Model-Based Search to Determine Minima in Molecular Energy Landscapes

Technical Report 04-48

TJ Brunette Oliver Brock
Laboratory for Perceptual Robotics
Department of Computer Science
University of Massachusetts Amherst

September 2004

Abstract

Search for the global minimum in a molecular energy landscape populated with numerous local minima is a difficult task. Search techniques relevant to such complex spaces can be classified as either global or local. Global search explores the entire space, guaranteeing the global extremum will be found. To accomplish this, the number of samples required grows exponentially with the number of dimensions. Since this is clearly not computationally tractable, global search is impractical in high-dimensional spaces. Local search, on the other hand, employs gradient descent to avoid searching the entire exponential space. Gradient descent methods are susceptible to getting stalled in local minima and consequently, no guarantees can be made about finding the global minimum. We propose a middle ground that minimizes the effects of exponential space and local minima by integrating domain knowledge and information generated during search into a model, and then using this model to focus computation on regions of increasing relevance. Directing resources to multiple relevant regions prevents oversampling local minima. At the same time the exploration of only significant regions avoids the intractable computational requirements of high-dimensional spaces. The proposed method, called Model-Based Search (MBS), is compared to the local search method Monte Carlo as implemented in Rosetta - currently considered the best computational protein structure prediction method. The results indicate that MBS is significantly better at finding lower energy minima than the Monte Carlo technique implemented as part of Rosetta. This effect is amplified as the dimensionality of the search space increases.

1 Introduction

Over the past five years, gene sequences have been discovered at a rate approximately 65 times faster than experimentalists have been able to determine the structure of the proteins they represent [1, 2, 3]. If the rate of protein structure determination remains static, it will take over 300 years to determine the native state of the proteins that have been identified from gene sequences so far. Not only are current experimental techniques labor-intensive, but they face challenges for large categories of proteins, such as membrane proteins [4, 5]. Without knowledge of the structure, it is difficult, if not impossible, to determine the biological function of proteins [6, 7]. Consequently, efficient computational approaches to structure determination represent an important tool that can enable researchers to more quickly establish a protein's function, thereby

gaining understanding into the mechanisms that underly diseases, which could ultimately lead to the design of drugs to cure them.

Unfortunately, *ab initio* protein structure prediction is a difficult problem. Levinthal's paradox gives a perspective on the size of conformational space [8]. Limiting each amino acid to only three distinct spatial placements, a protein consisting of 200 amino acids could assume 3^{200} different conformations. Assume it takes 10^{-12} seconds to evaluate the energy of each state. Conducting exhaustive search in the resulting conformation space would take $8.4 * 10^{75}$ years. That is far longer then life has existed on earth. Clearly, exhaustive search is not possible.

To avoid exhaustive search, existing methods for *ab initio* protein structure prediction (see Section 2) rely on local methods, such as gradient descent. These methods search the energy landscape associated with conformation space by descending its gradient. Since the energy landscape contains many local minima, local methods may fail to find the global extremum. During their search of the conformation space, much computation time is wasted with the exploration of local minima. Most importantly, these methods are generally "memory-less." By that we mean that they do not remember the conformations they have already explored and consequently cannot avoid repeated exploration of very similar conformations.

We propose a novel search method for conformation space that overcomes the shortcomings of purely local methods while avoiding the consequences of Levinthal's paradox. This is accomplished by restricting exploration of conformation space to only biologically relevant regions. Contrary to local methods, such as Monte Carlo (see Section 2), some information about previous evaluations of the energy function of a protein is maintained in what is referred to as a model. This model captures essential features of the energy landscape and can guide future exploration. Search then proceeds by iteratively refining the model with the most relevant information and by discarding information about regions that have been identified as local minima. Due to the use of a model, we refer to this method as *model-based search*.

It should be emphasized that the objective of this research is to devise a more efficient search method for biologically motivated energy landscapes. The underlying energy function searched by the proposed method is taken from the literature [9]. Consequently, a successful experiment using model-based search finds conformations of lower energy than other methods. The quality of the predicted structure with respect to the native structure of a protein, however, depends on the quality of the energy function. Therefore, finding a lower-energy conformation is not equivalent to determining the protein structure more accurately. Nevertheless, the understanding obtained by finding the global minimum of an energy function can be used to improve the energy function itself.

We present experiments in which model-based search can find lower-energy conformations in an approximated protein energy landscape than the leading protein structure predictor *Rosetta* [9].

2 Related Work

2.1 Local Search Methods

Locating the extremum in high dimensional continuous space with many local extrema can be difficult. Relevant search techniques can be classified as either global or local. Global search systematically explores the entire space, guaranteeing the extremum will be found. However, as dimensionality increases, the amount of time and memory to perform this search grows exponentially. In contrast, local search follows a single path through the energy landscape and thus typically has minimal memory requirements. The lack of global information causes local techniques to get stalled in extrema and flat regions of space. For local methods no guarantees about locating the extremum can be made. In practical applications, this disadvantage is

outweighed by the computational efficiency of local methods. When compared with global methods, local methods are capable of finding near-optimal extrema in relatively complex energy landscapes in a reasonable amount of time.

The simplest local search algorithm is hill climbing or gradient descent [10]. In discrete spaces, gradient descent proceeds in a greedy fashion, choosing the neighbor with the lowest value as the next state. In high-dimensional continuous space this is infeasible since all neighboring states can not be examined. Instead, one adjoining state is randomly sampled, and if it represents a lower value, it is accepted. This corresponds to a random walk through the landscape described by the searched function, in which only steps towards states with a lower value are acceptable. This random walk approximates gradient descent in high-dimensional spaces in which the computation of the gradient is not possible. In the domain of protein structure prediction, this search method is called the Monte Carlo (MC) algorithm [11]. Generally, many repeated runs of Monte Carlo are attempted, increasing the probability of finding the global extremum.

A local search that only accepts states that decrease in value will always become stuck upon encountering a local extremum. In molecular biology, the Monte Carlo algorithm has been augmented with the Metropolis criterion to allow search to escape local extrema [12]. As before, a randomly generated neighboring state with lower value—in this case the energy of a molecule— is accepted as a successor state. If energy increases, the step is accepted with a probability inversely proportional to the increase in energy. Consecutive successor states of increasing energy allow the search to escape a local extremum.

To further decrease the chances of local search getting stuck in an extremum, initial stages should examine a wide variety of states without becoming trapped. But, as search progresses and the states being examined are lower in energy, local extrema need to be more throughly explored. Simulated annealing accomplishes this by adjusting the likelihood of MC to accept a state that increases the energy [13]. A temperature variable in the Boltzmann equation used by Metropolis Monte Carlo is slowly lowered causing large jumps in energy to become increasingly unlikely as search progresses [14].

Every individual invocation of the Monte Carlo method tracks a single search trajectory by keeping a single state in memory. Tracking multiple trajectories and keeping multiple states in memory, however, allows computation to be biased towards the best state among those kept in memory. Beam search maintains a constant width frontier for search, adding states in a depth-first manner based upon a heuristic [15]. One of the big drawbacks of beam search is that over time, multiple searches become concentrated in a single region of space [13].

Genetic algorithms also track multiple paths by maintaining a population of states and applying the principles of evolution to this population. States in a new generation of the population are generated from the previous generation by combining two parent states [16].

2.2 Search Methods for Protein Folding and Protein Structure Prediction

The task of determining a protein's native state can be treated as a search through a vast conformational space for the lowest energy structure. All algorithms currently used for native state prediction combine domain information with local search techniques to avoid the complexity of high-dimensional conformation spaces.

Molecular dynamics (MD) is search that attempts to predict protein structure by simulating the folding process that occurs in nature [17, 18, 19, 20, 21]. This is done by simulating the interactions of all forces acting between atoms of the protein and solvent. The interactions are modeled using Newton's law at time steps equivalent to atomic thermal vibrations. Evaluating the forces acting upon all molecules at every time step of a folding protein is biologically accurate, but computationally prohibitively expensive. Like other local search techniques MD is susceptible to spending a significant amount of computation in local minima.

To escape from a minimum, thermal vibrations must sample a state outside of the minimum. Currently MD can simulate approximately 100 nanoseconds of protein folding. However, the fastest proteins fold in the high microsecond range, with many proteins taking over a second. Even with the continued exponential increase in computational power, it is unlikely MD will be a practical solution to protein folding for several decades.

To reduce the computational requirements of molecular dynamics, larger steps need to be taken between states. The Metropolis Monte Carlo algorithm uses an energetically biased random walk in the approximated energy landscape of a protein to find the native state [11, 22, 23, 24]. The large step size between states combined with randomly choosing the direction of the next step, as opposed to inferring the gradient from the atom-atom interactions, make MC a much more computationally tractable algorithm. A consequence of these shortcuts is that the resulting protein motion is less biologically plausible.

Several heuristics have been developed to help MC minimize oversampling local minima in the protein domain. The technique of simulated annealing [14, 25] is applicable as is interrupting the simulation to accept a high-energy state (jump walking) [26] or directly controlling the probability with which states at different energy levels are sampled (multi-canonical ensemble method) [22, 27]. Approaches that combine multiple techniques are also possible [28].

Integration between domain and search techniques can further enhance search. Fold recognition (also called threading) exploits the insight that amino acid sequences with a high degree of homology often fold into similar three-dimensional structures [29, 30]. Fold Recognition algorithms like *Rosetta* (currently considered the most accurate method of protein structure prediction [31]) search a database of previously folded proteins for short amino acid fragments with aligned sequences. Once these alignments are found the corresponding secondary structures can be assembled using Monte Carlo to find the native state [32]. Using these short alignments greatly reduces search space. This technique is currently the most accurate method to *ab initio* protein folding. However, like its underlying search technique (Monte Carlo), threading is still susceptible to the problem of oversampling local minima.

Molecular dynamics, Monte Carlo, and threading can be considered to be variations of local search. Local search algorithms are susceptible to oversampling local minima because there is no method of detecting if a lower minimum exists or how it can be reached.

2.3 Comparative Modeling for Protein Structure Prediction

Significant sequence homology between two proteins often imply structural similarity. Comparative modeling exploits this by predicting the 3-dimensional structure of an unknown protein using the known structure of a homologous protein as a template. By relying on databases of known structures, the aforementioned search of conformation space can be avoided. On the other hand, this method is only capable of making structure predictions for a protein if the structure of homologous proteins has been determined experimentally.

When the sequence is at least 40% identical the predicted structure can have an RMS error as low as 1 Å for 90% of the residues [33]. The success of comparative modeling points to the fact that the use of domain information is an important component of successful protein structure prediction.

2.4 Robotic Motion Planning

The search technique presented in this report is inspired by research is robotic motion planning. Motion planning is used to compute a collision-free path from an initial to final configuration of the robot through a set of obstacles [34]. Motion planners operate in configuration space, where the robot is represented as a

point. Every dimension in configuration space corresponds to a degree of freedom of the robot. Each point in configuration space represents either an obstacle or free space, depending on whether the matching state of the robot is in collision with the environment or not. Motion planning can then be viewed as a search problem in which the goal is to capture the connectivity of the configuration space.

Configuration space grows exponentially with the number of degrees of freedom, making path planning a difficult task. Probabilistic techniques use random sampling to make path planning tractable [35]. Random sampling produces a reliable estimate of the successful paths without a detailed examination of all possible paths. The probabilistic roadmap(PRM) technique proceeds by generating a set of random configurations. Configuration in collision are discarded. The remaining configurations represent vertices in a so-called roadmap. Edges between neighboring vertices are added into the roadmap when no obstacle exists between them. The resulting graph contains an approximation of the connectivity of free configuration space.

When covalent bonds between atoms are represented as joints, and atoms are represented as links, an amino acid sequence matches the definition of a robot [36]. The configuration of the robot is equivalent to the conformation of the protein. In the case of proteins, obstacles correspond to high energy conformations. A path through the conformation space of a protein that reaches the native state by only passing through low-energy regions can be viewed as a folding trajectory.

The application of probabilistic roadmap (PRM) techniques to proteins has been explored in the literature. A PRM planner was used as a way to capture information pertaining to pathways in the folding landscape [36, 37]. The use of the probabilistic roadmap has been extended to the Stochastic Roadmap Simulation (SRS). This technique analyzes all folding pathways concurrently [38, 39, 40]. This allows SRS to capture ensemble properties of the folding landscape.

PRM planners have also been used to generate unfolding paths of proteins [36, 37]. The native state has to be known for this method to be applicable. A roadmap of unfolding paths is computed with denser sampling in the vicinity of the native state. Knowing the native state, computational resources can be directed towards those regions of conformation space that are most likely to be important for unfolding. The ability of this algorithm to produce a large set of unrelated folding pathways provided insight into aspects of folding kinetics that could not be captured by other theoretical techniques [36].

Probabilistic roadmap methods applied to the protein folding domain have significantly improved our ability to computationally examine folding kinetics. However, roadmap strategies do not find the native state and waste significant computation time examining pathways through high energy regions.

2.5 Adaptive Sampling

Adaptive sampling in statistics is motived by the desire to sample rare, clustered populations. Adaptive sampling is also referred to as Sampling-Importance Resampling (SIR) [41, 42, 43]. To properly estimate the population size of a rare item, increased sampling density occurs in regions where the item has been located. This metaphor can also be applied to protein structure prediction: Low energy conformations in the molecular energy landscape can be considered rare items and are likely to contain other minima in close proximity. Hence, the area around low-energy conformation should be sampled more densely. Model-based search, the method proposed in this report, can be viewed as a adaptive sampling strategy whose goal it is to locate the lowest possible energy conformations, which are rare in conformation space.

3 Model-Based Search

The global search for an extremum of an arbitrary function has been proven to require an exponential number of samples [44] and consequently is intractable in high-dimensional spaces. In addition, we have seen that local approaches have no performance guarantee. Because local methods do not maintain the information obtained during search, they have no way of guiding exploration other than by purely local criteria and consequently these methods waste computational resources in local minima and by repeatedly exploring a particular region of the search space.

Model-based search is a sampling-based search method for finding extrema in high-dimensional search spaces. It is motivated by the fact that in many applications the problem domain itself as well as the information obtained during search can provide valuable insight into the relevance of regions with respect to the search task. Information from the problem domain, for example, may rule out certain portions of the search space as candidates for containing an extremum of interest. These regions should not be searched by the search method. Similarly, making only moderate assumptions about the continuity of the function we are searching, information obtained during the search can rule out the presence of the desired extremum in a particular region of the search space. As a consequence, additional exploration should be focused on other regions. Therefore, the general objective of model-based search is to exploit information from the problem domain and the problem instance to guide search space exploration in the most effective manner. To accomplish this, a model is used to compactly represent relevant information. During search, the model is used to direct further exploration. In this report we will only address the use of information obtained during the search process. Future work will be concerned with the inclusion of information from the problem domain.

Model-based search proceeds by interleaving exploration of the search space with incremental updates to a model that represents the information obtained so far. The model is used to guide exploration during each iteration. Figure 1 is a graphical illustration of the search for a global minimum in an arbitrary function using model-based search. In the following we will provide additional details about each of the steps shown in that figure.

1. Initially, the model (shown with dashed lines in the figure) contains no information about the function we are searching. If we were to incorporate information from the problem domain, it would be reflected in the model already at this stage and could provide guidance as to where to expend computational resources. The initialization of the model from domain information will be the subject of future work.

2. As the model does not provide any bias towards regions of the search space, the function is sampled uniformly at random. The resulting samples contain information about the search space. The model has to be updated to incorporate this information.

3. The model used for this illustration attempts to approximate the function at a very coarse level. Local minima among sets of neighboring samples are identified and used to approximate the function. While this is a very imprecise model of the original function, we will see later that the information suffices to significantly improve the overall search. It should be noted that the choice of the underlying representation for the model is critical in model-based search. The model has to be expressive enough to be able to guide the search, but cannot require much memory for storage, as memory limitations would quickly become prohibitive in high-dimensional space. Also, the model has to allow for a computationally efficient assessment of where future explorations should be performed.

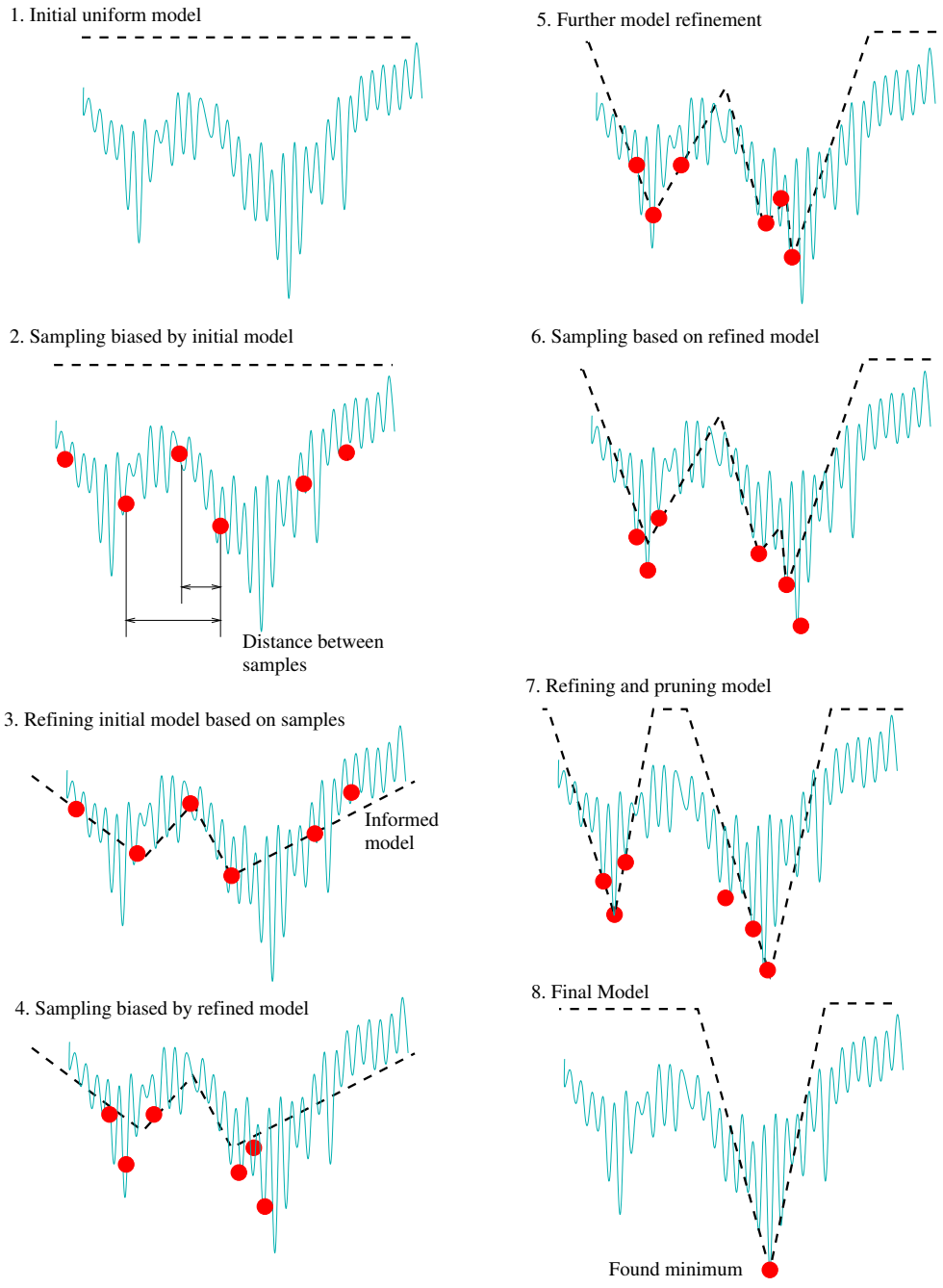


Figure 1: Model-based search: iterative refinement of an approximate model to find a global minimum.

4. Using the information contained in the model, additional samples are place in regions likely to contain a local minimum. Other regions have been eliminated from the search and will not be explored any further. Depending on the quality of the initial model, however, the ruled out regions may actually contain the desired global minimum. Therefore, an inaccurate model may prevent model-based search from finding the global minimum, implying that model-based search is an incomplete search technique. This is by design, since we know that complete search is intractable. The observation emphasizes that the quality of the model is highly relevant to the quality of the resulting search.

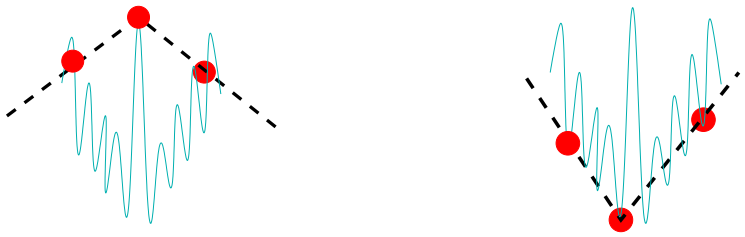
5. Based on the additional samples, the model is updated. In our example, unnecessary samples from previous iterations are discarded, reducing the memory requirements of the model representation and rendering it computationally efficient. Note that the number of local minima represented by the model depends on the samples placed during each iteration. The granularity of the model is adapted automatically.

6. Again, the model is used to guide further exploration of the search space.

7. The model is updated with the information obtained by exploration. If local minima represented in the model are assumed not to contain the global minimum, they are pruned from the model.

8. The global minimum has been identified.

From this description of model-based search, it is apparent that the quality of the model critically determines the quality of the resulting search. Figure 2 illustrates that the quality of the model also depends on the placement of samples. In Figure 2(a), samples are placed in such a way that the constructed model does not represent the underlying function at all. This is addressed by including the application of gradient descent methods into the process of placing samples. After a sample has been placed randomly, gradient descent is used to find the closest local minimum. The local minimum is used to construct the model, rather than the original sample. This procedure significantly improves the accuracy of the model, as illustrated in Figure 2(b).



(a) Bad placement of samples

(b) Good placement of samples

Figure 2: Sampling only locally low energy conformations minimizes the chance that a high energy sample from a low energy region will misinform the model.

4 Application of Model-Based Search to Molecular Energy Landscapes

Protein structure prediction can be viewed as search for the global minimum in a molecular energy landscape. The molecular energy landscape of a protein is assumed to be shaped like a high-dimensional funnel with the native state of the protein situated at the global minimum of the funnel [45, 46, 47]. An illustration of a three-dimensional funnel is shown in Figure 3. The energy landscape contains many local minima and consequently does not lend itself well to search using purely local methods. In this section we will describe how model-based search can be applied to protein structure prediction by searching molecular energy landscapes for global minima.

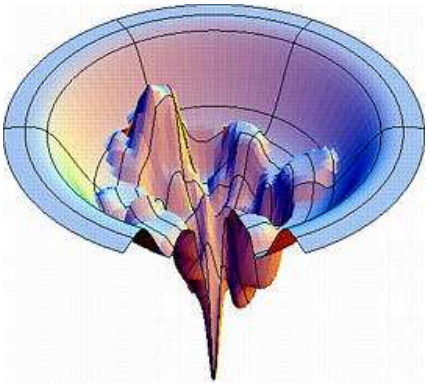


Figure 3: An illustration of the molecular energy landscape of a protein. The native state is represented by the global minimum of this funnel-like landscape. The walls of the funnel exhibit numerous local minima. The process of protein folding can be viewed as gradient descent in the molecular energy landscape, the process of protein structure prediction can bee seen as the search for the global minimum of the landscape. This image is from a paper by Hue Sun Chan and Ken A. Dill [46].

To perform protein structure prediction, an extensive software infrastructure is required. Since the focus of our work is on the search for minima in high-dimensional conformation spaces, we integrated an implementation of model-based search with Rosetta [9].

4.1 Rosetta

The name Rosetta [9] refers to a large suite of software tools related to protein structure and protein folding. Here, we will only refer to the Rosetta protein structure prediction software. More specifically, we will only be concerned with *ab initio* protein structure prediction, i.e., without the use of homology information. Over the last several years, Rosetta has repeatedly been demonstrated to outperform other computational methods of protein structure prediction in this category [31].

One of the fundamental building blocks of Rosetta is the energy function. Designed to approximate the true energy function of proteins, it computes an estimate of the energy for a particular protein conformation based on knowledge about interactions between portions of the backbone, side chains, and solvent. The design of this function was guided by observations from experimentally determined native structures [9]. During the search process performed by Rosetta, this energy function is used to evaluate the quality of generated conformations. The conformations that are considered prediction candidates for the native structure of the protein are called decoys.

The search for the global minimum in the energy landscape, described by Rosetta's approximate energy function, begins with an arbitrarily initialized protein structure. This structure is incrementally refined using a Monte Carlo insertion strategy to replace the angles of short fragments from the decoy with ones retrieved from a fragment library. The fragment library contains fragments of other proteins for which the structure has been determined experimentally. For fragment replacement, preference is given to low energy fragments with similar sequence, therefore increasing the likelihood of an insertion resulting in an overall low energy protein structure. Because the candidate fragments for replacement occur in nature, it is assumed that they represent an energetically favorable conformation.

As search progresses the size of the fragments to be replaced is reduced, Monte Carlo becomes increasingly restrictive in the acceptance of states that increase the energy, and a more complete and precise energy function is used. As a result, the step size of the Monte Carlo algorithm is reduced in low energy regions.

Rosetta distinguishes the candidate native state from other states using a clustering procedure. We do not investigate this aspect of Rosetta since our interests lie in the improvement of the search strategy.

4.2 Integration of Model-Based Search with Rosetta

We modified Rosetta to devise a protein structure predictor based on model-based search. To do so, the search used in Rosetta was replaced with a model-based search algorithm. Other elements of Rosetta, such as the energy function, remained unchanged. In this section we will detail how the general description of model-based search given in Section 3 was instantiated to yield a practical implementation. We will describe the underlying model used to represent information obtained during the search and we will explain how this information is used to guide future exploration through the process of sampling.

As we discussed before, the model used in model-based search will critically impact the quality of the overall search. In the preliminary implementation of model-based search described in this report, we chose a very simple model. This model maintains information about regions of the search space that might contain the global minimum. These regions are determined based on samples taken from conformation space. Each of these samples corresponds to a particular conformation with an energy value determined by Rosetta's energy function. If a sample has lower energy than its nearest neighbors, we consider the associated regions to contain a local, if not the global minimum.

Each of these regions is represented by the sample of locally minimal energy and its nearest neighbors. To guide future exploration of the search space, we associate a score with each region. This score is determined by the energy of the sample with locally minimal energy and by an estimate of the size of the local minimum, determined by the distance to its neighbors. This score will be used to determine how much exploration should be dedicated to a particular region during the ongoing search process. It biases exploration towards larger regions, because they are relatively unexplored, and towards regions with lower energy, because they are more likely to contain the global minimum. The scoring function used for the experiments described in Section 5 weighs the well size more heavily to bias exploration towards regions underrepresented in the model. While we have not explored the parameter space of our scoring function, this intuitively motivated choice performs well in practice.

During the search process, the information represented in the model is used to guide exploration. Since exploration is performed by sampling, the number of samples generated in each region during a particular iteration of the model-based search algorithm is proportional to the score of that region. These samples should be generated inside the region specified by the model.

In the case of protein structure prediction, the process of generating samples is rather difficult. If samples are generated randomly, the probability of sampling a low-energy conformation are almost zero. Nearly all samples will contain steric clashes that result in very high energy values. Therefore, we carry out sam-

pling based on the fragment replacement strategy employed by Rosetta in its first phase of the search. The fragment replacement routine is initialized with the conformation of locally minimal energy stored in the model. A series of random segment replacements then generates an alternative conformation in proximity to the initial conformation. The result is used as a sample to refine the model. By using Rosetta's fragment replacement strategy, the negative effects of placing locally bad samples shown in Figure 2(a) can be avoided.

4.3 Step-By-Step Description of Model-Based Search

This section illustrates how the components of model-based search described in this section are combined to result in an efficient search method for molecular energy landscapes. The description of model-based search given below refers to the graphical illustration of the method shown in Figure 4.

1. Starting with an arbitrary conformation for the protein, Rosetta's fragment replacement procedure is used to generate a set of initial samples. The Monte Carlo fragment insertion strategy used by Rosetta places samples only in local minima, which ensures maximum accuracy of the model.
2. Nearest neighborhood relations for the samples are computed.
3. After conformations with locally minimal energy have been identified, an approximate model of the energy function is computed. This model encodes the conformations of locally minimal energy and an estimate of the size of the associated well in the function. The size is estimated based on proximity and energy level its neighbors. For each well represented in the model, a score based on the size and energy level of the well is computed.
4. To reduce the memory requirements of model-based search, the model is pruned. Only wells with high scores are maintained.
5. The model is refined so as to be able to guide future exploration in the most efficient manner. This is accomplished by reducing the size of the wells, resulting in an exploration of the area around the conformation of locally minimum energy.
6. Rosetta's fragment replacement approach is initialized with the conformation of locally minimal energy. The conformations generated after repeated fragment replacements are used as samples.
7. The steps described here are repeated until no progress towards lower-energy states can be made, or until the allotted computational resources have been exhausted.

This iterative process interleaves model refinement based on information obtained from exploration with sampling based on information contained in the model; it focuses computational resources on regions likely to contain the global minimum. Model-based search avoids global, exhaustive search of the search space by exploiting information obtained during the search process itself. It avoids the computational requirements associated with global methods, and eliminates the shortfalls of purely local methods.

5 Experimental Results

The predictions from the implementation of model-based search (MBS) described in Section 4 were compared with predictions obtained using Rosetta [9]. We used nine proteins, varying in size from 60 to 414 amino acids. The experimentally determined native structure of these proteins is used in the comparison.

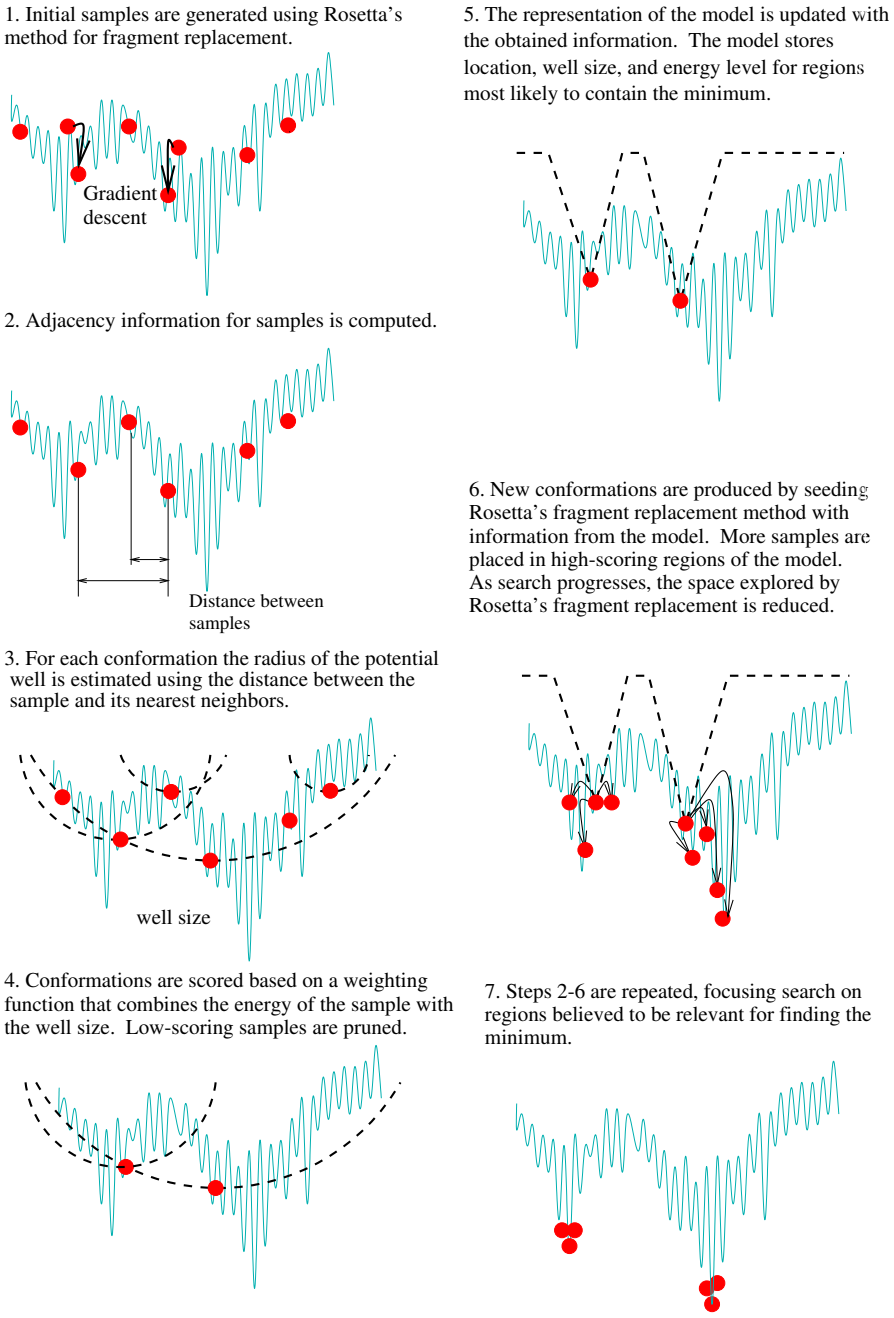


Figure 4: An implementation of model-based search in a molecular energy landscape of a protein for protein structure prediction. The implementation integrates model-based search with Rosetta [9].

For each prediction both algorithms produced 500 decoys (conformations of the protein considered candidates for the predicted native state). To generate these decoys, both algorithms were given an equal amount of computational resources. The results presented here show that MBS consistently outperforms the search strategy implemented by Rosetta, producing decoys with much lower energy (where the energy is determined using Rosetta's energy function). As dimensionality increases, the performance advantage of MBS becomes increasingly pronounced (Figures 5 and 8).

While decoys generated using MBS were consistently much lower in energy than the ones generated with the local search method used by Rosetta, this did not translate to the decoys being significantly closer to the experimentally determined native structure of the protein (Figure 7). This is reflected in the fact that the predictions using MBS do not achieve a significant reduction in RMSD. This can be attributed to two factors:

1. The energy function provided by Rosetta only represents an approximation to the true energy function of the protein. This is confirmed by the fact that for proteins with 60-216 amino acids MBS was able to find states with lower energy than the native state. Since it is generally assumed that the native state corresponds to the global minimum of nature's energy function, Rosetta's energy function must be inaccurate.

2. RMSD is not an accurate similarity measure between decoys. For example, two decoys with shared substructures that are hinged at one joint causing them to become shifted in space would be considered rather dissimilar, whereas two completely dissimilar decoys that spatially occupy a small region would be considered similar. In the future, we intend to explore the use of alternate similarity measures for proteins, such as VAST [48], DALI [49], CE [50], ProSup [51], or LGA [52].

Due to above arguments, the results presented here should be interpreted only along the dimension of energy. This dimension suffices to demonstrate the effectiveness of MBS, since we are only concerned with finding states of low energy within the energy landscape. By using more accurate energy functions (admittedly a tremendous challenge by itself), MBS will be able to make predictions closer to the native state.

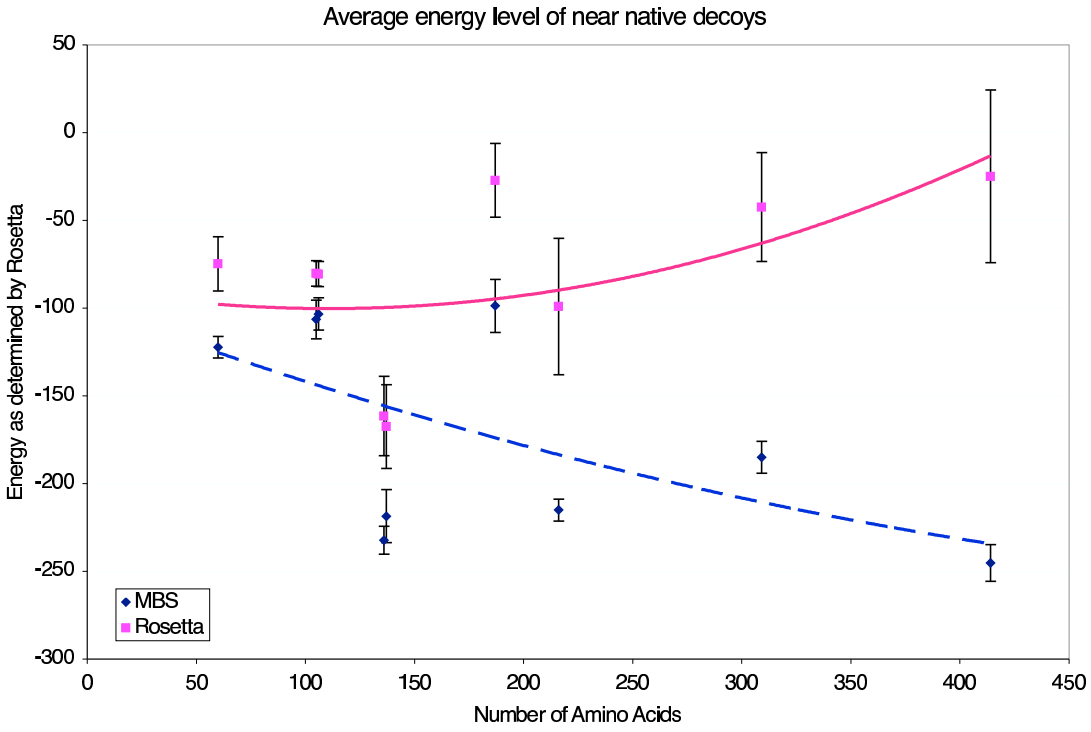


Figure 5: A comparison of the average energy of decoys produced by MBS and Rosetta for proteins of different sizes. It can be seen that model-based search outperforms the Monte Carlo method implemented by Rosetta by a large margin. The performance advantage increases as the length of the proteins increases. The trend lines indicate that this performance gap continues to widen for longer proteins.

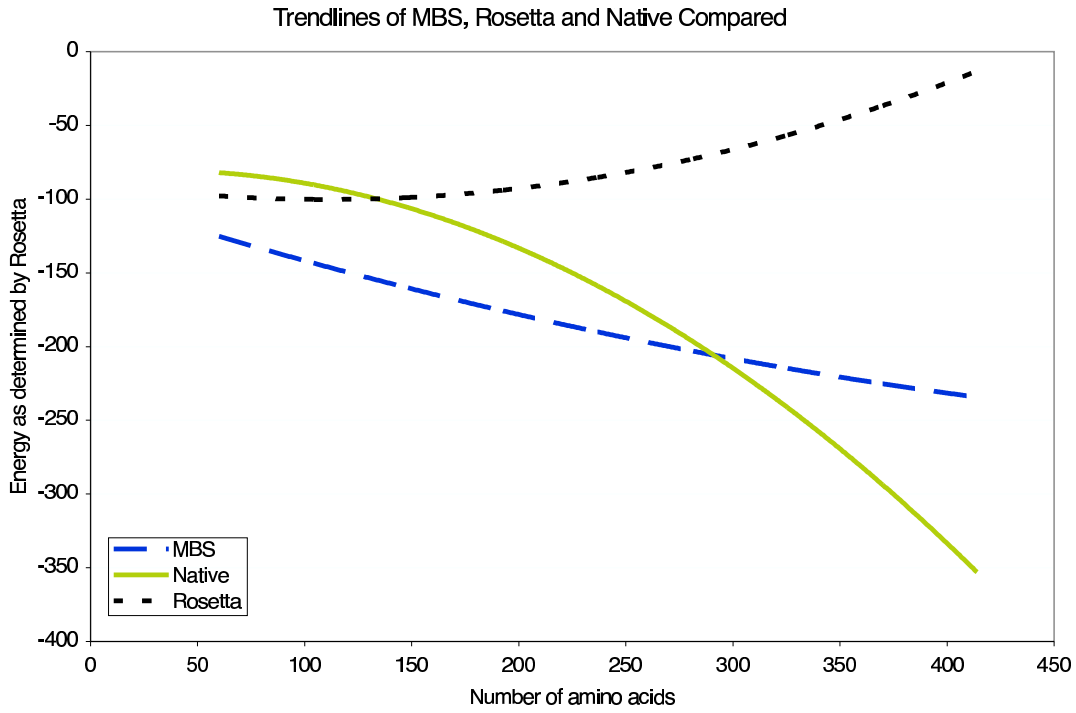


Figure 6: A comparison of the average energy of decoys produced by MBS and Rosetta, and the energy of the experimentally determined native structure as determined by the energy function used for these experiments. For short proteins, model-based search finds lower-energy decoys than the energy of the native structure. This points to inaccuracies in the energy function, since the native state should represent the global minimum of the energy function.

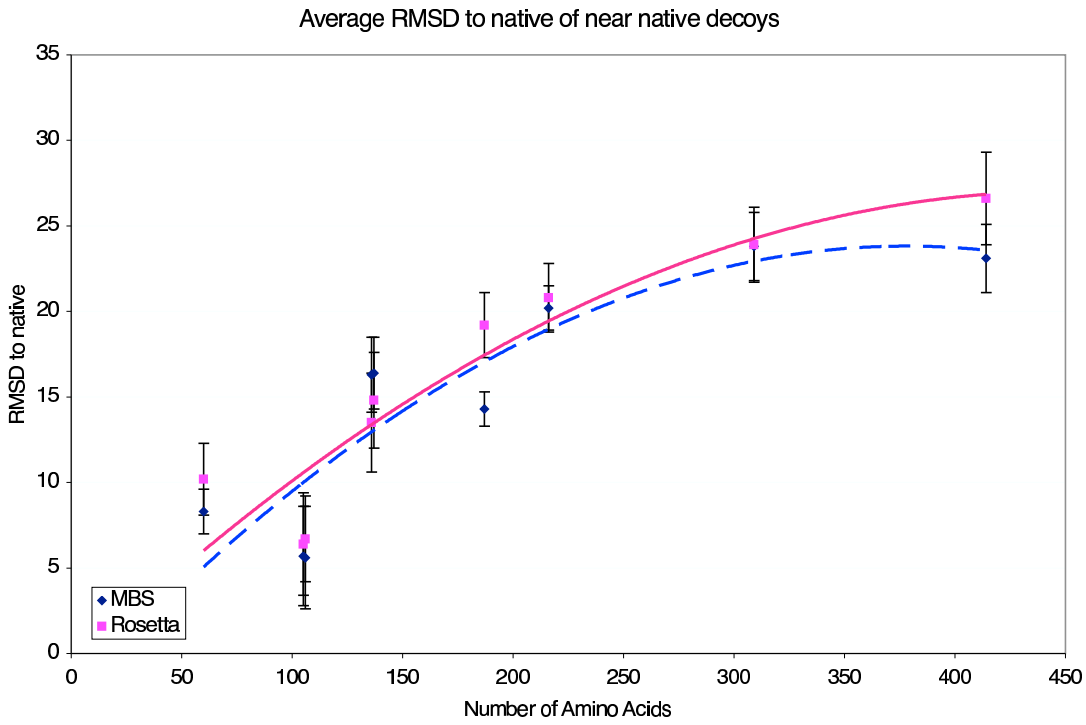
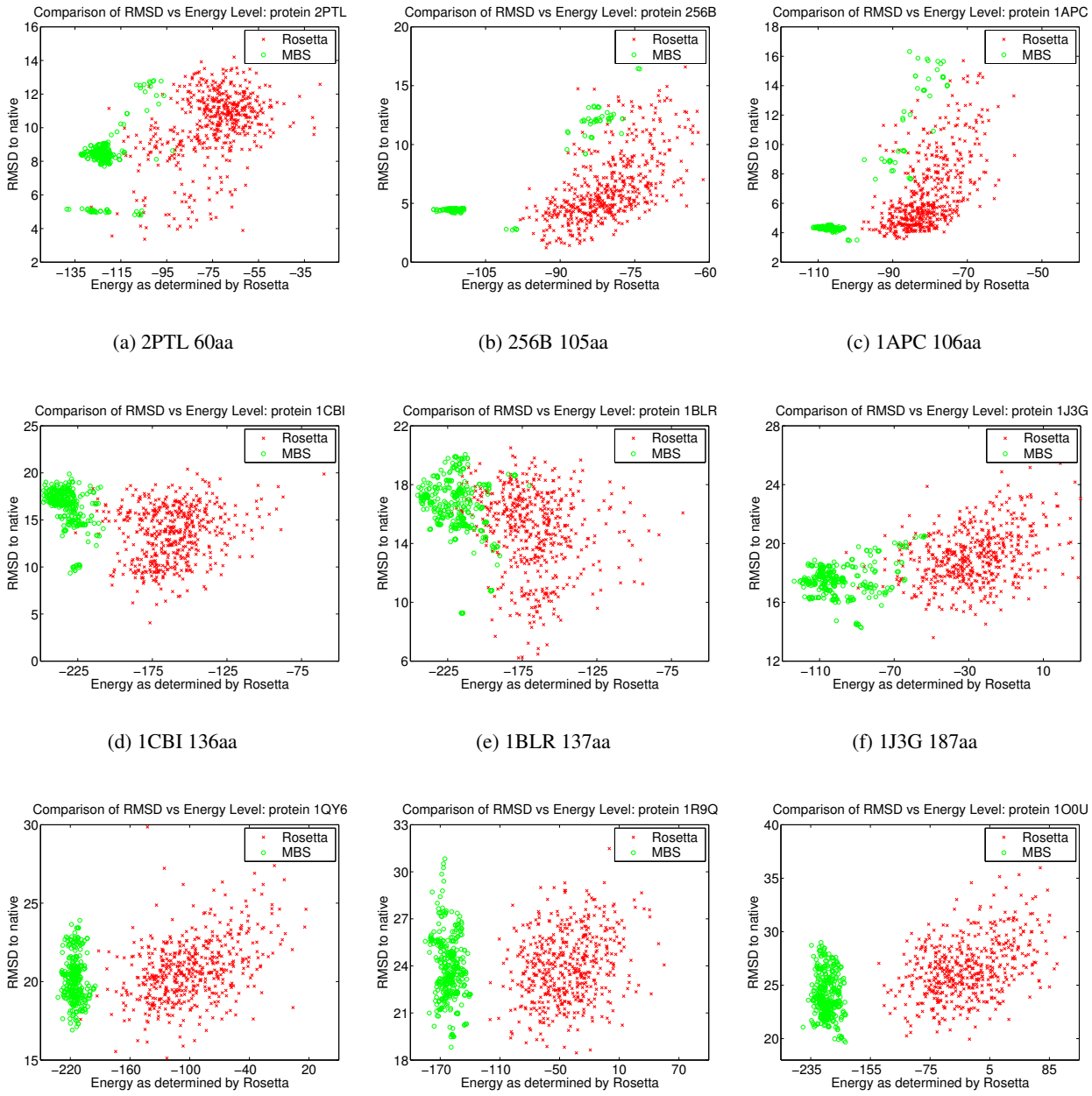


Figure 7: A comparison of RMSD to the native structure for decoys produced by MBS and Rosetta. Only the decoys produced by MBS for long proteins have significantly lower RMSD that those produced by Rosetta. This points to inaccuracies in the energy function and shows that RMSD may not be the correct metric to evaluate similarity between protein structures.



(a) 2PTL 60aa (b) 256B 105aa (c) 1APC 106aa

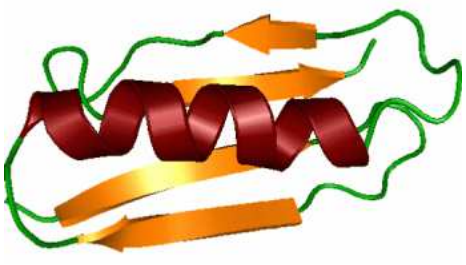
(d) 1CBI 136aa (e) 1BLR 137aa (f) 1J3G 187aa

(g) 1QY6 216aa (h) 1R9Q 309aa (i) 100U 414aa

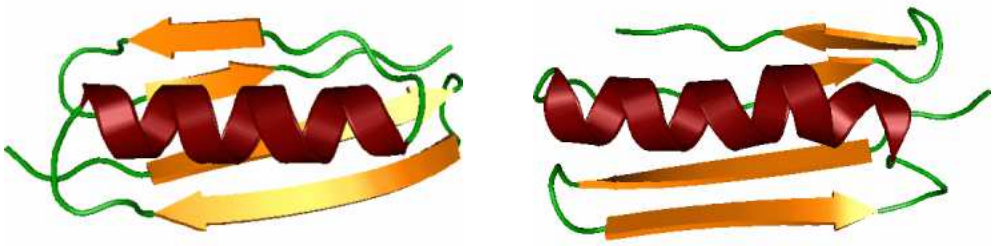
Figure 8: These graphs compare the 500 near native decoys produced by model-based search with those produced by Rosetta. Starting in the upper left with the smallest protein, it is easily visible that decoys produced using MBS are lower in energy then those found by Rosetta.

5.1 2PTL Protein

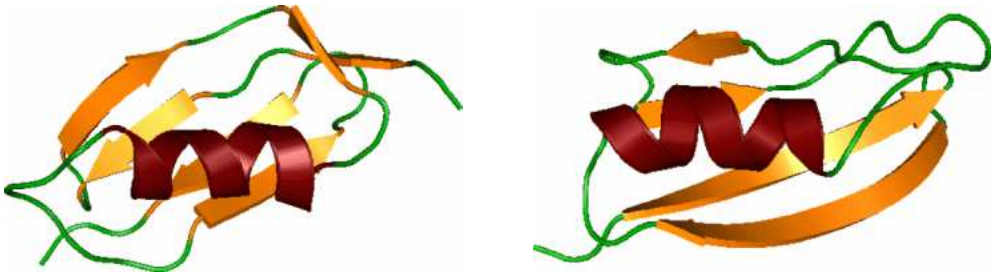
2PTL represents the smallest protein in this study. Figure 8(a) shows that samples generated by MBS are significantly lower in energy then those produced by Rosetta. MBS decoys appear to be more ordered than those of Rosetta (Figure 9(b), 9(c)). MBS generates two low-energy clusters at 5 and 8 RMSD to the native structure. In these clusters, the β-sheet is located on opposite sides of the α-helix. Note how the two MBS decoys are almost mirror image of one another, with exception of a reversal of the top two strands of the β-sheet(Figure 9(b)).



(a) Native



(b) Model-Based Search

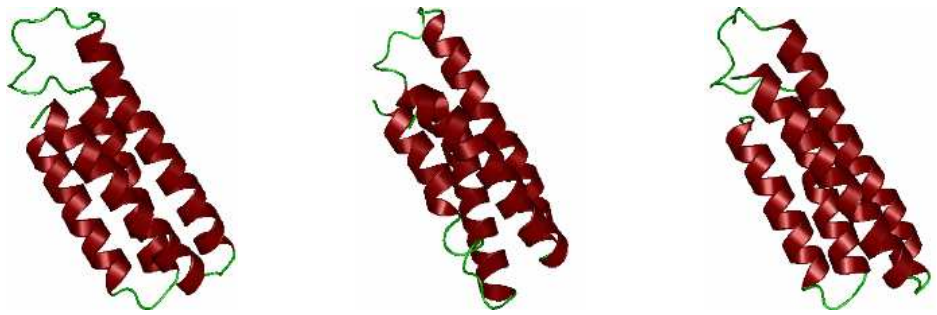


(c) Rosetta

Figure 9: 2PTL is a 60 amino acid binding protein from the Immunoglobulin L Chain. The native state was determined by Wilkstroem et al. using NMR [53].

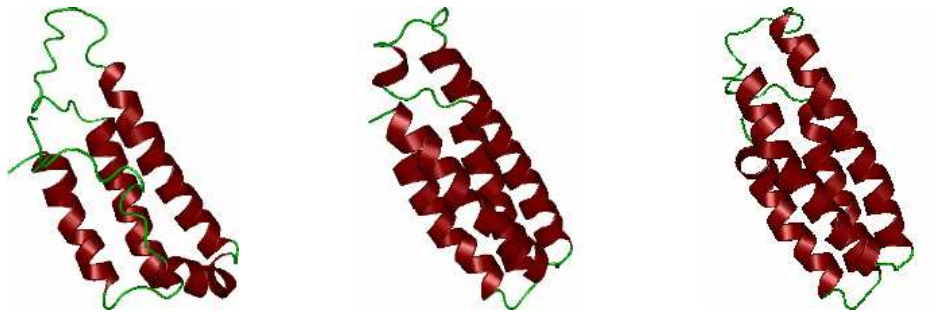
5.2 256B and 1APC Proteins

Similar to 2PTL, the energy levels of decoys produced by MBS are significantly lower then those produced by Rosetta. However, this does not translate into decoys significantly closer to the native state. We claim this is caused by inaccuracies in the energy function. The native states of 256B and 1APC have energy levels of -70.0 and -33.1 , while the lowest energy decoys found by MBS are -115.5 and -111.3. One possible reason for this is that the energy function may favor interactions local to the α-helix over global interactions. The low-energy decoys produced by MBS seem to have longer α-helices with less alignment between them than the native state (Figure 10, 11).



(a) Native (b) Model-Based Search (c) Rosetta

Figure 10: Protein 256B is a 105 amino acid protein form of cytochrome B562. The native state of 256B was determined by Lederer et al using x-ray crystallography [54].

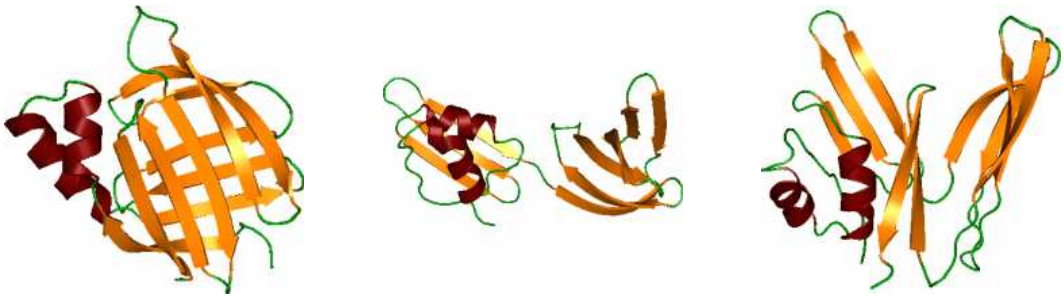


(a) Native (b) Model-Based Search (c) Rosetta

Figure 11: Protein 1APC is a 106 amino acid protein form of cytochrome B562. The native state of 1APC was determined by Wand et al. [55].

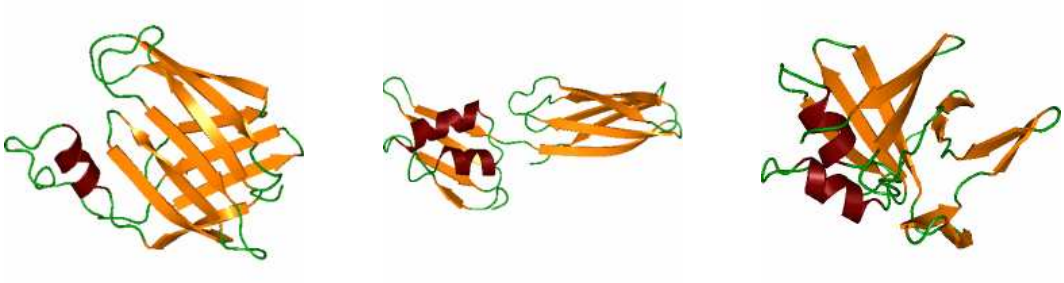
5.3 1CBI and 1BLR Proteins

The native state for both 1CBI and 1BLR has a β-sheet domain attached to a smaller α-helix domain (Figure 12(a), 13(a)). As with 256B and 1APC, the native state is much higher in energy then the decoys generated by MBS. The native states of 1CBI and 1BLR have energy levels of -188.5 and -123.1, while the lowest energy decoys found by MBS are -250.9 and -245.2 The low energy decoys for both proteins incorrectly associate some of the β-sheets with the α- helix (Figures 12(b), 13(b)). Low energy decoys produced by MBS for both 1CBI and 1BLR have very similar structure, while those produced by Rosetta have diverse structure that appears disordered (Figure 12(b), 12(c), 13(b) and 13(c)). Combining these insights indicates that there is likely a systematic error in the energy function that rates interactions between α-helices and β-sheets too strongly.



(a) Native (b) Model-Based Search (c) Rosetta

Figure 12: Protein 1CBI is a 136 amino acid retinoic acid binding proteins. The native state was determined by Thompson et al. [56].

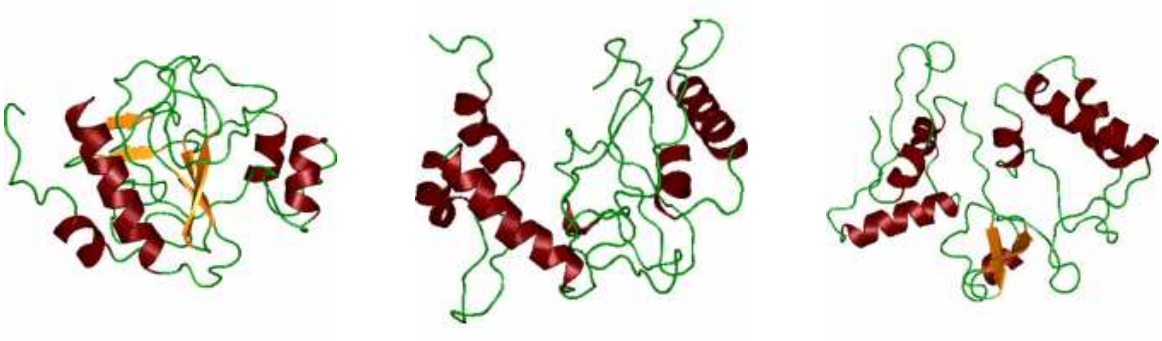


(a) Native (b) Model-Based Search (c) Rosetta

Figure 13: Protein 1BLR is a 137 amino acid retinoic acid binding proteins. The native state was determined by Wang et al. [57].

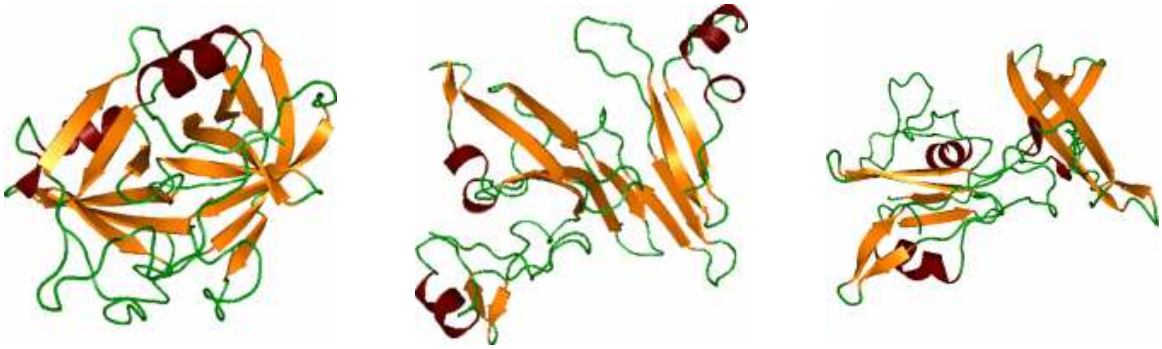
5.4 1J3G and 1QY6 Proteins

As the dimensionality of the proteins continues to increase, it becomes harder to visually distinguish inaccuracies in the fold. Decoys produced by MBS appear to have more compact cores than those of Rosetta (Figure 14(b), 14(c), 15(b) and 15(c)). Increased dimensionality also results in almost all MBS decoys to be of lower energy than any Rosetta decoy (Figure8(f), 8(g)).



(a) Native (b) Model-Based Search (c) Rosetta

Figure 14: Protein 1J3G is a 187 amino acid hydrolase. The native state was determined by Liepinish et al. [58].



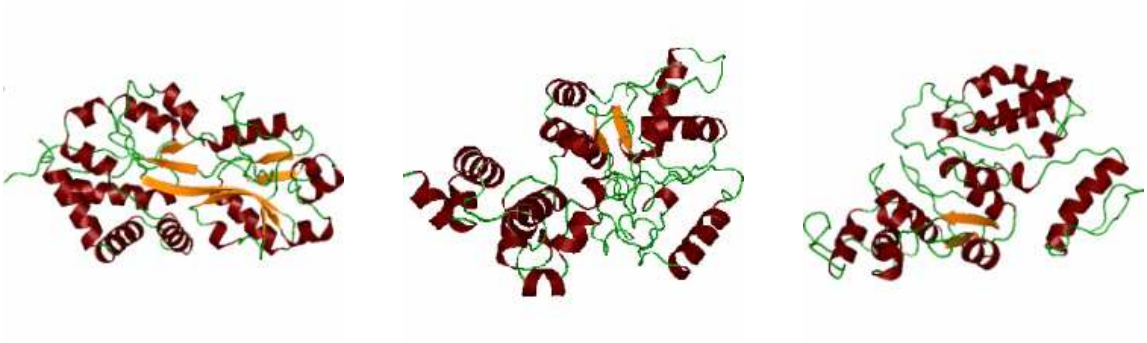
(a) Native (b) Model-Based Search (c) Rosetta

Figure 15: Protein 1QY6 is a 216 amino acid protease from Staphylococcus Aureus. The native state was determined by Prasad et al. [59].

5.5 1R9Q and 1O0U Proteins

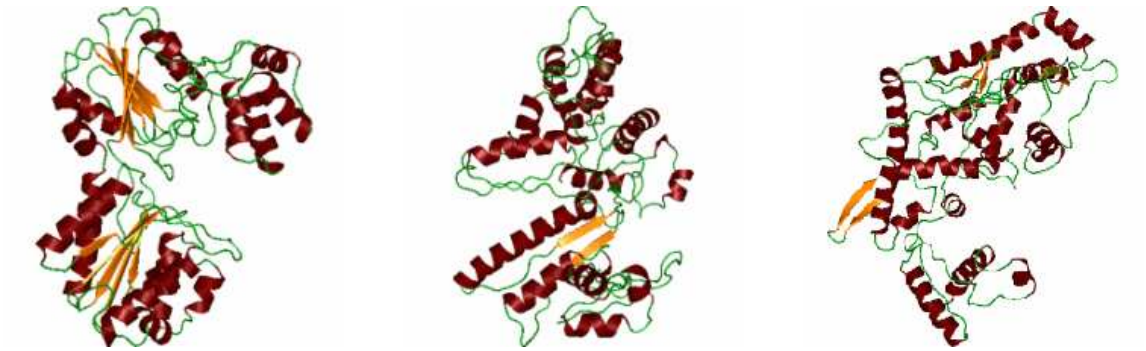
For the largest proteins in our study (1R9Q and 1O0U), the highest energy MBS decoy is lower in energy then the lowest energy Rosetta decoy. (Figure 8(h), 8(i)). Through visual inspection it is hard to say much about the final structures of the proteins, but cores of the MBS decoy look more compact.

For 1R9Q and 1O0U the native state is lower in energy than most if not all the decoys. If we assume the trends from smaller proteins apply here, then most likely many conformations exist with much lower energy then native. Since these low energy states are not being found, MBS is not fully searching the space.



(a) Native (b) Model-Based Search (c) Rosetta

Figure 16: The native state of, the 309 amino acid protein, 1R9Q was determined by Scheifner et al. [60].



(a) Native (b) Model-Based Search (c) Rosetta

Figure 17: The largest protein in our study, 414 amino acid 1O0U, had it's native state determined by a research group at the Joint Center For Structural Genomics center [61].

5.6 The Effect of the Model in Model-Based Search

The underlying model used to represent the information obtained during search is of critical importance to the quality and efficiency of model-based search. This is demonstrated by experiments with different underlying models. We compared the model described in Section 4 with a model that only maintains the minimal energy samples, and a model that only maintains the radius of the relevant regions, but not its energy level. A model solely based on energy level causes MBS to degenerate to a form of adaptive beam search, whereas a model based on the radius effectively performs uniform sampling in relevant regions by adapting the number of samples to the size of the region. The comparison of the decoys generated by MBS with the respective models is shown in Figure 18. Not surprisingly, the most expressive model results in the lowest-energy decoys. This result provides motivation to explore alternative models in the future.

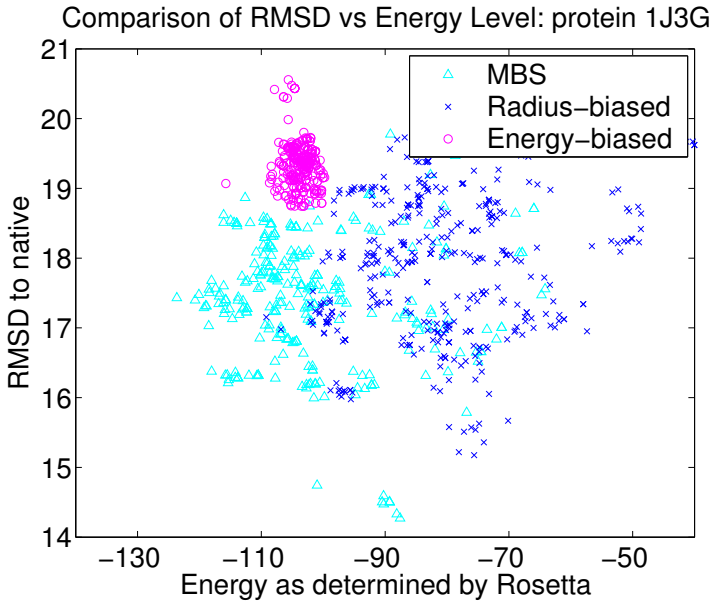


Figure 18: This graph shows that model-based search is sensitive to the model chosen to represent local features of the energy function. A model that incorporates both the energy level of a region as well as an approximation to the size of the region achieves the best results (labeled MBS). A model based solely on energy level causes MBS to degenerate to a form of adaptive beam search (labeled energy-biased) and results in susceptibility to local minima. Biasing the search in the favor of samples with high radius (labeled radius-based) does not focus computation on the most promising regions of the energy function.

6 Conclusions and Future Work

Search for the global extremum in a high-dimensional space will always be a challenging problem. We feel the best solution will require the integration of all possible sources of information from the problem domain and the particular problem instance to reduce the amount of search space that has to be explored. Model-based search (MBS) lays out a method for the integration of domain knowledge with information generated during search. In this report we only demonstrated how model-based search can exploit the information

obtained during search, i.e., the particular problem instance. By exploiting this information, MBS was significantly better at finding low energy states in a molecular energy landscape than the Monte Carlo-based technique implemented in Rosetta, currently considered to be the leading protein structure predictor. This was demonstrated by predicting the structure of nine proteins, ranging in size between 60 and 414 amino acids, for which the native structure had been determined experimentally.

The results also indicate that the performance of MBS degrades in higher-dimensional search spaces. While this degradation occurs much slower than with other search methods, it is apparent that a near-optimal extremum cannot be found in those spaces. We attribute this to the fact that the amount of information obtained during the search process is not sufficient to focus search on relevant regions. We will therefore explore the inclusion of domain information to initialize the model. This additional information should result in an additional reduction of the search space and consequently significantly improve the performance of model-based search.

Furthermore, we have shown that the expressiveness of the model used in model-based search can have a significant effect on the quality and efficiency of the search. We will investigate various models, including models that apply dimensionality reduction techniques, to improve the performance of the proposed implementation of model-based search.

Model-based search identifies structures with energy levels far below the energy level of the native structure (energies have been determined based on the energy function of Rosetta). The conclusion of these findings must be that the energy function is not accurate. We would like to use model-based search to evaluate the quality of several energy functions in order to investigate and potentially reduce their inaccuracies.

Another outcome of our experiments was the insight that RMSD is a rather bad indicator of structure similarity in proteins. This affects the quality of the model built by MBS during search. Future work will explore whether better distance metrics can increase the quality of information used for building the model. Candidates for this investigation are: VAST [48], DALI [49], CE [50], ProSup [51], or LGA [52].

In addition, we intend to explore the application of model-based search to other high-dimensional problem domains, such as structure learning in Bayesian Networks [62]. In these different applications, we foresee difficulty in the creation of good distance metrics to determine nearest neighbors. Distance metrics other than Euclidean distance will have to be explored, as determining the nearest neighbor in Euclidean space becomes increasingly meaningless in high-dimensional spaces [63].

In the future, we expect search methods for practical problems in high-dimensional spaces to include various sources of information to reduce the amount of search required. Model-based search can be viewed as a general framework for accomplishing this objective. The empirical evidence presented in this report shows that even simple models can have a significant effect on the performance of search in high-dimensional spaces.

Acknowledgments

We would like to acknowledge the help of Carol Rohl [9] at the University of California Santa Cruz and Mehmet Serkan Apaydin [39] at Stanford University. We thank both of them for making their software available available to us and for providing valuable suggestions during the implementation phase of model-based search. We would also like to thank Lila Gierasch, Arnie Hagler, Ken Rotondi, and Marc Vogt from the Department of Biochemistry and Molecular Biology at the University of Massachusetts Amherst for their insightful comments.

References

[1] Protein Data Bank, "http://www.rcsb.org/pdb/holdings.html," July 23 2004.

[2] National Center for Biotechnology Information, "http://www.ncbi.nlm.nih.gov/refseq/," July 23 2004.

[3] K. Pruitt and D. Maglott, "Refseq and locuslink:ncvi gene-centered resources," *Nucleic Acids Research*, 2001.

[4] K. Wüthrich, *NMR of Proteins and Nucleic Acids*. John Wiley & Sons, 1986.

[5] ——, "Biological crystallography," *Acta Cristallographica Section D*, vol. 51, p. 249, 1995.

[6] C.-I. Branden and J. Tooze, *Introduction to Protein Structure*, 2nd ed. Garland Publishing, 1999.

[7] D. Voet and J. G. Voet, *Biochemistry*, 2nd ed. John Wiley & Sons, 1995.

[8] C. Levinthal, "Are there pathways for protein folding?" *Journal de Chimie Physique*, vol. 65, no. 1, pp. 44–45, 1968.

[9] C. A. Rohl, C. E. M. Strauss, K. M. S. Misura, and D. Baker, "Protein structure predictionusing rosetta," *Methods in Enzymology*, vol. 383, pp. 66–93, 2004.

[10] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 2nd ed. Prentice Hall, 2002.

[11] J. M. Hammersley and D. C. Handscomb, *Monte Carlo Methods*. Methuen and Co., Ltd., 1964.

[12] N. Metropolis and S. Ulam, "The monte carlo method," *Journal of the American Statistical Association*, 1949.

[13] S. Russell and P. Norvig, *Artificial Intelligence A Modern Approach, Second Edition*. Pearson Education Inc., 2003.

[14] S.Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," *Science*, 1983.

[15] B. T. Lowerre, "The harpy speech recognition system," Ph.D. dissertation, Carnegie Mellon University, 1976.

[16] J. Holland, "Adaptation in natural and artificial systems," *Artificial Intelligence*, 1975.

[17] M. Levitt, "Molcular dynamics of native protein – I. Computer simulation of trajectories," *Journal of Molecular Biology*, vol. 168, pp. 595–620, 1983.

[18] ——, "Molcular dynamics of native protein – II. Analysis and nature of motion," *Journal of Molecular Biology*, vol. 168, pp. 621–657, 1983.

[19] C. D. Snow, H. Nguyen, V. S. Pande, and M. Gruebele, "Absolute comparision of simulated and experimental protein-folding dynamics," *Nature*, vol. 6911, no. 40, pp. 102–106, 2002.

[20] W. F. van Gunsteren, "Computer simuation by molecular dynamics as a tool for modelling of molecular systems," *Molecular Simulation*, vol. 3, pp. 187–200, 1989.

[21] W. Wang, O. Donini, C. M. Reyes, and P. A. Knollman, "Biomolecular simulations: Recent developents in force fields, simulations of enzyme catalysis, protein-ligand, protein-protein, and protein-nucleic acid nonconvalent interactions," *Annual Reviews Biophysics and Biomolecular Structure*, vol. 30, pp. 211–243, 2001.

[22] J. Lee, "New monte carlo algorithm: Entropic sampling," *Physical Review Letters*, 1993.

[23] N. Metropolis, "The beginning of the Monte Carlo method," *Los Alamos Science*, vol. 21, pp. 1087–1092, 1987.

[24] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. N. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *Journal de Chimie Physique*, vol. 21, pp. 1087–1092, 1954.

[25] Y. Okamoto, "Protein folding problem as studied by new simulation algorithms," *Recent Research Developments in Pure & Applied Chemistry*, vol. 1, 1998.

[26] D. D. Frantz, D. L. Freeman, and J. D. Doll, "Reducing quasi-ergodic behavior in Monte Carlo simulations by j-walking: Applications to atomic clusters," *Journal of Chemical Physics*, vol. 93, no. 4, pp. 2769–2784, 1990.

[27] B. A. Berg and T. Neuhaus, "Multicanoncial ensemble: A new approach to simulate first-oder phase transitions," *Physical Review Letters*, vol. 68, no. 1, pp. 9–12, 1992.

[28] H. Xu and B. J. Berne, "Mulitcanonical jump walking: A method for efficiently sampling rouch energy landscapes," *Journal of Chemical Physics*, vol. 110, no. 21, pp. 10 299–10 306, 1999.

[29] D. T. Jones and J. M. Thornton, "Potential energy functions for threading," *Current Opinion in Structural Biology*, vol. 6, pp. 210–216, 1996.

[30] A. R. Leach, *Molecular Modelling – Principle and Applications*, 2nd ed. Prentice Hall, 1991.

[31] J. Moult, K. Fidelis, A. Zemla, and T. Hubbard, "Critical assessement of methods of protein structure prediction (CASP)-round V," *Proteins: Stucture, Function and Genetics*, vol. 53, no. S6, pp. 334–339, 2003.

[32] P. Bradley, D. Chivian, J. Meiler, K. M. S. Misura, C. A. Rohl, W. R. Schief, W. J. Wedemeyer, O. Schueler-Furmann, P. Murphy, J. Schonbrun, C. E. M. Strauss, and D. Baker, "Rosetta predictions in CASP5: Successes, failures and prospects for complete automation," *Proteins: Stucture, Function and Genetics*, vol. 53, pp. 457–468, 2003.

[33] A. Šali, L. Potterton, F. Yuan, H. van Vlijmen, and M. Karplus, "Evaluation of comparative protein modeling by modeller," *Protein*, 1995.

[34] J.-C. Latombe, *Robot Motion Planning*. Boston: Kluwer Academic Publishers, 1991.

[35] L. E. Kavraki, P. Švestka, J.-C. Latombe, and M. H. Overmars, "Probabilistic roadmaps for path planning in high-dimensional configuration spaces," *IEEE Transactions on Robotics and Automation*, vol. 12, no. 4, pp. 566–580, 1996.

[36] N. Amato and G. Song, "Using motion planning to map protein folding landscapes and analyze folding kinetics of known native structures," Department of Computer Science at Texas A and M University, Tech. Rep. TR01-001, October 15th,2001.

[37] G. Song, S. Thomas, K. Dill, J. M. Schjoltz, and N. M. Amato, "A path planning-based study of protein folding with a case study of hairpin formation in protein g and l," in *Proceedings of the Pacific Symposium of BioComputing*. PSB2003, Jan 2003.

[38] M. Apaydin, A. P. Singh, D. L. Brutlag, and J.-C. Latombe, "Capturing molecular energy landscapes with probabilistic conformational roadmaps," *International Conference on Robotics and Automation*, 2001.

[39] M. S. Apaydin, D. Brutlag, C. Guestrin, D. Hsu, J.-C. Latombe, and C. Varma, "Stochastic roadmap simulation: An efficient representation and algorithm for analyzing molecular motion," Stanford University, Tech. Rep., 2002.

[40] M. S. Apaydin, D. Brutlag, C. Guestrin, D. Hsu, and J.-C. Latombe, "Stochastic roadmap simulation: An efficient representation and algorithm for analyzing molecular motion," *In Proc. ACM Int. Conf. on Computational Biology(RECOMB)*, vol. International Conference on Research in Computational Molecular Biology, pp. 12–21, 2002.

[41] D. B. Rubin, "Using the SIR algorithm to simulate posterior distributions," *Bayesian Statistics 3: Edited by J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F. M. Smith*, 1988.

[42] A. Smith and A. Gelfand, "Bayesian statistics without tears: A sampling-resampling perspective," *American Statistics*, 1992.

[43] S. Thompson and G. Seber, *Adaptive Sampling*. Wiley series in probability and statistics, 1996.

[44] A. G. Sukharev, "Optimal strategies of the search for an extremum," *U.S.S.R. Computational Mathematics and Mathematical Physics*, vol. 11, no. 4, pp. 119–137, 1971, translated from Russian, *Zh. Vychisl. Mt. i Mat. Fiz.*, 11(4):910-924.

[45] J. Bryngelson, J. Onuchic, N. Socci, and P. Wolynes, "Funnels, pathways and the energy landscape of protein folding: A synthesis," *Proteins*, 1995.

[46] H. S. Chan and K. A. Dill, "Protein folding in the landscape perspective: Chevron plots and non-arrhenius kinetics,," *Proteins: Structure, Function, and Genetics*, 1998.

[47] A. Šali, E. Shaknovich, and M. Karplus, "How does a protein fold," *Nature*, 1994.

[48] J.F.Gibrtat, T. Madej, and S. Bryant, "Suprising similarities in structure comparison," *Current Opinion in Structural Biology*, 1996.

[49] L. Holm and C. Sander, "Protein structure comparison by alignment of distance matrices," *Journal of Molecular Biology*, 1993.

[50] I. Shindyalov and P. Bourne, "Protein structure alignment by incremental combinatorial extension(ce) of the optimal path," *Protein Engineering*, 1998.

[51] Z. Feng and M. Sippl, "Optimum superimposition of protein structures: ambiguities and implications," *Fold Des.*, 1996.

[52] A. Zemla, "LGA: a method for finding 3d similarities in protein structure," *Nucleic Acid REsearch*, 2003.

[53] M. Wlkstroem, T. Drakenberg, S. Forsen, U. Sjoebring, and L. Bjoerck, "Three-dimensional solution structure of an immunoglobulin light chain-binding domain of protein l. comparison with the igg-binding domains of protein g," *to be published*, 1994.

[54] F. Lederer, A. Glatigny, P. Bethge, H. Bellamy, and F. Mathews, "Improvement of the 2.5 angstrom resolution of cytochrome b562 by redetermining the primary structure and using molecular graphics," *Journal of Molecular Biology*, 1981.

[55] A. Wand, Y.Feng, and S.G.Sligar, "Solution structure of apocytochrome b562," *Nature Sructural Biology*, 1994.

[56] J. Thompson, J. Bratt, and L. Banaszak, "Crystal structure of cellular retinoic acid binding protein I shows increased access to the binding cavity due to formation of an intermolecular beta-sheet." *Journal of Molecular Biology*, 1995.

[57] L. Wang, Y. Li, J. Markley, and H.Yan, "Nmr solution structure of type II human cellular retinoic acid binding protein: implications for ligand binding," *Biochemistry*, 1998.

[58] E. Liepinsh, C. Genereux, D. Dehareng, B. Joris, and G. Otting, "NMR structure of citrobacter freundii ampd, comparison with bacteriophage t7 lysozyme and homology with pgrp domains," *Journal of Molecular Biology*, 2003.

[59] L. Prasad, Y. Leduc, K. Hayakawa, and L. Delbaere, "The structure of a universally employed enzyme: V8 protease from staphylococcus aureus," *Acta Crystallogr D Biol Crystallogr*, 2004.

[60] A. Schiefner, J. Breed, L. Bosser, S. Kneip, J. Gade, G. Holtmann, K. D. snd W. Welte, and E.Bremer, "Cation-pi interactions as determinants for binding of the compatible solutes glycine betaine and proline betaine by the periplasmic ligand-binding protein prox from escherichia coli," *Journal of Biol Chem*, 2004.

[61] Joint Center For Structural Genomics, "Crystal structure of putative glycerate kinase (tm1585) from thermotoga maritima at 2.95 a resolution," *to be published*, 2002.

[62] N. Friedman, "The Bayesian structural EM algorithm," in *Proceedings of the International Conference on Uncertainty in Artificial Intelligence (UAI)*, Madison, USA, 1998, pp. 129–138.

[63] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is "nearest neighbor" meaningful," *Proceeding of the 7th International Conference on Database Theory*, 1999.