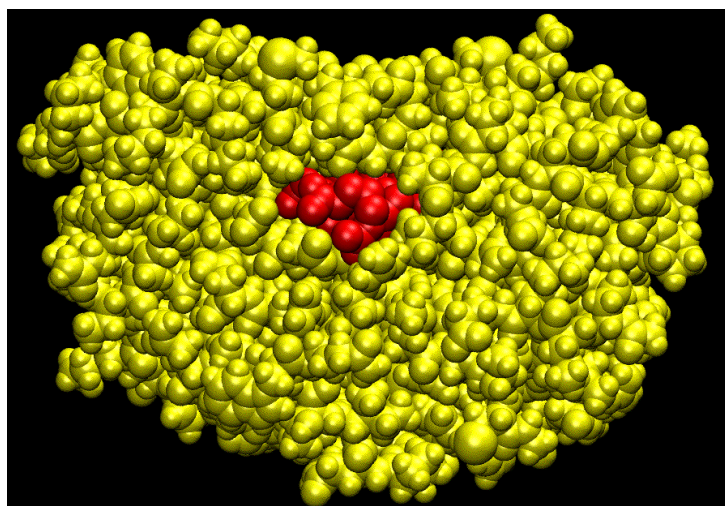

Figure 1: HIV-1 protease, a molecule involved in the replication of the HIV virus. Each atom is represented by a sphere with the appropriate Van der Waals radius. The ligand is shown in red.

ligand complex as would be determined experimentally using X-ray crystallography (Rhodes, 1993) or Nuclear Magnetic Resonance (NMR) (Wüthrich, 1986) methods. Effective docking methods could be used to scan databases of potential ligands and can shorten the drug design cycle (currently 10 years) and reduce its cost (currently 800 million dollars) by generating better leads faster.

One of the first paradigms for the docking problem was the so-called lock-and-key model (Fischer, 1894). Within this paradigms the docking problem reduces to a shape complementarity problem: for a given shape of the active site of a protein we want to find a ligand that can take on a matching shape. Eventually, it was observed that in some cases protein flexibility cannot be ignored. Koshland (1958) proposed the induced-fit model. In this model the protein and the ligand adapt to each other during the binding process. A more modern view on protein flexibility is that the binding process selects a stable conformation from an ensemble of metastable states (Bursavich and Rich, 2002; Ma et al., 1999, 2002, 2001). Some examples of systems that exhibit significant flexibility include streptavidin (Weber et al., 1989), HIV-1 protease (Wlodawer et al., 1989), DHFR (Bystroff and Kraut, 1991), aldose reductase (Urzhumtsev et al., 1997), and maltose binding protein (Duan and Quiocho, 2002).

Several approaches have been proposed for modeling the flexibility of a protein for the analysis of biomolecular interactions. They can be roughly divided into five different categories: (a) the use of soft proteins which relax energetic penalties due to steric clashes, e.g., (Jiang and Kim, 1991), (Schnecke et al., 1998), (Apostolakis et al., 1998), (b) the selection of a few critical degrees of freedom in the protein binding site, e.g., (Leach, 1994), (Leach and Lemon, 1998), (c) the use of multiple protein conformations either individually or by combining them using an averaging scheme, e.g., (Pang and Kozikowski, 1994), (Knegtel et al., 1997), (Sudbeck et al., 1998), (d) the use of modified molecular dynamics methods, e.g., (Di Nola et al., 1994; Mangoni et al., 1999), (Nakajima et al., 1997a,b), and (e) the use of collective degrees of freedom as a new basis of representation for protein flexibility, e.g., (Levy and Karplus, 1979), (Levitt et al., 1985), (García, 1992), (Teodoro et al., 2003). The latter approach is gaining increasing popularity.

Collective degrees of freedom can be determined using different methods. The simplest method of them is the calculation of normal modes for the protein (Levy and Karplus, 1979; Go et al., 1983; Levitt et al., 1985). Normal modes are simple harmonic oscillations about a local energy minimum which depends on the structure of the protein and the energy function. Normal modes assume a purely harmonic energy function and by considering that the protein is at an energy minimum, its flexibility can be represented by using the low frequency normal modes

as degrees of freedom for the system. Zacharias and Sklenar (1999) applied a method similar to normal mode analysis to derive a series of harmonic modes that were used to account for protein flexibility in the binding of a small ligand to DNA. In practice this reduced the number of degrees of freedom of the DNA molecule from 822 (3×276 atoms $- 6$) to approximately 5 to 40. Keserû and Kolossváry (2001) also used a normal mode based model (Kolossváry and Guida, 1999; Kolossváry and Keserû, 2001) to study inhibitor binding to HIV integrase.

An alternative and more widely used method of calculating collective degrees of freedom for macromolecules is the use of dimensional reduction methods. Such methods do not assume a purely harmonic energy function and have thus wider applicability. Among them, a popular choice is principal component analysis (PCA) (Jolliffe, 1986). This method was first applied by García (1992) in order to identify high-amplitude modes of fluctuations in macromolecular dynamics simulations. It has also been used to identify and study protein conformational substates (Romo et al., 1995; Caves et al., 1998; Kitao and Go, 1999), as a possible method to extend the time-scale of molecular dynamics simulations (Amadei and Berendsen, 1993; Amadei et al., 1996; Abseher and Nilges, 2000) and as a method to perform conformational sampling (Abseher and Nilges, 2000; de Groot et al., 1996b,a). The most significant principal components have a direct physical interpretation. They correspond to a concerted motion of the protein where all the atoms move in specific spatial directions and with fixed ratios in overall displacement. Recently, we have presented a protocol (Teodoro et al., 2003) to derive a reduced basis representation of protein flexibility which can be used to reduce the complexity of modeling protein/ligand interactions. By considering only the most significant principal components as the valuable degrees of freedom of the system, it is possible to reduce an initial search space of thousands of degrees of freedom to less than fifty. This is achievable because the fifty most significant principal components usually account for 80–90% of the overall conformational variance of the system. For several systems even the use of the first ten modes offers significant advantages (Teodoro, 2003).

The development of effective docking methods is inhibited by two major considerations. One is the existence of scoring functions that can identify complexes that minimize free energy, a measure of the affinity with which two molecules interact. A second consideration is the huge combinatorial complexity of the docking problem. Both the protein and the ligand are flexible and when they interact they change their shape to produce a minimum free energy 'perfect' fit. The flexibility of a protein is encoded with a few hundred to a few thousand parameters. So docking involves searches in combined conformational spaces of very high

dimension. This paper addresses the combinatorial aspect of the docking problem. We define a reduced basis for the representation of the protein using the main modes of motion of the protein (García, 1992). We use a robotics-inspired search method based on Expansive Space Trees (Hsu et al., 1999) to explore the conformational space of the protein. In this report *we ignore the conformational search for the ligand.* It is assumed that the best-ranked conformations produced by our search are passed to a traditional docking program which will try to dock the ligand to each of the conformations. This can be done relatively quickly; *the hard problem is to quickly find distinct low-energy conformations of the protein.* This problem is the focus of this report.

The outline for the rest of this report is as follows. In the next section we describe how to obtain a compact representation of a protein's flexibility using the main modes of motion. Section 3 describes a docking algorithm for the efficient exploring the conformational space described by the main modes of motion. In section 4 we give several different algorithm-independent performance evaluation criteria that are used in section 5 to describe the performance of our algorithm on some test cases. In section 6 we discuss the contributions of this report and outline directions for future research.

2 Preliminaries

2.1 Generation of Molecular Trajectories

Given an initial conformation of a protein and a temperature, a molecular dynamics program like NAMD (Kalé et al., 1999) or Amber (Cornell et al., 1995) can simulate the motion of the protein using Newton's second law, or the equation of motion, $F = m\boldsymbol{a}$, where F is the force exerted on the particle, m is its mass and $\boldsymbol{a}$ is its acceleration. From a knowledge of the force on each atom (as this is specific by the potential energy of the molecule), it is possible to determine the acceleration of each atom in the system. Integration of the equations of motion then yields a trajectory that describes the positions, velocities and accelerations of the particles as they vary with time. Unfortunately the integration step is limited by the highest-frequency motion that must be simulated since fast vibrations imply rapidly changing velocities and accelerations and is typically in the order of a femtosecond. As a result, molecular dynamics simulations are time consuming and computationally expensive. From a simulated trajectory we would like to compute the most accurate model possible of the molecule's flexibility.

2.2 Major Mode Analysis

The main modes of motion of a protein can be determined from a molecular dynamics simulation. Suppose we have a trajectory containing m conformations for a molecule with n atoms. Let A be a $(3n) \times m$ matrix containing the atom displacements for each atom in each conformation. The displacements are measured with respect to the average conformation. The main modes of motion (or principal components) can then be computed using the Singular Value Decomposition (SVD). The SVD of matrix A is defined as

$$A = U \Sigma V^T, \tag{1}$$

where U and V are orthonormal matrices and Σ is a nonnegative diagonal matrix whose diagonal elements are the singular values of A in decreasing order. The columns of matrices U and V are called the left and right singular vectors, respectively. The left singular vectors corresponding to the largest singular values reflect the major modes of motion.

Care should be taken to align each conformation with a reference conformation to remove any variation due to translations and rotations: we are only interested in how the shape of the protein changes. The matrix A therefore contains the displacements after each conformation has been aligned with the starting conformation. Two conformations can be aligned using SVD (Golub and Loan, 1996, sec. 12.4.1). First, the conformations are translated so that their geometric centers are at the origin. Next, we find a rotation that minimizes the RMSD between two conformations. The RMSD is a distance measure, defined as the square root of the average squared distance between corresponding atoms. Let C_1 and C_2 be two $n \times 3$ matrices containing the Cartesian coordinates of two conformations. Then the orthogonal matrix R that minimizes the Frobenius norm $\|C_1 - C_2 R\|_F$ is given by $R = U V^T$, where U and V are matrices containing the left and right singular vectors of $C_2^T C_1$, respectively. If R has determinant -1, then it corresponds to an improper rotation. The minimizing proper rotation R' is then given by $R' = U' V^T$, where U' is equal to U except that the signs of the entries in the last column are flipped.

3   Expansive Search

In this section we will present a new algorithm for exploring the conformational space of a flexible protein using major modes. A docking program that searches

the conformational space of a flexible protein should have the following two properties:

- It should be biased towards low-energy conformations.
- It should efficiently explore 'unknown' parts of the search space.

There is some tension between these two requirements. If the algorithm is too biased toward low-energy conformations, it will get stuck in a local energy minimum. On the other hand, if exploration is emphasized too much, we may end up with many physically unrealistic conformations. The two requirements need to be carefully balanced.

We have developed an algorithm that has the two properties mentioned above. It is based on an algorithm that solves the motion planning problem, a problem in robotics. The goal in motion planning is to find a path between two configurations of a robot such that no configuration along the path is in collision with any obstacles. This problem is known to be *PSPACE*-hard, but randomized algorithms have been shown to be very practical to solve complicated motion planning problems in practice (Kavraki et al., 1996; Hsu et al., 1999; LaValle and Kuffner, 2001; Akinc et al., 2003). Our algorithm is based on one such algorithm (Hsu et al., 1999). The original algorithm builds up a tree of collision-free configurations. New leafs are added to the tree by sampling near existing nodes that are likely to be near unexplored parts of the space. After several iterations the tree represents a roadmap of the environment: a simplified 1-dimensional structure that has the same structure as the underlying high-dimensional configuration space.

To use a motion planning algorithm we need a representation of the degrees of freedom of the protein. We can consider a protein a very complex robotic system, where each dihedral angle is considered a degree of freedom. If we view docking as a motion planning problem, a motion planning algorithm would try to find collision-free path from a starting conformation to other stable conformations. A path specifies the values of the dihedral angles along the path. A conformation is considered in collision if its energy exceeds a threshold. Motion planning has been applied before to protein folding (see e.g., (Apaydin et al., 2004; Amato et al., 2003)). For *docking* most proteins of interest this approach would not work, because the number of degrees of freedom is simply too large (typically much larger 100). So instead, we use the major modes as approximate degrees of freedom.

Major modes are a linear approximation of the most important motions of a protein. However, the *real* motions are combinations of dihedral rotations (plus

6

some bond angle and bond length stretching). Therefore, a displacement along the major modes is only valid in a small neighborhood. To move beyond this neighborhood we need to perform some corrections to maintain a physical conformation. These corrections are applied each time the conformation moves more than a distance δ along the major modes. If the perturbation in the expansion step is of size x, then corrections are applied $\lceil x/\delta \rceil$ times. The corrections are applied in two stages. During the first stage the bond angle and bond length energy is minimized. These energy terms take time linear in the number of atoms to compute, so minimization can be done very efficiently. We do not perform a full minimization, but merely take a limited number steps along the gradient of the bond energy function. This is done in part because of efficiency reasons, but there is also the risk that with full minimization the conformation may move back to its original position. The first stage should fix up most of the distortion. During the second stage we minimize the total energy. Since this takes quadratic time, we only take a small number of steps along the gradient of the energy function.

Now that the representation of the degrees of freedom is defined we can present the algorithm. Our algorithm is easily defined inductively. It iteratively creates a set of conformations. The base case is one known conformation taken from, e.g., the Protein Data Bank. The inductive case is as follows. Suppose we have a set of $n-1$ conformations. To generate the nth conformation, we randomly select one of the previous $n-1$ conformations and generate a Gaussian perturbation of that conformation. The perturbation is computed in major mode space. This is called an *expansion* step. The key to the algorithm is how to select a conformation. This selection is based on a weighting function. Conformations with low weight are considered "good", whereas ones with high weight are considered "bad". Assume the conformations are sorted by increasing weight. We select conformation i for expansion according to a geometric distribution: the probability that we pick i for expansion is given by

$$P(i) = p(1-p)^{i-1}, \qquad i = 1, \ldots \tag{2}$$

The reason we use this distribution rather than just selecting the conformation with the lowest weight is that usually the top ranking conformations are all equally good. Figure 2 shows a simple visual representation of how the algorithm explores the conformational space. It creates an outward growing tree in the space spanned by two major modes, biased towards low-energy areas.

The weighting function should assign low weights to low-energy conformations and to conformations in the unexplored parts of the search space. Let the
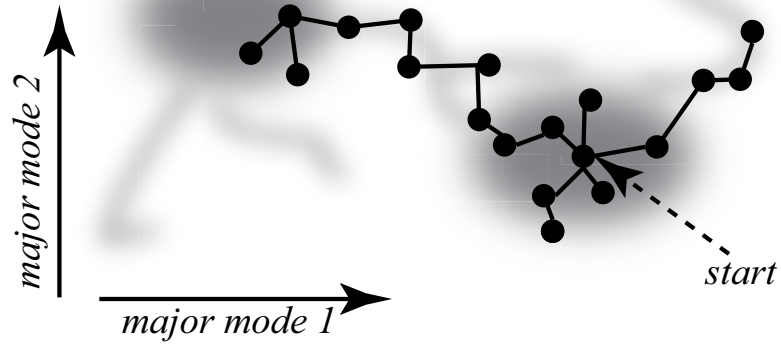
Figure 2: An expansive search for low-energy conformations. Conformations are sampled along the first two major modes. Low-energy areas in this two-dimensional space are indicated by the gray shaded areas. The starting point of the search is indicated by the arrow.

weight w_i of conformation i be defined as:

$$w_i = \big(1.1 - \exp(-\gamma(E_i - E_{\min}))\big) \cdot c_i / (1 + d_i), \tag{3}$$

where

$$\gamma = \text{a constant controlling the sensitivity to energy,} \tag{4}$$
$$E_i = \text{the energy of conformation } i, \tag{5}$$
$$E_{\min} = \min_{i=1,\dots,n} E_i, \tag{6}$$
$$c_i = \text{number of times conformation } i \text{ has been selected, and} \tag{7}$$
$$d_i = \text{sum of distances to } k \text{ nearest neighbors.} \tag{8}$$

Let us break down the weighting function in different parts and analyze how they contribute to the desired behavior. The energy term $\big(1.1 - \exp(-\gamma(E_i - E_{\min}))\big)$

ranges from 0.1 when $E_i = E_{\min}$ to 1.1 when E_i approaches infinity. In our implementation we use the CHARMM energy model (MacKerell et al., 1998), but the algorithm is not specific to this model. The energy function is very non-linear. Low-energy conformations are often very close to high-energy conformations. By using an exponential scaling function of the energy we emphasize differences in low-energy conformations and de-emphasize differences in high-energy conformations. Although we are searching the conformational space and not simulating molecular dynamics, one can think of the γ parameter as being equal to $1/(k_B T)$, i.e., the inverse of the Boltzmann constant times the temperature. This energy rescaling function was shown to be very effective in (Mancera et al., 2004). This technique is known as *stochastic tunneling* (Merlitz and Wenzel, 2002).

The number of times a conformation has been selected is used in the weighting function to prevent the algorithm from exhaustively exploring the space around low-energy conformations. The nearest neighbor distance serves a similar purpose, but it also forces the algorithm to select conformations with a large nearest neighbor distance and, thus, a low weight. Conformations with a large nearest neighbor distance are in sparsely sampled part of the search space, so the weighting function forces expansion in these areas.

After a conformation has been selected, we generate a new conformation by perturbing the selected conformation. We then compute the weight for the new conformation and insert it in the sorted list of conformations. Since the weight depends on the nearest neighbor distance, we also need to update the weights of the nearest neighbors of the new node. If the value of $E_{\min}$ changes, the value of all weights change. This should not happen too often, but this update can also be done lazily without deviating too much from the 'right' sorted order. By 'lazily' we mean in this case that only the weights of new nodes and their nearest neighbors are updated.

3.1   Random Bounce Walk

In this section and the next section we describe two changes that will improve the performance and the scalability of the algorithm, respectively. Typically, a conformation is energetically very constrained in moving to nearby conformations. Most of the search space corresponds to high-energy conformations. To move around in such a constrained space it may be helpful to generate new conformations with a short random walk rather than simply with a random perturbation. This may improve coverage of the search space, but obviously generating a neighbor with a random walk is more expensive. The random walk is generated as follows. First
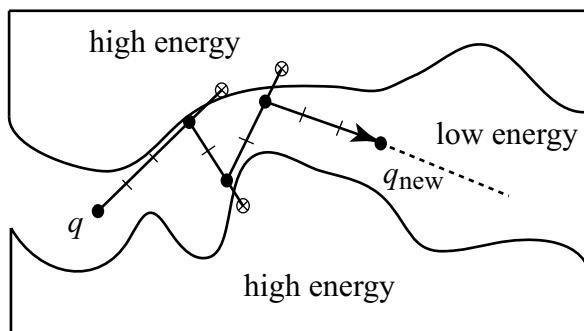
Figure 3: Illustration of the random bounce walk.

a random direction (in major mode space) is chosen. The conformation is moved along this direction in small increments until the energy exceeds a threshold. If a threshold is exceeded, the algorithm backtracks to the last good position and picks a new direction. This process is repeated until either the distance between the endpoint of the path and the starting conformation exceeds some threshold or a maximum number of steps is exceeded. The random walk is illustrated in figure 3. The walk starts at conformation q, bounces three times, and stops at q_{new} once the distance along the walk has reached a maximum threshold. The small dashes along the path indicate the intermediate conformations. The open circles denote high-energy conformations from which the algorithm backtracks to the conformations marked with a solid circle.

3.2 Multi-Level Search

To perform very large searches, the expansive search algorithm needs to be extended to run in parallel on a cluster of machines. To minimize the communication between subprocesses, we partition the search space and run the algorithm separately on different parts of the space. Unfortunately, we do not know a priori how to partition the search space into promising regions for a search. However, once we have run a search, we have some information about where to search next. This leads to the following multi-level search algorithm. The algorithm keeps track of a collection of *seeds*: promising starting points for a search. Initially, the only seed is a known crystal structure. After one search we select a fixed number of conformations whose energy is below some threshold and who are all at least some distance apart from each other and from the seeds, and add them to the set of seeds. We then start a search from each of the new seeds. To prevent the expan-

sive search from covering the exact same part of the space over and over again, we need to make a small modification. At the start of each search, the set of all seeds is read in to initialize the set of generated conformations. The nearest neighbor distance component of the weighting function will guide the search toward unexplored parts of the space. Nevertheless, it is inevitable that some part of the search space covered in previous levels will be covered again at the current level.

4 Performance Evaluation

To judge how well our algorithm is doing, we need to define a number of ways to measure the performance. Due to the complexity of the problem and the number of parameters involved, the result of a particular instantiation of the algorithm cannot be characterized by one number. The algorithm takes as input a start conformation and produces as output a set of conformations with their CHARMM energy. In our experiments we have used the following two performance measures, each of which highlights a different aspect of the search results:

- number of different low-energy conformations, and

- distance to known crystal structures.

Below we will describe these performance measures in more detail.

4.1 Number of Different Low-Energy Conformations

One of the main goals of our docking program is to find low-energy conformations, so it is only natural to consider the number of low-energy conformations it generates. Many conformations that the search produces will be so close together that for all practical purposes they are the same. We would like to cluster all low-energy conformations that are close together. For this we have developed a very simple algorithm. It takes as input the conformations produces by the search, sorted by energy. The conformations with energy larger than some energy threshold are rejected immediately. The algorithm incrementally constructs a kd-tree of all remaining conformations that are at least some minimum distance apart. The conformation with the lowest energy is always added to the kd-tree. For subsequent conformations we determine the distance to the nearest neighbor in the kd-tree. If this distance greater than the threshold distance, we add it to the kd-tree. From a geometric point of view this is not the optimal solution to finding

11

conformations that are furthest apart from each other. But for our purposes the fact that our algorithm is 'greedy' is actually a desirable property. It biases the outcome towards low-energy conformations.

Another approach to find different low-energy conformations is to use a 'traditional' clustering approach and take the centroids to be the representative conformation of a cluster. In theory this could work, because we expect to see a higher density of conformations in low-energy areas of the search space. In our experience, however, this difference in density was not enough in practice to overcome the fact that clusters tend to blend together. Conformations are generated as deviates from previously generated conformations. We suspect that our conformations form a high-dimensional star-shaped set. For our experiments we used k-means and x-means clustering Pelleg and Moore (2000). K-means clustering is an iterative procedure that determines a clustering with k clusters. X-means is a variant of k-means that automatically determines the statistically optimal value for k. Running x-means clustering on our search results produced hundreds of clusters, which leads us to believe that the conformations do not form well-separated clusters.

4.2   Distance to Known Crystal Structures

The second performance measure we used is the distance to crystal structures. For many large, flexible proteins there are several different crystal structures in the Protein Data Bank (PDB). In each crystal structure the protein is bound to a different ligand. Ignoring the ligand data, we are interested in measuring the distance between each crystal structure to its nearest neighbor in the set of conformations produced by the search. If for many crystal structures there is a nearby conformation in the search results, then the algorithm produces biologically plausible structures and is a useful tool in predicting docking targets. For this performance measure we ignore energy. We are assuming that the search results will be ranked according some other energy/scoring function. The top structures are then docked with ligands using a docking program that assumes a rigid protein. There is one practical issue with using crystal structures from the PDB. The crystal structures sometimes differ in a few residues. We replace those residues with the corresponding residue in the molecule used for the search. We follow this substitution by a simple local energy minimization to resolve steric clashes resulting from the substitution. Finally, we find the optimal alignment with the start conformation of the search using SVD as explained in section 2.2.
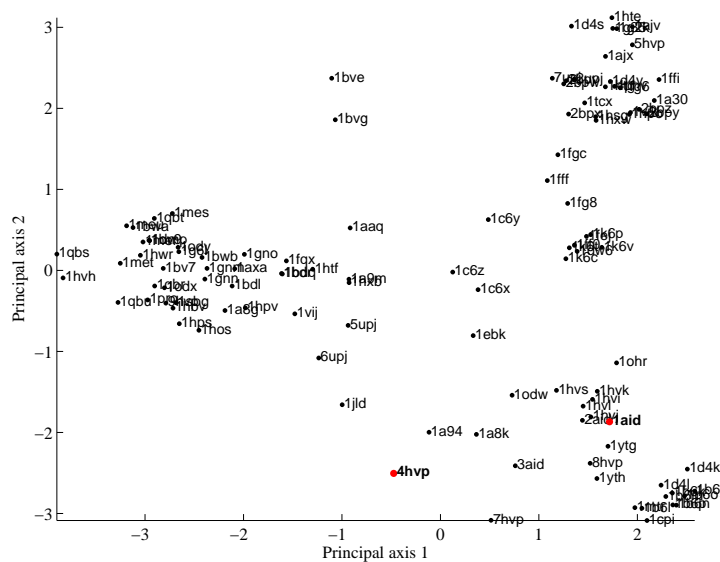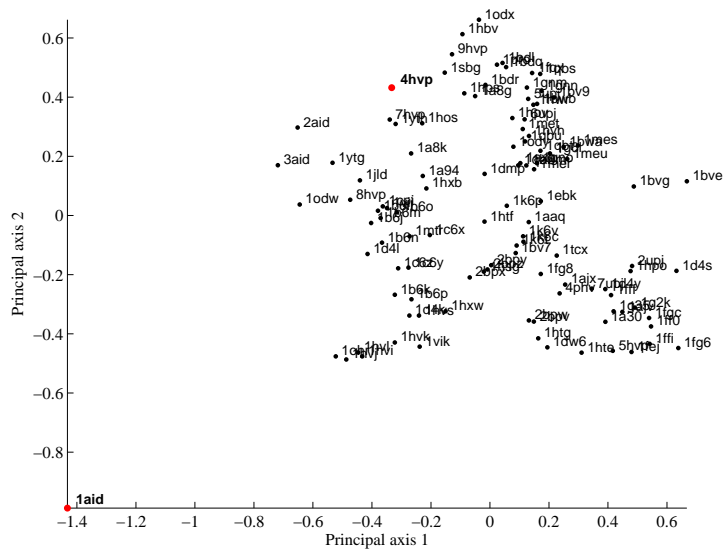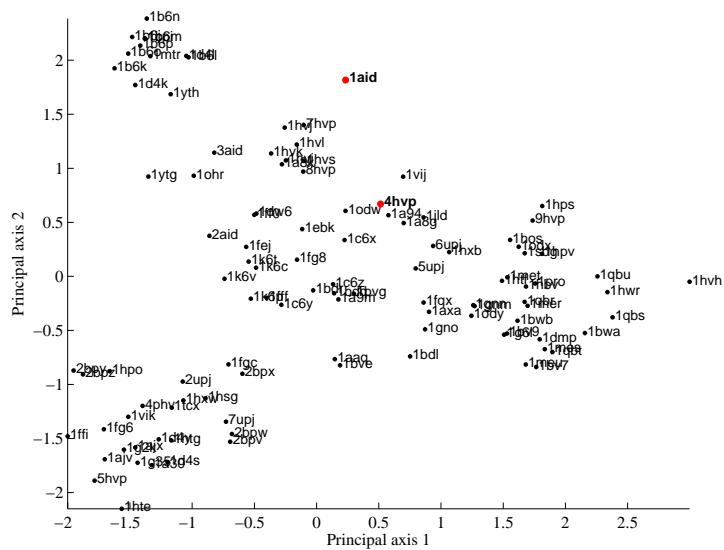
(a) All-atom projection
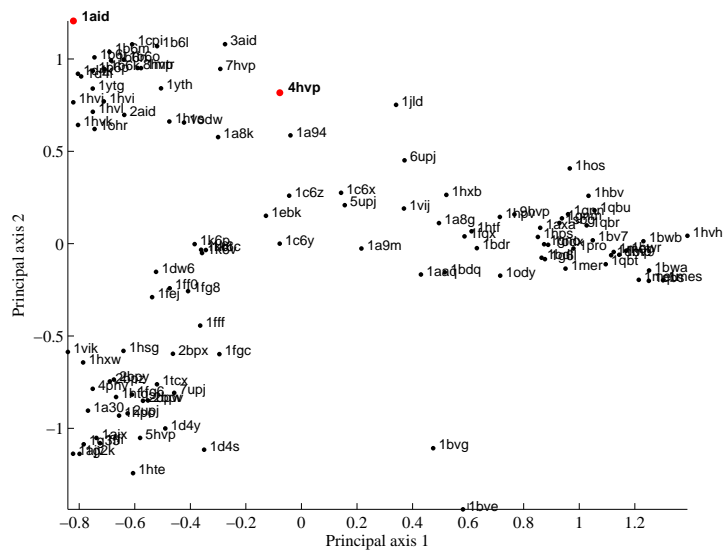
(b) Backbone projection

Figure 4: Projection of 111 crystal structures on the first two principal vectors of four different sets of atoms.

13

(c) Binding site projection

(d) Extended binding site projection

Figure 4: (Continued) Projection of 111 crystal structures on the first two principal vectors of four different sets of atoms.

14

Measuring distance between the crystal structures and the search results is only useful if we know how large the distance is in major mode space, since this is the space we will actually be exploring. The local energy minimization will help us explore a small part of the space that is not in the span of the major modes, but this is generally only a small displacement. Let us first consider the relative distances between crystal structures of HIV-1 protease. Figure 4 shows the projection of 111 crystal structures onto the first two principal vectors after residue substitution, local energy minimization and alignment with 4hvp, the crystal structure we used as a docking target. The principal components can be determined for all atoms, the backbone atoms (594 atoms), the binding site atoms (266 atoms), or the extended binding site atoms (948 atoms), resulting in slightly different projections. In figure 4 we have shown all four projections and highlighted 4hvp and 1aid. This last structure was chosen as the start conformation for our expansive search. In the 1aid structure the flaps of HIV-1 protease are wide open, whereas in 4hvp they are almost closed. In figure 5 show how well these crystal structures can be approximated with major modes, starting from 1aid. The major modes in the search were computed from a 2ns NAMD simulation started at 4hvp. Note that the curves for the backbone RMSD and (extended) binding site RMSD are not necessarily monotonically decreasing. Also note that the docking algorithm could produce conformations that are closer to the crystal structures than one would expect based on figure 5. This is possible because the energy is locally minimized when a conformation is moving along major modes. This minimization can cause a displacement that is not in the span of the major modes.

5   Implementation

5.1   Computation of Major Modes

We have implemented a parallel version of PCA that can compute the major modes of large trajectories very efficiently. For example, computing 20 major modes of 10,000 conformations of HIV-1 protease, a molecule with 3120 atoms and 198 residues, takes about 118 seconds on 6 dual-processor nodes of our cluster of 1600MHz Athlon CPUs. Our implementation is based on the P_ARPACK library (Maschhoff and Sorensen, 1996), a parallel version of ARPACK (Lehoucq et al., 1998). It allows for efficient computation of the first couple of singular vectors without explicitly computing the covariance matrix.
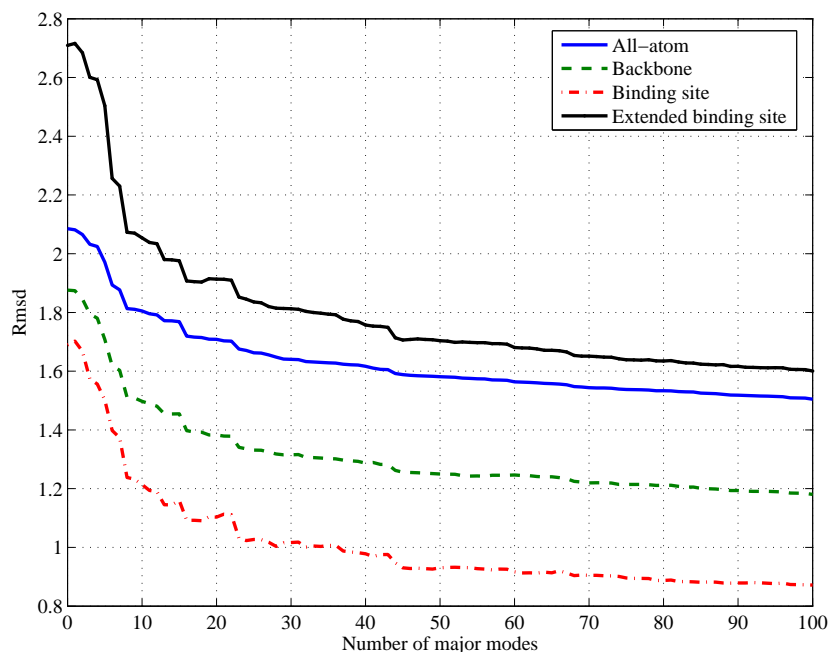
15

Figure 5: Average distance between crystal structures and their projection onto major modes, measured over different sets of atoms.

5.2 Search Results

We tested our conformational search on two systems that are known to be very flexible. The first one is HIV-1 protease. It has 3120 atoms, and there are about 110 crystal structures available. The second one is FK506 binding protein, a system with 1663 atoms, and 70 available crystal structures. In our searches we used the first five main modes of motion, computed from NAMD simulations. Certain parameters such as number of nearest neighbors to use and local minimization parameters were optimized once, and kept fixed throughout the different tests. For other parameters we performed a parameter sweep over many combinations. The search algorithm turned out to be relatively insensitive to a lot of these parameters. Below we report on some of the interesting patterns we observed. Results are averaged over five runs.

One of the performance evaluation criteria we described was the number of low-energy conformations. In our tests we defined this as the number of conformations who are all at least 1Å RMSD apart and whose energy was roughly equal to the average crystal structure energy or lower. Figure 6 shows the results for

16
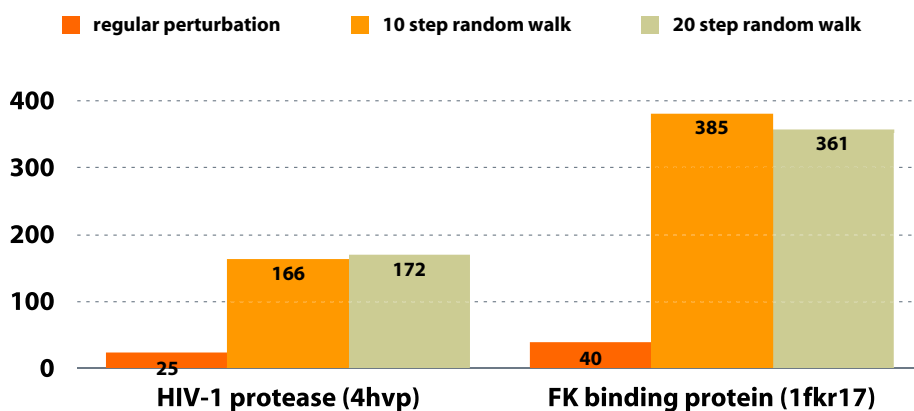
(a) Number of distinct low-energy conformations.
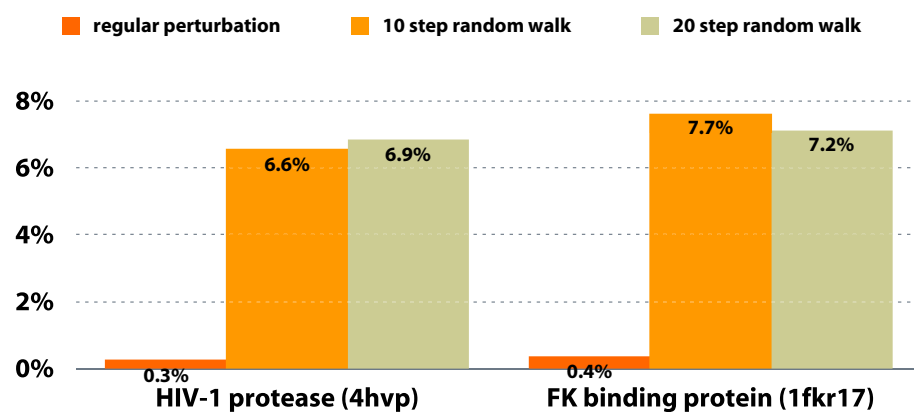
(b) Percentage of distinct low-energy conformations w.r.t. total number of conformations generated.

Figure 6: Number of low-energy conformations with different neighbor selection strategies, measured after 20 hours.

the two different systems and for different neighbor selection schemes. These results clearly show that the random walk neighbor selection scheme helps significantly in producing more low-energy conformations. Not only does the random walk neighbor selection produce more low-energy conformations (figure 6a), it is also more efficient in finding them (figure 6b). In the case of HIV-1 protease, the regular perturbation produced about 10,000 conformations and the random walk search between 2,500 and 3,000. For FK506 binding protein, the regular perturba-

		min (cal/mol)	average crystal structs. (cal/mol)
	regular	-3968	
4hvp	10 steps	-3436	-2000
	20 steps	-3413	
	regular	-1885	
1fkr17	10 steps	-1725	-1000
	20 steps	-1813	

Table 1: Comparison between minimum energies produced by the search algorithm and the energies of crystal structures.

tion produced about 10,000 conformations and the random walk search between 5,000 and 6,000. The differences between HIV-1 protease and FK506 binding protein are mostly due to the difference in size of the molecules.

In terms of approximation of known crystal structures, our algorithm did not fare as well. The nearest neighbor in the search results for each crystal structure was only marginally closer to the corresponding crystal structure than the original starting point. The average distance between the starting conformation and the crystal structures was 1.76Å RMSD for HIV-1 protease with 4hvp as starting point. For FK506 binding protein the average distance to the starting conformation 1fkr17 was 2.70Å RMSD. The typical experimental resolution is 1.50Å, so this explains in part the difficulty in getting much closer to the crystal structures. Another mitigating factor is that RMSD is measured over the whole molecule, including parts that may not matter for a protein's function. Nevertheless, in future work we investigate further how our search results can be compared with experimental data.

In terms of finding low-energy conformations the algorithm performed very well. Table 1 shows that the searches produced conformations with energies significantly lower than the average crystal structure energy. In other words, the algorithm is very effective at finding conformations with very low energy. It also shows that the limitations of the CHARMM energy model we used: despite the fact that we can find many low-energy conformations, the search results are not that much closer to the crystal structures than the starting point for the search.

6 Discussion

This report describes a new approach to generating low-energy conformations of a flexible protein for the purpose of docking flexible ligands. Previous docking approaches only allow for limited flexibility of the protein, whereas our approach can model large conformational changes. For many molecules this flexibility is essential to their function. The flexibility of the protein is represented by the main modes of motion, which are computed from a molecular dynamics trajectory. The main modes form a reduced basis. We explore conformations in this reduced basis using an algorithm that is based on a robotics motion planning algorithm. The algorithm simultaneously tries to explore unknown parts of the search space and around low-energy conformations.

Our simulations showed that we are able to produce many different low-energy conformations for HIV-1 protease and FK506 binding protein. On the downside, the conformations produced by a search are only marginally closer to other known crystal structures than the starting conformation for a search. This means that the algorithm performs well within the model, but the model has its limitations. Improving the model is the main focus of future research.

In the near future we plan to extend the model in the following ways. First, we plan to perform searches with a large number of major modes. To prevent the algorithm from exploring parts of the search space that are unlikely to contain low-energy conformations, the search will be biased towards the first couple of major modes. This can be accomplished by using the singular values obtained during the major mode computation as weights. We also expect to further improve the results by using symmetry preserving major modes (Shah and Sorensen, 2005; Moll et al., 2005) for systems that exhibit symmetry such as HIV-1 protease. In (Moll et al., 2005) we showed that symmetry preserving major modes are better in explaining the variability in crystal structures. Exploiting the symmetry reduces the number of degrees of freedom. A second, more important extension is to use *free* energy rather than energy to guide the search. Exact computation of the free energy is impractical, but a reasonable estimate might be sufficient to arrive at physically more realistic results. A third route that we plan the explore in conjunction with the previous two ideas is to split the search up in two stages. In the first stage we treat the problem purely geometrically. The output of this stage consists of conformations without steric clashes. In the second stage, the "best" conformations are ranked with a (free) energy function. Since in the current version of the algorithm energy calculations dominate the run-time, the multi-stage approach promises to deliver a huge speed-up.

19

Acknowledgments

References

Abseher, R. and Nilges, M. (2000). Efficient sampling in collective coordinate space. *Proteins*, 39:82–88.

Akinc, M., Bekris, K. E., Chen, B. Y., Ladd, A. M., Plaku, E., and Kavraki, L. E. (2003). Probabilistic roadmaps of trees for parallel computation of multiple query roadmaps. In *Proc. of Eleventh Intl. Symp. on Robotics Research*, Siena, Italy.

Amadei, A. and Berendsen, A. B. L. H. J. (1993). Essential dynamics of proteins. *Proteins*, 17:412–425.

Amadei, A., Linssen, A. B., de Groot, B. L., van Aalten, D. M., and Berendsen, H. J. (1996). An efficient method for sampling the essential subspace of proteins. *J Biomol Struct Dyn*, 13:615–625.

Amato, N. M., Dill, K. A., and Song, G. (2003). Using motion planning to map protein folding landscapes and analyze folding kinetics of known native structures. *J. Comp. Bio.*, 10(3–4):239–255.

Apaydin, M. S., Brutlag, D. L., Guestrin, C., Hsu, D., and Latombe, J.-C. (2004). Stochastic conformational roadmaps for computing ensemble properties of molecular motion. In Boissonnat, J. D., Burdick, J., Goldberg, K., and Hutchinson, S., editors, *Algorithmic Foundations of Robotics V*, pages 131–147. Springer.

Apostolakis, J., Pluckthun, A., and Caflisch, A. (1998). Docking small ligands in flexible binding sites. *J. Comp. Chemistry*, 19:21–37.

Bursavich, M. G. and Rich, D. H. (2002). Designing non-peptide peptidomimetics in the 21st century: Inhibitors targeting conformational ensembles. *Journal of Medicinal Chemistry*, 45:541–558.

Bystroff, C. and Kraut, J. (1991). Crystal structure of unliganded Escherichia coli dihydrofolate reductase. ligand-induced conformational changes and cooperativity in binding. *Biochemistry*, 30(8):2227–2239.

Caves, L. S., Evanseck, J. D., and Karplus, M. (1998). Locally accessible conformations of proteins: multiple molecular dynamics simulations of crambin. *Protein Science*, 7:649–666.

Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., K. M. Merz, J., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., and Kollman, P. A. (1995). A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.*, 117:5179–5197.

de Groot, B. L., Amadei, A., Scheek, R. M., van Nuland, N. A., and Berendsen, H. J. (1996a). An extended sampling of the configurational space of HPr from E. coli. *Proteins*, 26:314–322.

de Groot, B. L., Amadei, A., van Aalten, D. M., and Berendsen, H. J. (1996b). Toward an exhaustive sampling of the configurational spaces of the two forms of the peptide hormone guanylin. *Biomol Struct Dyn*, 13:741–751.

Di Nola, A., Roccatano, D., and Berendsen, H. J. (1994). Molecular dynamics simulation of the docking of substrates to proteins. *Proteins*, 19:174–182.

Duan, X. and Quiocho, F. (2002). Structural evidence for a dominant role of nonpolar interactions in the binding of a transport/chemosensory receptor to its highly polar ligands. *Biochemistry*, 41:706–712.

Fischer, E. (1894). Einfluss der Configuration auf die Wirkung der Enzyme. *Ber. Dtsch. Chem. Ges.*, 27:2985–2993.

García, A. E. (1992). Large-amplitude nonlinear motions in proteins. *Physical Review Letters*, 68(17):2696–2699.

Go, N., Noguti, T., and Nishikawa, T. (1983). Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proceedings of the National Academy of Science USA*, 80:3696–3700.

Golub, G. H. and Loan, C. F. V. (1996). *Matrix Computations*. Johns Hopkins University Press, third edition.

Hsu, D., Latombe, J.-C., and Motwani, R. (1999). Path planning in expansive configuration spaces. *International Journal of Computational Geometry and Applications*, 9(4-5):495–512.

Jiang, F. and Kim, S. H. (1991). "Soft docking": Matching of molecular surface cubes. *J. of Molec. Bio.*, 219:79–102.

Jolliffe, I. T. (1986). *Principal Components Analysis*. Springer-Verlag, New York.

Kalé, L., Skeel, R., Bhandarkar, M., Brunner, R., Gursoy, A., Krawetz, N., Phillips, J., Shinozaki, A., Varadarajan, K., and Schulten, K. (1999). Namd2: Greater scalability for parallel molecular dynamics. *Journal of Computational Physics*, 151:283–312.

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

Keserû, G. M. and Kolossváry, I. (2001). Fully flexible low-mode docking: application to induced fit in HIV integrase. *J Am Chem Soc*, 123:12708–12709.

Kitao, A. and Go, N. (1999). Investigating protein dynamics in collective coordinate space. *Current Opinion in Structural Biology*, 9(2):164–169.

Knegtel, R. M., Kuntz, I. D., and Oshiro, C. M. (1997). Molecular docking to ensembles of protein structures. *J. of Molec. Bio.*, 266:424–440.

Kolossváry, I. and Guida, W. C. (1999). Low-mode conformational search elucidated: Application to C39H80 and flexible docking of 9-deazaguanine inhibitors into PNP. *J. Comp. Chemistry*, 20:1671–1684.

Kolossváry, I. and Keserû, G. M. (2001). Hessian-free low-mode conformational search for large-scale protein loop optimization: Application to c-jun N-terminal kinase JNK3. *J. Comp. Chemistry*, 22:21–30.

Koshland, D. E. (1958). Application of a theory of enzyme specificity to protein synthesis. *Proceedings of the National Academy of Science USA*, 44:98–104.

LaValle, S. M. and Kuffner, J. J. (2001). Randomized kinodynamic planning. *International Journal of Robotics Research*, 20(5):378–400.

Leach, A. R. (1994). Ligand docking to proteins with discrete side-chain flexibility. *J. of Molec. Bio.*, 235:345–356.

Leach, A. R. and Lemon, A. P. (1998). Exploring the conformational space of protein side chains using dead-end elimination and the A∗ algorithm. *Proteins*, 33:227–239.

Lehoucq, R. B., Sorensen, D. C., and Yang, C. (1998). *ARPACK Users' Guide: Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods*. SIAM.

Levitt, M., Sander, C., and Stern, P. S. (1985). Protein normal-mode dynamics: trypsin inhibitor, crambin, ribonuclease and lysozyme. *J. of Molec. Bio.*, 181:423–447.

Levy, R. M. and Karplus, M. (1979). Vibrational approach to the dynamics of an alpha-helix. *Biopolymers*, 18:2465–2495.

Ma, B., Kumar, S., Tsai, C. J., and Nussinov, R. (1999). Folding tunnels and binding mechanisms. *Protein Engineering*, 12:713–720.

Ma, B., Shatsky, M., Wolfson, H. J., and Nussinov, R. (2002). Multiple diverse ligands binding at a single protein site: A matter of pre-existing populations. *Protein Science*, 11:184–197.

Ma, B., Wolfson, H. J., and Nussinov, R. (2001). Protein functional epitopes: Hot spots, dynamics and combinatorial libraries. *Curr Opin Struct Biol*, 11:364–369.

MacKerell, Jr., A. D., Bashford, D., Bellott, M., Dunbrack Jr., R. L., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F. T. K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D. T., Prodhom, B., Reiher, III, W. E., Roux, B., Schlenkrich, M., Smith, J. C., Stote, R., Straub, J., Watanabe, M., Wiórkiewicz-Kuczera, J., Yin, D., and Karplus, M. (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. *Journal of Physical Chemistry B*, 102:3586–3616.

Mancera, R. L., Källblad, P., and Todorov, N. P. (2004). Ligand-protein docking using a quantum stochastic tunneling optimization method. *J. Comp. Chemistry*, 25:858–864.

Mangoni, M., Roccatano, D., and Di Nola, A. (1999). Docking of flexible ligands to flexible receptors in solution by molecular dynamics simulation. *Proteins*, 35:153–162.

Maschhoff, K. J. and Sorensen, D. C. (1996). P_ARPACK: An efficient portable large scale eigenvalue package for distributed memory parallel architectures. In *Proceedings of the Third International Workshop on Applied Parallel Computing, Industrial Computation and Optimization*, pages 478–486. Springer-Verlag.

Merlitz, H. and Wenzel, W. (2002). Comparison of stochastic optimization methods for receptor-ligand docking. *Chemical Physics Letters*, 362(3–4):271–277.

Moll, M., Shah, M., Sorensen, D. C., and Kavraki, L. E. (2005). A study of molecular flexibility using main modes of motion. In *Proceedings of the 13th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB)*. Submitted.

Nakajima, N., Higoa, J., Kiderab, A., and Nakamura, H. (1997a). Flexible docking of a ligand peptide to a receptor protein by multicanonical molecular dynamics simulation. *Chem Phys Lett*, 278:297–301.

Nakajima, N., Nakamura, H., and Kidera, A. (1997b). Multicanonical ensemble generated by molecular dynamics simulation for enhanced conformational sampling of peptides. *J Phys Chem B*, 101:817–824.

Pang, Y. P. and Kozikowski, A. P. (1994). Prediction of the binding sites of huperzine a in acetylchilinesterase by docking studies. *Journal of Computer-Aided Molecular Design*, 8:669–681.

Pelleg, D. and Moore, A. W. (2000). X-means: Extending k-means with efficient estimation of the number of clusters. In *Seventeenth Intl. Conf. on Machine Learning*, pages 727–734.

Rhodes, G. (1993). *Crystallography Made Clear*. Academic Press, London.

Romo, T. D., Clarage, J. B., Sorensen, D. C., and G. N. Phillips, J. (1995). Automatic identification of discrete substates in proteins: singular value decomposition analysis of time-averaged crystallographic refinements. *Proteins*, 22:311–321.

Schnecke, V., Swanson, C. A., Getzoff, E. D., Tainer, J. A., and Kuhn, L. A. (1998). Screening a peptidyl database for potential ligands to proteins with side-chain flexibility. *Proteins*, 33:74–87.

Shah, M. and Sorensen, D. C. (2005). A symmetry preserving singular value decomposition. CAAM TR 05-01, Dept. of Computational and Applied Math., Rice University.

Sudbeck, E. A., Mao, C., Vig, R., Venkatachalam, T. K., Tuel-Ahlgren, L., and Uckun, F. M. (1998). Structure-based design of novel dihydroalkoxybenzylox-opyrimidine derivatives as potent nonnucleoside inhibitors of the human immunodeficiency virus reverse transcriptase. *Antimicrob Agents Ch*, 42:3225–3233.

Teodoro, M. L. (2003). *Modeling Protein Flexibility Using Collective Modes of Motion: Applications to Drug Design*. PhD thesis, Rice University.

Teodoro, M. L., George N. Phillips, J., and Kavraki, L. E. (2003). Understanding protein flexibility through dimensionality reduction. *J. Comp. Bio.*, 10(3-4):617–634.

Urzhumtsev, A., Tête-Favier, F., Mitschler, A., Barbanton, J., Barth, P., Urzhumt-seva, L., Bielmann, J.-F., and Moras, D. (1997). A 'specificity' pocket inferred from the crystal structures of the complexes of aldose reductase with the pharmaceutically important inhibitors tolrestat and sorbinil. *Structure*, 5(5):601–612.

Weber, P. C., Ohlendorf, D. H., Wendoloski, J. J., and Salemme, F. R. (1989). Structural origins of high-affinity biotin binding to streptavidin. *Science*, 243(4887):85–88.

Wlodawer, A., Miller, M., Jaskolski, M., Sathyanarayana, B. K., Baldwin, E., Weber, I. T., Selk, L. M., Clawson, L., Schneider, J., and Kent, S. B. (1989). Conserved folding in retroviral proteases: crystal structure of a synthetic hiv-1 protease. *Science*, 245(4918):616–621.

Wüthrich, K. (1986). *NMR of Proteins and Nucleic Acids*. Wiley & Sons, New York.

Zacharias, M. and Sklenar, H. (1999). Harmonic modes as variables to approximately account for receptor flexibility in ligand- receptor docking simulations: Application to DNA minor groove ligand complex. *J. Comp. Chemistry*, 20:287–300.