Predicting Protein Structure with Guided Conformation Space Search

Technical Report 05-63

Oliver Brock TJ Brunette
Bioinformatics Research Laboratory
Department of Computer Science
University of Massachusetts Amherst

November 1, 2005

Abstract

Protein structure prediction is one of the great challenges in structural biology. The ability to accurately predict the three-dimensional structure of proteins would bring about significant scientific advances and would facilitate finding cures and treatments for many diseases. We propose a novel computational framework for protein structure prediction. The novelty of the framework lies in its approach to conformation space search. Conformation space search is considered to be the primary bottleneck towards consistent, high-resolution prediction. The proposed approach to conformation space search represents a major conceptual shift in protein structure prediction, made possible by combining insights and algorithms from robotics and machine learning with techniques from molecular biology in an innovative manner. The key innovation comes from the insight that target-specific information can effectively guide conformation space search towards biologically relevant regions. We propose a framework for protein structure prediction that achieves biological accuracy and computational efficiency by guiding conformation space search using target-specific information. The proposed framework exploits information about the characteristics of the target's energy landscape acquired continuously during search. As search progresses, the continuous integration of these sources of information will tailor conformation space search to the particular characteristics of the target. This tailored conformation space exploration can overcome the current bottleneck, yielding highly accurate and efficient structure prediction.

1 Introduction

Consistent high-resolution protein structure prediction remains one of the most important challenges in molecular biology. A solution to this problem would allow the development of an understanding of cellular processes, thereby facilitating the development of cures and treatments for many diseases. The difficulty of protein structure prediction can be attributed to the vastness of the protein's conformation space. It is therefore not surprising that conformation space search is believed to represent the primary bottleneck towards consistent high-resolution protein structure prediction [8, 50]. We propose a novel approach to protein structure prediction that specifically targets this bottleneck. Compared with existing approaches, it represents a significant conceptual shift in how conformational space search is viewed and thus in how protein structure prediction is achieved.

Computational protein structure prediction can be understood as the search for an energy minimum in the conformation space of the protein. Exhaustive exploration of conformation space is computationally intractable. Existing approaches avoid exhaustive search by using biological information to restrict exploration

to specific regions of conformation space. Within these regions, however, these approaches most commonly perform *random exploration*, based on the Metropolis Monte Carlo method or one of its derivatives. We propose to replace this uninformed, random exploration with an intelligent search procedure that *continuously acquires information* about the search space and then uses this information to *guide the search towards biologically relevant regions*. This search procedure interprets samples in conformation space as observations that reveal information about the specific prediction target. This information is used to guide the search. As search progresses, exploration is continuously adapted based on highly relevant, target-specific information.

In Section 4, we show that target-specific information obtained during search is highly effective in guiding conformation space exploration towards biologically relevant regions. The effectiveness of conformation space search can be further improved by including *additional* target-specific information, such as experimental measurements of the target obtained from nuclear magnetic resonance (NMR) spectroscopy. These measurements represent spatial constraints on the target's structure. Knowledge of these constraints can be used to complement other sources of information in order to guide conformation space search towards regions that are relevant for the prediction of the target's native structure.

In this report, we develop a principled and general computational framework for protein structure prediction, capable of integrating information obtained during search and from experimental measurements. Protein structure prediction proceeds by identifying characteristics of the energy landscape and exploiting the obtained information as well as experimental data to guide conformation space search towards region of biological relevance. This process tailors the search for each structure prediction based on relevant information pertaining to the specific target. To achieve this goal, the proposed approach exploits insights from robotics, machine learning, data mining, and molecular biology. We believe that the combination of techniques from these disciplines will lead to a computational framework for protein structure prediction that surpasses existing approaches in their biological accuracy and generality, as well as in their computational efficiency. To accomplish these objectives, we present in this report an algorithmic framework for protein structure prediction that achieves computational efficiency and prediction accuracy by **guiding conformation space search towards biologically relevant regions** based on information obtained during the search itself. We present preliminary experimental evidence that this novel framework holds the potential to outperform existing *de novo* protein structure prediction approaches.

2 Background and Significance

To obtain an accurate understanding of protein folding is one of the most important challenges in molecular biology [9]. A solution to this problem would enable the efficient functional annotation of the genomes determined by the Human Genome Project and other genome initiatives. It furthermore would facilitate rational drug design and thereby accelerate the finding of cures or treatments for many diseases. Despite its importance, however, the mechanisms of protein folding are still relatively poorly understood and no efficient, general, and accurate procedure to produce the structure of a protein has been established. As a consequence, the number of sequenced genes exceeds the number of known protein structures by almost two orders of magnitude. This gap is widening quickly and it seems impossible to determine the structure of all existing proteins experimentally [23].

Acknowledging the importance and difficulty of protein structure determination, the Protein Structure Initiative (PSI) [54] has been created, complementing to the Human Genome Project [77], to advance the state of the art in structural genomics [23, 74, 84]. The ambitious goal of this initiative is to obtain structural models of the proteins represented by all sequenced genes. The technical approach taken in this initiative is based on the observation that proteins with sequence similarity also exhibit structural similarity. By experimentally determining the structure of a set of carefully chosen proteins using high-throughput experimental techniques, the goal is to obtain complete coverage of protein sequence space [32, 77]. This would bring

every existing protein within "comparative modeling distance" of one of the sequences with experimentally determined structure. The task of general protein structure prediction can then be addressed by comparative modeling techniques [33, 55, 75].

Structural genomics and the PSI hold great promise to advance our understanding of protein structure. However, several fundamental difficulties pertaining to protein structure prediction cannot be overcome by an exclusive application of experimental structure determination and comparative modeling methods.

- Recent studies estimate the number of protein families to be in the high tens of thousands [32, 61]. This would necessitate the experimental structure determination of at least that many proteins within the Protein Structure Initiative. Since its initiation five years ago, the structural genomics centers associated with the PSI have determined the structures of about one thousand proteins [56]. While certainly an important milestone, these results make previous predictions about the availability of tens of thousands of structures within the next few years [74] appear overly optimistic [23].
- Most of the one thousand newly predicted structures come from target sequences that can easily be predicted [6] and do not represent genuinely novel homologous families [41]. This suggests that the high-throughput technology used by structural genomics groups is optimized to determine the structures of proteins that exhibit significant sequence similarity to known structures [23]. It is therefore difficult to assess how much progress towards complete coverage of protein sequence space has been made over the last five years [32, 77].
- An increasing number of genome sequences belong to single-member protein families, i.e., they exhibit no sequence similarity with known sequences. These sequences are called orphan open reading frames (ORFans). They make up about 25-30% of each newly sequenced genome and in some cases up to 60% [65]. If it is assumed that some of these ORFans have novel unique folds, the challenge associated with achieving coverage of protein space would increase dramatically [23].
- There are several categories of proteins that cannot be addressed within the framework of structural genomics [41]. This includes the important category of membrane proteins, as their structure can generally not be determined using experimental techniques [80, 81], and disordered or unstructured proteins [20, 48, 78, 79]. For the latter category, no well-defined native structure exists. Instead, the function of these proteins is facilitated by the partial lack of precise three-dimensional structure.

Due to these challenges, structural genomics researchers predict that the eventual success of structural genomics will require a synergetic integration of experimental techniques and computational approaches, including *de novo* protein structure prediction methods [23].

Computational structure prediction is a well-studied problem and a large number of approaches have been presented in the extensive literature [18, 26, 31, 38]. All of these approaches face a common, critical difficulty, namely that of the tremendous size of conformation space [45]. Because structure prediction requires an adequate sampling of this space, most approaches are only practical for relatively small proteins, yield inaccurate results for larger proteins, and require substantial computational resources. We propose to overcome these limitations by developing methods for protein structure prediction based on the following hypothesis:

> *The exploration of conformation space yields information about the characteristics of the protein's energy landscape. By using this information to direct further exploration of conformation space towards biologically relevant regions, only a small fraction of the overall conformation space has to be explored. This renders protein structure prediction more accurate, more generally applicable, and more computationally efficient.*

To demonstrate the novelty of this approach, we examine the sources of information used in computational approaches to protein structure prediction. In each case, information is exploited to avoid the computationally intractable exhaustive search of conformation space.

- *Comparative or homology modeling* relies on information about experimentally determined structures [38, 55, 75, 62, 73] to almost entirely avoid search in conformation space. These approaches achieve high prediction accuracy, because they rely on highly accurate information. On the other hand, the information is very specific and does not permit generalization to structures without sequence similarities.
- *Threading or fold recognition*, similarly to comparative modeling, only considers structural elements that have been experimentally observed. In contrast to comparative modeling, however, a prediction is determined by assembling small structural components. The assembly is guided by an energy function [26, 33, 36, 42, 87, 83]. By considering small structural components, threading is able to predict previously unknown structures. This advantage comes at the cost of having to perform search during the assembly process.
- *Ab initio or de novo methods* rely on information about atomic physiochemical interactions [1, 31, 68, 86] captured in energy functions (force fields). Exhaustive search of conformation space is replaced by Monte Carlo methods [27, 49] or simulated annealing [37]. These methods exploit the gradient of the energy potential as a source of information to guide the search towards the native state.

  Search in conformation space can be simplified further by considering an additional source of information. The fragment assembly approach [5, 59, 66] yields significant improvements in prediction accuracy and computational efficiency by restricting conformation space search to parameters given by short protein fragments retrieved from the protein data bank (PDB) [57].
- *Molecular dynamics* [30, 35, 69] relies on information about atomic physiochemical interactions, in conjunction with a simulation of the equations of motion of the entire physical system [34, 46, 47, 76]. The folding trajectories obtained by molecular dynamics simulations are generally in agreement with experimental evidence. The simulation is so computationally expensive, however, that for most proteins only a small fraction (on the order of hundreds of nanoseconds [71]) of the folding trajectory can be computed.

Threading and *de novo* structure prediction use biological information to reduce the search space, but still depend on efficient conformation space search to make accurate predictions. The most commonly used search method is the Monte Carlo method [49]. Many improvements to the Monte Carlo method have been proposed; they include the replica Monte Carlo method [70], the multiplexed-replica exchange method [58], the multi-canonical ensemble method [4], parallel tempering [29], jump walking [22], multi-canonical jump walking [82], smart walking [89], entropic sampling [43], methods based on weighted histograms [39], local energy flattening [85], importance sampling [72], and sampling-importance resampling [67].

All of the aforementioned search methods make only limited use of the information obtained during the exploration of the solution space. This becomes obvious when one considers the small amount of state information they maintain (conformation, temperature, etc.). Noteworthy exceptions are conformational space annealing [44] and tabu search [25]. These methods maintain more significant state information during the search. Conformational space annealing maintains information about multiple concurrent Monte Carlo runs to ensure coverage of the solution space. Tabu search labels regions of the search space as tabu if they do not contain the desired solution. Neither of these methods, however, uses information to explicitly guide search towards the correct solution.

Given this discussion of the state of the art in protein structure prediction and associated conformation space search methods, where are opportunities for improvement?

A recent paper by Bradley, Misura, and Baker states that "the primary bottleneck to consistent high-resolution prediction appears to be conformational sampling" [8] (see additional statements by Prof. Baker in enclosed letter of support). This statement is also supported in a recent paper by John Moult [50], evaluating a decade of CASP (Critical Assessment of Techniques for Protein Structure Prediction) competitions. One of the bottlenecks for template-free modeling discussed in this paper pertains to the difficulty of selecting the most accurate structure from a large set of candidates [50]. The structural variability among those many candidates has been attributed to energy potentials that are not funneled all the way to the native-like structures, but instead are shaped like a caldera around the native state [31]. Our results show that this caldera might be much narrower than assumed, or may not even exist (see Figure 12). Instead, the perceived caldera is due to inadequate conformation space search.

Another bottleneck for this category of structure prediction approaches are scoring functions that consider atomic interactions at adequate levels of detail [50]. Increasing this level of detail to complete all-atom models, however, will result in more complex and jagged energy landscapes, further increasing the need for improved conformation space search techniques.

The effectiveness of conformation space sampling can be increased through the use of biological information. This has been demonstrated by the introduction of the fragment assembly approach [51, 52, 53, 59], which uses the biological information contained in structural fragments retrieved from the PDB. Since its introduction during CASP3 in 1998, the fragment assembly approach has dominated the field of *de novo* protein structure prediction [50]. Recent results seem to show that this approach is capable of predicting any biologically plausible, novel fold, despite the fact that the structural fragments are limited to those contained in the PDB [88].

If the effectiveness of the search can be improved through the use of information, then the degree of improvement will depend on the relevance of the information to the particular prediction problem. Given this obvious statement, it is surprising that *de novo* structure prediction approaches have made very limited use of the information obtained *during* the prediction process itself. Instead, they rely almost exclusively on information that was available *a priori*. But the search in conformation space continuously reveals new information about the protein in question. This information could be used to direct the search towards regions of the search space that are biologically relevant *for the specific protein*.

We believe that critical improvements in protein structure prediction will result from exploiting the information obtained during search to guide future explorations of conformation space. This claim is supported by preliminary experimental evidence presented in Section 4.

Apart from designing better search algorithms, as proposed here, one might expect improvements in conformation space search to come from increasing processing speeds of computers, or from exploiting massive computational parallelism, as in the Folding@Home project [40, 64]. While both of these factors yield incremental improvements, they will not be able to overcome the fundamental bottlenecks of conformation space search. This can be seen by examining the size of conformation space. Even if one considers biologically relevant proteins to have a maximum length and therefore the maximum relevant conformation space to be of finite size, this space is too large to be searched exhaustively with any conceivable computational means. This is illustrated very well by the Levinthal paradox [45]. A successful attempt to overcome existing bottlenecks therefore has to focus on novel algorithmic approaches instead of more powerful computers.

In conclusion, we believe that consistent high-resolution protein structure prediction can only be achieved by improving conformation space search. This will require the exploitation of protein-specific information during the search. This information can be obtained during the search itself.

3 Protein Structure Prediction Framework Based on Guided Search

3.1 Algorithmic Framework

In this section we discuss the algorithmic principles that underly the proposed framework. In Section 3.2 we then show how these principles can be translated into a specific implementation of the protein structure prediction framework.

Relation to existing approaches: We develop a *general* and *efficient* algorithmic framework for protein structure prediction. We define generality as the ability to make template-free predictions [50], i.e., predictions of folds that have not been previously observed, for proteins of any biologically plausible length. Efficiency refers to the computation time required to make predictions with maximum accuracy for a given prediction method. We begin by classifying existing approaches according to these characteristics.

Figure 1 graphs the generality and efficiency of the most common structure prediction methods. Homology modeling is computationally efficient, because it predicts the structure of a target by finding homologous proteins with known structure. Since a prediction requires the existence of homologous folds, however, this approach is not general. Molecular dynamics covers the opposite end of the spectrum. It folds proteins from first principles and therefore is very general. But due to the computational cost of simulating folding trajectories from first principles, its efficiency is poor.

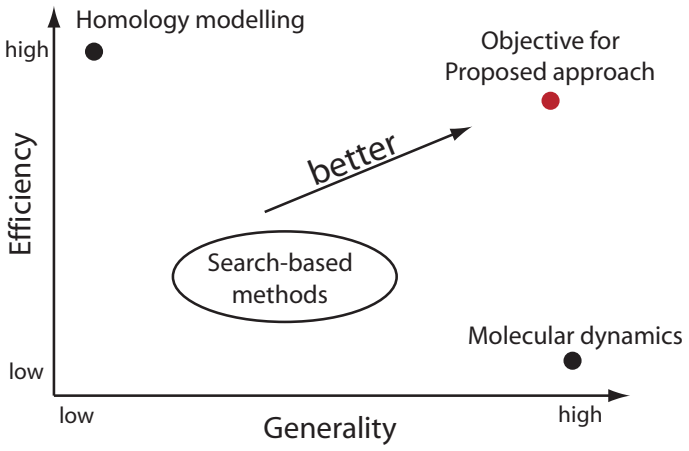


Figure 1: Generality and efficiency of prediction methods.

A third group of structure prediction approaches, including threading and the fragment assembly method, strikes a balance between molecular dynamics and homology modeling: they use less specific information than homology modeling to achieve generality and use less general knowledge than molecular dynamics to achieve efficiency. Instead, these search-based methods generally rely on information about short segments of the protein. This new type of information does not allow to predict the structure directly, making it necessary to search the solution space for the most accurate prediction. The size of this solution space increases dramatically with the size of the protein, requiring increasingly efficient search methods for structure prediction of large proteins. This implies that the efficiency of the search method does not only impact the computational resources necessary for structure prediction, but also determines the generality of the prediction method.

We propose the development of a protein structure prediction framework based on a novel and highly efficient search procedure. The arguments above imply that such a framework would achieve better efficiency and more generality than existing search-based methods.

To understand how a more efficient search procedure can be devised, we examine the relationship between available information and the size of the resulting search space (see Figure 2). Homology modeling uses information that effectively indicates the structure of the target and therefore requires very minimal search. Molecular dynamics uses very detailed information about the dynamic evolution of the folding process to limit the search space to local minima along the folding pathway. Existing search-based methods use very general information to reduce the size of the search space, but only achieve good prediction results for

short proteins [8, 50].

Our approach uses additional information to further reduce the search space so that template-free structure prediction becomes biologically accurate and computationally efficient, even for large proteins. The approach accomplishes this by exploiting the information that it obtains during the search itself. Every step of the search uncovers information about the prediction target. This information can be used to further guide the search. The information is highly relevant, because it is target specific, and it does not incur a significant computational cost, since it exploits information that is obtained anyway. Existing approaches predominantly rely on random exploration and discard the information obtained during search.

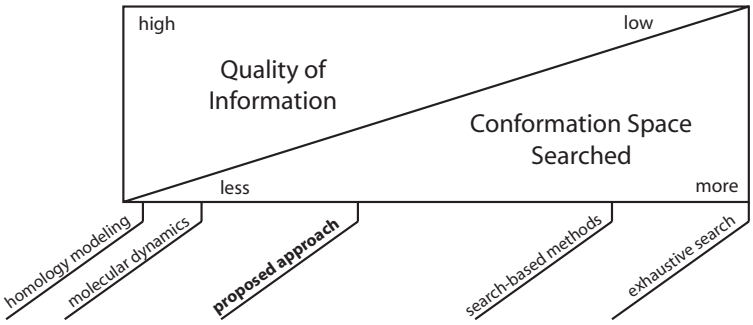


Figure 2: Quality of used information and size of search space for prediction methods.

Overview: Protein structure prediction can be viewed as the search for the global minimum of the protein's energy function. Due to the complexity of the energy landscape, search-based methods determine minima in this landscape by sampling. Each sample placed during search contains a small amount of information about characteristics of the underlying energy function. Monte Carlo-based methods do not exploit this information—they only remember information about the current location of the search. If it were possible to efficiently extract and maintain the information obtained throughout the entire search, this information could be used to direct the ongoing search towards important regions of the search space. As search progresses, the information would become more and more accurate and search would be directed in increasingly effective and accurate ways, yielding a highly efficient search procedure. This is the principle underlying the proposed framework for protein structure prediction.

Figure 3 depicts the general idea behind the proposed structure prediction framework. Samples taken during search can be seen as observations in the energy landscape. By aggregating observations over conformation space and time (as search progresses), the proposed framework can extract information about characteristics of the energy landscape. These characteristics indicate regions of the energy landscape that are more likely to contain good local minima. The proposed framework capture these characteristics as a probabilistic bias defined over the entire search space. By performing search guided by this bias, the effectiveness and accuracy of structure prediction are optimized.

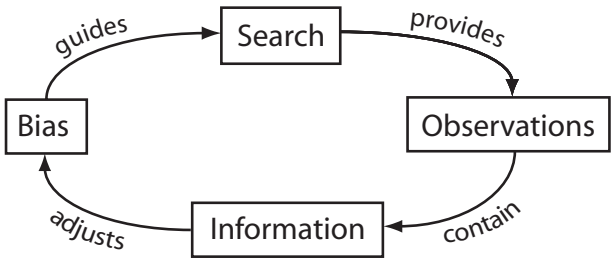


Figure 3: Guiding search by information obtained during search.

The general concept of exploiting feedback from past actions to guide future actions has been used extensively—and very successfully—in adaptive feedback control in robotics [3, 17, 21] and in incremental learning [19, 24] in the context of machine learning. An incremental learner updates its belief about the learning task after each training example it sees. A particular type of incremental learning is active learning [15, 16, 19]. An active learner does not only learn incrementally, but is also able to chose its training

data based on information obtained in the past. This situation corresponds exactly to the proposed approach depicted in Figure 3: the search algorithm uses its past experience to ask for a sample (training data), which in turn is used to update its hypotheses about the energy landscape. Therefore, active learning provides a firm theoretical foundation for the proposed approach to structure prediction. This theoretical foundation indicates that—under certain circumstances—active learning acquires models of functions exponentially faster than random sampling [60]. In other research, we have applied these principles with great success to robot motion planning in high-dimensional configuration spaces [11].

In the remainder of this section, we provide details about the components of the diagram in Figure 3.

Guiding search by biasing sampling: In the context of this proposal we consider search to be performed by sampling from conformation space. Guiding such a search amounts to varying the sampling density in different regions of conformation space.

Before we describe how the proposed framework biases sampling, we consider biasing in the context of a Metropolis Monte Carlo search. A particular Monte Carlo run starts at a random location in conformation space. We refer to locations in conformation space as decoys. The initial decoy is modified incrementally and randomly to generate subsequent samples. Effectively, search explores the local region around the initial decoy. It proceeds by finding a starting point *randomly* and subsequently performing a *random* structural modification to a *randomly* chosen locale on the decoy. The resulting search is only guided by the Metropolis criterion [49], producing a bias towards low-energy regions of the conformation space relative to the randomly chosen starting decoy. Since the energy landscape contains a large number of spurious minima, search expends substantial computational resources in biologically irrelevant regions.

The proposed structure prediction framework also relies on Monte Carlo-like search, i.e., search that progresses by taking small steps in the search space starting at previously sampled locations. However, in contrast to Monte Carlo, which performs these steps randomly, the proposed framework uses a bias to direct the choice of the initial decoy (conformation space *region*), modification *locale* on the decoy, and the type of modification performed to the decoy (*structure*). This bias is computed based on the information contained in previously taken samples. Mathematically speaking, the framework computes a joint conditional probability distribution $P_t(R, L, S|O_t)$ over conformation space regions R, candidate structures S, and locales L, given the observations $O_t = \{o_1, \cdots, o_t\}$. Candidate structures represent all possible search steps by providing a set of structural fragments that can replace a particular region of the decoy. Locale refers to different positions on the decoy; for each locale a different set of candidate structures is used. These candidate structures are obtained from homologous segments in the PDB [57]. The distribution $P_t(\cdot)$ is updated to $P_{t+1}(\cdot)$ when a new observation o_{t+1} is made.

Figure 4 illustrates these biases graphically. We give a high-level overview here, and describe the biases in more detail in Section 3.2. The bias for conformation space regions R is represented by a non-parametric model [2, 63], consisting of a set of decoys indicated as horizontal lines in the figure. A non-parametric model does not attempt to estimate a parametric distribution for the observed data, but instead stores the observed data itself. In our non-parametric model, the distribution of decoys in a conformation space region r indicates the probability of picking a starting point for search from r. The non-parametric model can be maintained by only keeping the most "promising" decoys. As a consequence, the distribution of decoys will change, causing the region bias to change over time.

For each decoy in the non-parametric model of conformation space, we maintain a bias for *where* to perform a modification on the decoy (locale) and for *how* to modify the decoy (structure). A structural bias represents a distribution over a set of considered structural changes, capturing the belief that certain structures are more likely than others to result in favorable modifications. Similarly, the locale bias indicates positions on the decoy in which sampling is more likely to result in a favorable replacement. For example, one conceivable bias could reduce the frequency of changes in regions with well-formed secondary structure.

Similarly to the region bias, the locale and structure bias are updated over time to reflect new information acquired during search. In contrast to the region bias, however, we only consider a finite set of locales and structures, allowing us to represent the bias explicitly as a discrete probability distribution.

The bias $P_t(R, L, S|O_t)$ over conformation space regions R, locales L on the decoy, and structural replacements S for locales on the decoy captures the information about promising conformation space regions and search directions obtained from previously obtained samples. The bias can thus be used to guide search towards regions of conformation space that have been identified as biologically relevant for the specific prediction target.

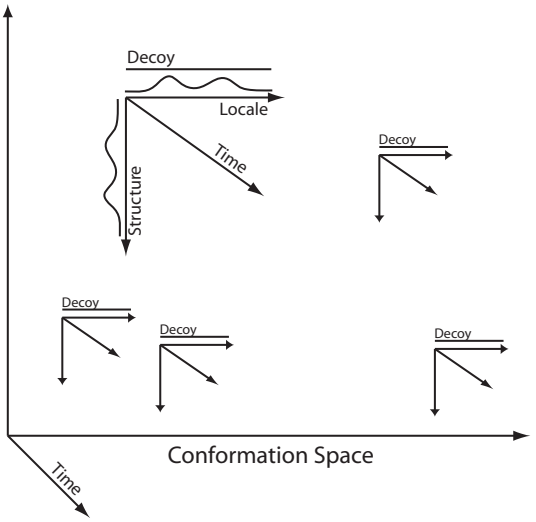


Figure 4: Biasing search by conformation space region, decoy structure, and decoy locale, as a function of time.

Observations: We would like to incrementally adjust the bias $P_t(R, L, S|O_t)$ based on available observations $O_t = \{o_1, \cdots, o_t\}$. An observation is obtained by performing a small step in search space, starting from a previously sampled location. Each location corresponds to a particular structure of the protein and is referred to as a *decoy*. An observation provides us with specific information about the decoy, including its structure, its energy, and whether or not the new decoy has lower energy than the original one. To guide search effectively, we have to interpret observations about decoys to allow appropriate adjustments to the bias.

Obtaining information from observations: Each individual observation only provides limited information about the energy landscape. As a consequence, search methods that rely exclusively on information obtained from individual observations cannot effectively guide the exploration of conformation space. This is particularly evident for Monte Carlo-based approaches, which search conformation space by attempting random search steps. But when multiple observations are viewed together, they may reveal important characteristics of the energy landscape that indicate how search should proceed.

The extraction of information from multiple observations determines accuracy and generality of the proposed protein structure prediction framework. More relevant, accurate, specific, and useful information can guide the search more effectively, leading to more accurate predictions for a wider range of proteins.

Information can be obtained by either aggregating multiple observations or by analyzing a specific decoy. Multiple observations can be aggregated in different ways, with each type of aggregation providing different kinds of information. If multiple observations are aggregated for all decoys currently under consideration (see Figure 4), we obtain information about the current state of the search. This information can be used to adjust search between a broad sampling of the conformation space in early stages and a fine-grained search of specific local minima. In the former case, one would expect large steps in search space to be successful; in later stages of the search, smaller steps are more likely to succeed. Similarly, it is possible to aggregate information for a particular decoy, obtaining information about a particular region of the conformation space, or for locales on the decoy, providing information about specific dimensions of the conformation space.

In Section 3.2 we will describe how a subset of these sources of information were exploited successfully for protein structure prediction, obtaining the preliminary results presented in Section 4.

Adjusting the bias: The information obtained from observations is used to adjust the bias. This bias influences which search steps are performed next. If the information we extracted from the observations is

accurate, the bias will focus search on biologically relevant regions. The adjustment of the bias closes the loop in Figure 3, and the process iterates, incrementally improving and updating the bias.

3.2 An Implementation of the Proposed Framework

We now describe a specific implementation of the algorithmic framework described in the previous section. In this implementation, we have made specific choices about how to represent the bias, how to extract information from observations, and how to adjust the bias based on this observation. The present, preliminary implementation serves as a proof-of-concept to illustrate that the proposed framework for protein structure prediction is able to improve the accuracy and generality of existing approaches.

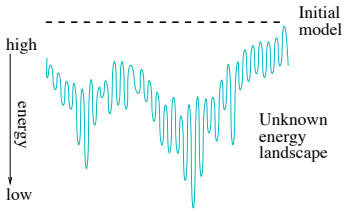
Integration with Rosetta: Rosetta [5, 7, 8, 59] is the most successful *de novo* protein prediction package [50]. Rosetta predicts protein structures using the fragment assembly approach. In this approach, structural fragments (9-mers and 3-mers) matching the amino acid sequence of the prediction target are retrieved from the PDB [57]. A list of such candidate structures is retrieved for each position (locale) of the target sequence. Rosetta predicts a structure for the target by searching the space of all possible fragment assemblies that match the amino acid sequence of the target. This search is performed using a large number of Monte Carlo runs, each starting from an extended initial decoy. The lowest-energy decoys found by the Monte Carlo runs are clustered; the center of the largest, low-energy cluster is most commonly considered to be the predicted structure.

Our proposed structure prediction framework specifically addresses the problem of conformation space search. To validate this framework, we have integrated a preliminary implementation with the C++ version of the Rosetta software package. Our implementation was forked from the Rosetta source tree in January of 2005. By virtue of source code integration, this implementation uses Rosetta's fragment assembly approach. Therefore, the search space of Rosetta as well as our proposed method consists of all possible fragment assemblies, rather than the full space of possible bond angles. In our implementation, we rely on Robetta [14] to retrieve appropriate fragments from the PDB. Furthermore, we rely on Rosetta's energy function to determine the energy of decoys. We also utilize Rosetta's incremental search approach: search progresses in *stages*, each of which uses an increasingly accurate energy function. However, the remaining aspects of structure prediction, namely those relating to conformation space search, have been replaced by an implementation of the algorithmic framework described in Section 3.1. Given that all conformation space search by Rosetta and by the proposed methods is performed in an identical search space, searching an identical energy function, the experimental evaluation in Section 4 truly measures the ability to search conformation space.

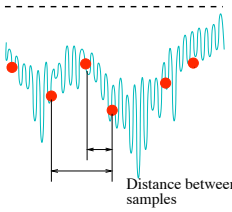
Region bias: We begin by describing a specific approach to conformation space search based on the region bias, called model-based search (MBS) [10], without considering structure bias and locale bias (see Figure 4). The objective of the region bias is to direct search towards low-energy regions of the search space. To compute a region bias, we represent a very rough approximation of the energy landscape by a set of decoys; this method of representing an approximation to a function is referred to as a non-parametric model [2, 63]. By examining the relationship of nearby decoys in our model, we identify low-energy regions and subsequently focus search on these regions.

The following description of our implementation of the region bias by model-based search (MBS) refers to Figure 5, illustrating the search for a global minimum of an unknown, jagged energy landscape.

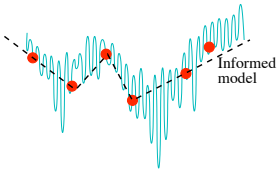
1. Initially, the model (shown as a dashed line in the figure) contains no decoys and hence no information about the energy landscape. Our region bias is uniform, favoring no specific region.



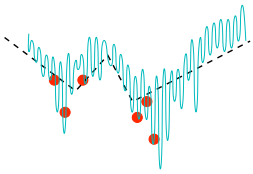
1. Initial model contains no decoys and therefore represents uniform bias.



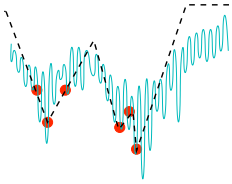
2. First set of decoys is generated by random sampling; resulting decoys are improved by short Monte Carlo runs.



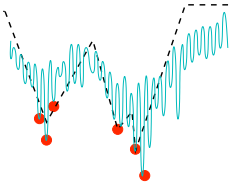
3. Decoys are used to estimate shape of energy landscape.



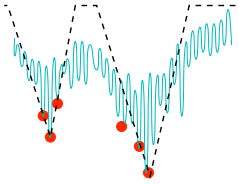
4. New decoys are generated by performing short Monte Carlo runs from local minima of the approximate model.



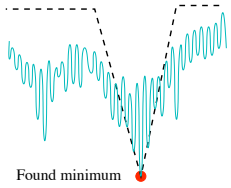
5. Decoys are used to improve estimate of the shape of energy landscape.



6. New decoys are generated by performing short Monte Carlo runs from local minima of the approximate model.



7. Decoys are used to improve estimate of the shape of energy landscape; irrelevant minima are removed from model.



8. Lowest-energy decoy found.

Figure 5: Region bias implemented with model-based search (MBS): iterative refinement of an approximate, non-parametric model of the energy landscape to find its global minimum.

2. Based on a uniform region bias, the energy landscape is sampled uniformly at random to obtain a set of initial decoys. Short Monte Carlo runs are used to find decoys in nearby local minima. This improves the quality of the information contained in the decoys.
3. The resulting decoys contain information about the energy landscape that we want to use to direct future search towards promising regions of conformation space. To determine where these promising regions lie, we compute distances between all decoys. The chosen distance metric is of critical importance for the accuracy of the resulting information. An adequate metric has to capture the true distances in the physical energy landscape; in our preliminary implementation we use RMSD between decoys. Based on the relative distances between decoys and their energy, we can find decoys that are of lower energy than their nearest neighbors. Effectively, we are determining a rough, non-parametric model of local minima in the energy landscape: the decoy of locally lowest energy represents an estimate of the bottom of the well and its neighbors provide an estimate of the width of the well. In the figure, this estimate is indicated by a dashed line. We can interpret this approximate model of the energy landscape as a region bias. Subsequent search is directed towards local minima regions in the model.
4. Search now proceeds according to the region bias. Starting from the local minima represented in the model, we generate new decoys by performing short Monte Carlo runs. The number of decoys generated per local minimum is determined by its energy level and its width. Lower-energy minima generate more decoys, as they represent more promising regions. Wide minima also generate a larger number of decoys, because larger regions require more detailed exploration. Effectively, the region bias indicates in which regions search should proceed as well as how much additional exploration per region should be performed.
5. The approximate, non-parametric model is updated with the newly generated decoys. This is accomplished by simply adding the decoys to the set of decoys present in the model. Based on this new set of decoys, the location, depth, and width of local minima are estimated. Decoys that do not play a role in this estimation can be discarded, keeping the size of the model small.
6. Based on the improved region bias, new decoys are generated.
7. The approximate, non-parametric model is updated with the newly generated decoys. We only maintain the most promising local minima, discarding other parts of the model. By discarding some of the local minima, the region bias is further refined and the size of the model is reduced.
8. The "global" minimum has been identified. While it is impossible to verify that the found minimum is the global minimum, experimental evidence presented in Section 4 shows that conformation space search based on the region bias alone finds significantly lower minima than other methods.

Effectively, model-based search controls the starting points of a large number of dependent, short Monte Carlo runs. Monte Carlo runs in high-energy regions of the landscape are aborted in favor of additional runs in low-energy regions. The effectiveness of search based on the region bias significantly exceeds that of search based on pure Monte Carlo (see Section 4). This confirms the hypothesis underlying the proposed research, namely that search can be guided intelligently based on information acquired during search.

Note that search based on the region bias, as implemented by model-based search, still generates search steps randomly using the Monte Carlo method. By adding a structure and a local bias, search steps will be guided by additional information, further increasing the efficiency of conformation space search.

Structure bias: The region bias described above determines *where* in conformation space the search is performed. We now describe a bias that influences *how* the search proceeds. More specifically, the structure bias imposes a bias on the choice of search steps to perform conformation space exploration.

The space of all possible search steps is given by all vectors in d dimensions, where d is the dimensionality of the search space. Each vector specifies a direction for the search and a length, indicating the size of the search step. The fragment assembly approach discretizes the space of all vectors by only considering vectors that correspond to a fragment replacement in the decoy (see Figure 6). The search vectors allowed by the fragment assembly approach are distributed non-uniformly if several similar fragments result in similar search step vectors. This non-uniformity can bias search towards conformation space regions that are not relevant to the prediction target.

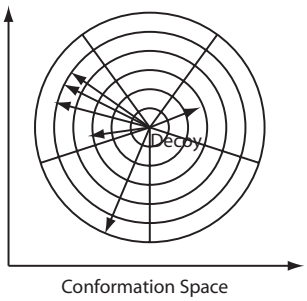


Figure 6: Starting from a specific decoy, fragment replacements represent search steps in conformation space.

We have implemented a structure bias that overcomes the unwanted bias present in the fragments. The goal of this bias is to explore all biologically relevant search vectors uniformly. A search vector is biologically relevant if it corresponds to a fragment replacement (this is the assumption underlying the fragment assembly approach). However, the number of similar fragments should not determine the sampling weight given to the search vector. Instead, we determine the set of distinct vectors represented by the fragments and sample uniformly from them, rather then from the fragments themselves.

To extract biologically relevant directions for search from the fragments, we perform greedy agglomerative clustering [28]. These clusters represent biologically relevant and distinct directions. The clusters are illustrated by the regions formed by the concentric circles in Figure 6. Each region corresponds to a cluster, aggregating one or multiple search vectors obtained from the fragments. By sampling uniformly from the clusters we establish a structure bias that guides search uniformly according to information contained in the fragment clusters. Combining this structure bias with the region bias of model-based search, the effectiveness of conformation space search is further improved, as evidenced by the results presented in Figures 9 and 10 in Section 4. This illustrates that additional sources of information afford further improvements in conformation space search, confirming the premise of the proposed protein structure prediction framework.

The success of the structure bias is highly dependent upon the distance metric used for clustering. Our distance metric combines information about the secondary structure of the fragments' residues as well as their spatial structure. The secondary structure distance between two fragments is given by the number of positions in which the secondary structure labels in the PDB differ. Structural distance between two fragments is determined by the RMS distance matrix error (dme) [12], which evaluates the distances of corresponding C_α atoms. The distance metric used for clustering weighs secondary structure more heavily than the distance based on spatial structure so that clusters with identical secondary structure distance are further differentiated based on the RMS distance matrix error.

Locale-dependent structure bias: The structure bias described above enables uniform exploration of the search space, providing improvements over the unwanted bias introduced in the fragment assembly approach. The structure bias is identical for every locale on every decoy, enforcing uniform exploration everywhere. This stands in contrast with some of our experimental observations, indicating that the most promising search vectors vary by conformation space region and by locale on the decoy. To exploit this insight, we devise a decoy- and locale-dependent structure bias (see Figure 4).

Our goal is to find the search vectors that are more likely to lead to successful search steps, i.e., to a reduction in energy. The graph in Figure 7 shows the average secondary structure distance for search steps aggregated per stage of the search. (Recall that Rosetta performs structure prediction in stages.) A distance of zero corresponds to a perfect secondary structure match between the fragment in the decoy and its replacement; a distance of one corresponds to a perfect mismatch. The graph shows that successful search steps are more likely to occur when the secondary structure of the replacement fragment closely resembles

the secondary structure of the fragment it replaces. This holds true for all but the first stage of the search, in which it is more important to explore the conformation space broadly. The spike in stage 14 is caused by the fact that search changes from replacing 9-mers in the decoy to replacing 3-mers.

This insight immediately suggests search should be biased towards fragments of similar secondary structure, starting at stage two. Such a bias varies with the fragment currently present at each locale of each decoy. An appropriate bias is illustrated in Figure 8, giving more sampling weight to search vectors in the vicinity of the original decoy. The height of the cone indicates the strength of the bias.

We have implemented an adaptive, locale-based structure bias. For every locale on each decoy we keep track of the average secondary structure distance for all search steps and for all successful search steps. Based on these aggregated observations, we adjust the bias such that the expected distance for future search steps selected with the bias is equal to the distance observed in successful search steps. This can be done by increasing the sampling weight of clusters with high secondary structure similarity to the current decoy. Our current implementation determines the height of the cone in Figure 8 only once. An obvious improvement is to adjust this bias as search progresses.

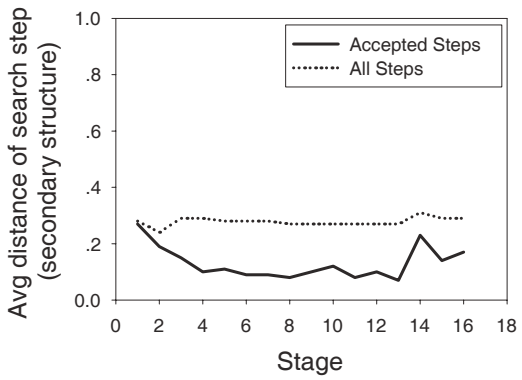


Figure 7: Average length of search steps for successful and unsuccessful fragment insertions per stage for antitermination factor NusB chain A (1EYVA, 139aa).

When the current, simplistic locale-based structure bias is combined with the region bias provided by model-based search, we observe significant performance gains compared to the pure structure bias described above. This is confirmed by the experimental results shown in Figures 9 and 10 in Section 4. This additional increase in search efficiency underlines that the information extracted from observations during search can successfully guide future search.

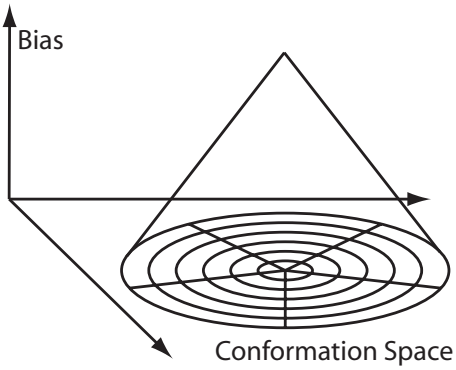


Figure 8: The locale-dependent structure bias causes small search steps to be sampled more frequently.

Summary: Our preliminary implementation of the proposed framework for protein structure prediction illustrates three important points:

- We are able to extract meaningful information from observations obtained during search.
- This information can guide search towards biologically relevant regions, substantially improving the effectiveness of conformation space search.
- The effectiveness of search continues to improve when multiple biases are combined and when increasingly meaningful information is obtained to determine the appropriate settings for these biases.

These findings validate the hypotheses that form the foundation of the proposed prediction framework.

4 Experimental Results

The main contribution of the our approach lies in effective conformation space search for an energetic minimum in a protein energy landscape. Consequently, we regard the **energy level of the lowest-energy decoys** found by a search method to be the **main criterion for evaluation**.

In our preliminary studies we compare the proposed protein structure prediction framework with the Rosetta package from the Baker laboratory [5, 8, 59]. Rosetta is regarded as the most successful *de novo* protein prediction package [50]. The source code for Rosetta is freely available and we have integrated a preliminary implementation of the proposed conformation space search method with the Rosetta package. In all experiments, the parameters are identical. We only vary the conformation space search method. This means that the same energy function is searched and the same fragment assembly approach with identical sets of fragments is used. Each experiment uses approximately the same amount of computational resources. Furthermore, experiments are fully automated.

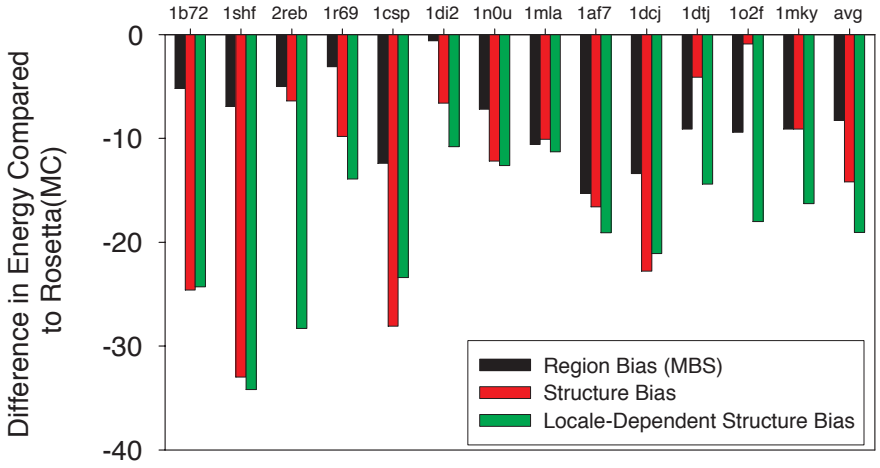


Figure 9: Difference in energy for decoys obtained by the preliminary implementation of the region bias, region/structure bias, and region/locale-dependent structure bias, relative to the energy of decoys found by Rosetta; small proteins test set (49–81aa). The rightmost column indicates average over all proteins.

To perform biologically accurate structure prediction using Rosetta, the package is generally run on high-performance compute clusters with many dozens of compute nodes. Our laboratory currently has only personal computers at its disposal. (We will install a 10 node high-performance compute cluster in November 2005.) Consequently, the objective for these preliminary studies cannot be to compete with the biological accuracy of massively parallel computations, but instead to demonstrate the improved efficiency of conformation space search based on the proposed methods. To perform experiments suited to our computing environment, we significantly reduce the number of decoys from 20,000–30,000 [8] to 350–500. In addition, we do not use the all-atom energy function of Rosetta, since its evaluation requires substantial computation time. These steps are necessary to allow us to perform experiments, given the computational resources available to us. But these steps will also affect the biological accuracy of predictions. Nevertheless, the presented experiments permit an accurate comparison of conformation space search methods. The gains we obtain in these comparisons will only increase for experiments with large number of decoys, due to the intrinsic properties of the compared methods: whereas Monte Carlo-based search methods, such as the one used in Rosetta, simply run additional, *independent* Monte Carlo runs, the proposed methods will ben-

efit from the acquisition of additional information by adding decoys, and consequently guide conformation space search even more effectively.

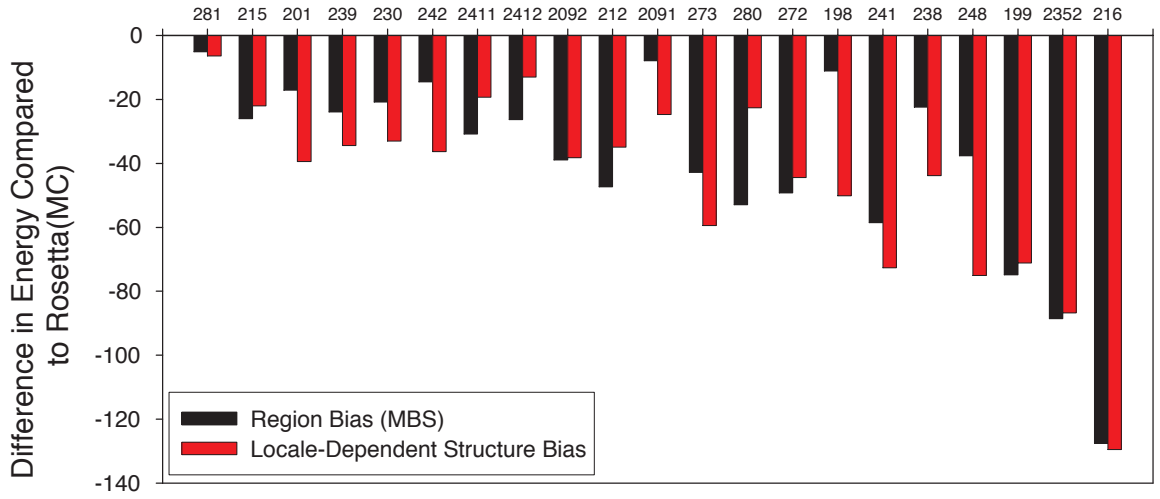


Figure 10: Difference in energy for decoys obtained by with the preliminary implementation of the region bias and the region/locale-dependent structure bias, relative to the energy of decoys found by Rosetta; CAFASP proteins test set (70–435aa); proteins are shown by increasing length.

We present experiments with thirty-four proteins, ranging in length from 49 to 435 amino acids. The first set consists of thirteen small proteins, taken from a recent paper by the Baker group [8]. These proteins range in length between 49 and 81 amino acids. From the 16 proteins discussed in the paper, we chose thirteen for which we were able to infer the cropped primary structure. The second group of targets were chosen from the Critical Assessment of Fully Automated Structure Prediction (CAFASP) 4 competition [13]. We chose ten targets from the New Fold category [50] with lengths between 94 and 435 amino acids (targets 201, 2092, 216, 238, 2411, 2412, 241, 242, 248, 273) and eleven targets from the difficult Fold Recognition category [50] with lengths between 70 and 410 amino acids (targets 198, 199, 209-1, 212, 215, 230, 235-1, 239, 272, 280, 281). These two categories represent protein fragments with very little sequence similarity to known structures and therefore are well-suited for the evaluation of a *de novo* prediction method.

(a) Native (b) Rosetta (c) Region bias (d) Structure bias

Figure 11: Native structure and predictions of a binding protein from the Immunoglobulin L Chain.

The proposed research is based on the hypothesis that it is possible to guide conformation space search with information obtained during the search. We present experimental results for a preliminary implementation of the proposed conformation space search method. This implementation exploits three types

of information: *1)* information about the relevance of certain conformation space regions (region bias), *2)* information about the direction in which search progresses (the type of fragment inserted into the decoy, structure bias), and *3)* information that allows us to vary the search direction for different locales on the decoy (locale-dependent structure bias). Based on these sources of information we present results for three different implementations, each including an additional source. Referring to Figures 9 and 10, *region bias* is an implementation based on the first source of information (previously, we called this method model-based search (MBS) [10]). *Structure bias* includes the region bias and in addition the structure bias; *locale-dependent structure bias* includes all three sources of information.

Figures 9 and 10 show the reduction in energy obtained with these three preliminary implementations, relative to the decoys generated by Rosetta. The energies were averaged over the lowest 5% of the generated decoys for each method. The experiments clearly indicate that the inclusion of information obtained during search is able to guide conformation space search towards relevant regions, resulting in lower-energy decoys. In all of our experiments, our novel techniques outperform the conformation space search implemented in Rosetta. The results show that the inclusion of additional sources of information yields additional improvements in performance. It is particularly noteworthy that the performance improvements become more pronounced as the size of the proteins increase (see Figure 10). For some of the experiments, however, the inclusion of *additional* sources of information did not result in a performance increase. This can be attributed to the fact that—due to the preliminary nature of our implementation—the available information is not fully leveraged. For example, in the current implementation, the locale-dependent structure bias is only adjusted once during the entire search. This means that this type of information is only used a single time during each search. We are currently extending this implementation to adjust the bias in an ongoing fashion as search progresses. Nevertheless, viewed over the entire test set of proteins, it is obvious that the inclusion of additional information results in improved conformation space search.

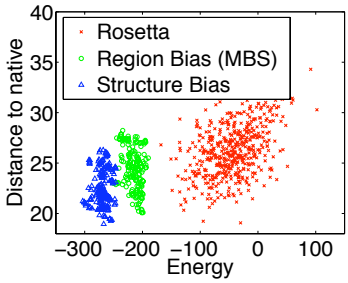


Figure 12: The decoys generated, plotted by their energy and distance to the native structure for Glycerate Kinase (1O0U, 410aa).

For a highly accurate energy function, lower-energy decoys would correspond to more accurate predictions of the native state. However, we did not use the more accurate all-atom energy function of Rosetta, but the simpler energy function that is used during the first sixteen stages of the Rosetta search. Based on this simpler energy function we were able to achieve some accurate predictions for short proteins, such as the one shown in Figure 11. For longer proteins, however, even though we achieved a significant reduction in energy of the decoys, this reduction in energy did not necessarily correspond to a more accurate prediction. This is illustrated by Figure 12, where 500 decoys obtained by Rosetta, the region bias, and the structure bias are shown. In our future work, we will incorporate more accurate energy functions into our framework.

5 Conclusion

This report describes a novel approach to conformation space search in the context of protein structure prediction. Protein structure prediction is viewed as the search for the global minimum in the in the protein's energy landscape. The efficiency and accuracy of structure prediction is currently believed to depend primarily on the effectiveness of conformation space search. Our approach to this problem differs from existing methods in that it obtains information about the energy landscape during conformation space search. This information is subsequently used to guide the search towards biologically relevant regions of the landscape. In contrast, existing conformation space search method—predominantly based on the Monte Carlo

method—perform this search in an uninformed, random fashion, ignoring information obtained during the search. We demonstrate in preliminary experiments that the use of information to guide search significantly improves the effectiveness of conformation space search. This raises hopes that our method is able to overcome the most critical difficulty of existing protein structure prediction approaches, permitting significant improvements in the accuracy and efficiency of protein structure prediction.

References

[1] P. Aloy, A. Stark, and C. H. adn Robert B. Russell. Predictions without templates: New folds, secondary structure, and contacts in CASP5. *Proteins: Stucture, Function, and Bioinformatics*, 53(S6):436–456, 2003.

[2] C. G. Atkeson, A. W. Moore, and S. Schaal. Locally weighted learning. *Artificial Intelligence Review*, 11(1-5):11–73, 1997.

[3] A. G. Barto, S. J. Bradtke, and S. P. Singh. Learning to act using real-time dynamic programming. *Artificial Intelligence*, 72(1-2):81–138, 2005.

[4] B. A. Berg and T. Neuhaus. Multicanoncial ensemble: A new approach to simulate first-order phase transitions. *Physical Review Letters*, 68(1):9–12, 1992.

[5] R. Bonneau, C. E. M. Strauss, C. A. Rohl, D. Chivian, P. Bradley, L. Malmström, T. Robertson, and D. Baker. De novo prediction of three-dimensional structures for major protein families. *Journal of Molecular Biology*, 322(1):65–78, 2002.

[6] P. E. Bourne, C. K. J. Allerston, W. Krebs, W. Li, I. N. Shindyalov, A. Godzik, I. Friedberg, T. Liu, D. I. Wild, and S. I. Hwang. The status of structural genomics defined through the analysis of current targets and structures. In R. B. Altman, A. K. Dunker, L. Hunter, T. A. Jung, and T. E. Klein, editors, *Proceedings of the Pacific Symposium on Biocomputing*, Hawaii, USA, January 2004. World Scientific Press.

[7] 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 Bioinformatics*, 53(S6):457–468, 2003.

[8] P. Bradley, K. M. S. Misura, and D. Baker. Toward high-resolution de novo structure prediction for small proteins. *Science*, 309(5742):1868–1971, 2005.

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

[10] T. Brunette and O. Brock. Improving protein structure prediction with model-based search. *Bioinformatics*, 21(Suppl. 1), June 2005. Proceedings of the International Conference on Intelligent Systems for Molecular Biology (ISMB), Detroit, USA.

[11] B. Burns and O. Brock. Toward optimal configuration space sampling. In *Proceedings of Robotics: Science and Systems*, Cambridge, USA, June 2005.

[12] C. Bystroff and D. Baker. Prediction of local stucture in proteins using a library o fsequence-structure motifs. *Journal of Molecular Biology*, 281(3):565–577, 1998.

[13] CAFASP 4. Critical assessment of fully automated structure prediction. http://www.cs.bgu.ac.il/ dfischer/CAFASP4/, 2004.

[14] D. Chivian, D. E. Kim, L. Malmström, P. Bradley, T. Robertson, P. Murphy, C. E. M. Strauss, R. Bonneau, C. A. Rohl, and D. Baker. Automated prediction of CASP-5 structures using the Robetta server. *Proteins: Stucture, Function, and Bioinformatics*, 53(S6):524–533, 2003.

[15] D. Cohn, L. Atlas, and R. Ladner. Improving generalization with active learning. *Machine Learning*, 15(2):201–221, 1994.

[16] D. A. Cohn, Z. Ghahramani, and M. I. Jordan. Active learning with statistical methods. *Journal of Artificial Intelligence Research*, 4:129–145, 1996.

[17] J. J. Craig. *Introduction to Robotics: Mechanics and Control*. Addison-Wesley, second edition, 1989.

[18] Y. Duan and P. A. Kollman. Computational protein folding: From lattice to all-atom. *IBM Systems Journal*, 40(2):297–309, 2001.

[19] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley-Interscience, second edition, 2001.

[20] A. L. Fink. Natively unfolded proteins. *Current Opinion in Structural Biology*, 15, 2005. In press.

[21] G. F. Franklin, J. D. Powell, and A. Emami-Naeini. *Feedback Control of Dynamic Systems*. Addison-Wesley, third edition, 1994.

[22] 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*, 93(4):2769–2784, 1990.

[23] I. Friedberg, L. Jaroszewski, Y. Ye, and A. Godzik. The interplay of fold recognition and experimental structure determination in structural genomics. *Current Opinion in Structural Biology*, 14(3):207–312, 2004.

[24] C. Giraud-Carrier. A note on the utility of incremental learning. *AI Communications*, 13(4):215–233, 2000.

[25] F. W. Glover and M. Laguna. *Tabu Search*. Kluwer Academic Publishers, 1998.

[26] A. Godzik. Fold recognition methods. *Methods of Biochemical Analysis*, 44:525–546, 2003.

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

[28] D. Hand, K. Mannila, and P. Smyth. *Principles of Data Mining*. MIT Press, 2001.

[29] U. H. E. Hansmann. Parallel tempering algorithm for conformational studies of biological molecules. *CPL*, 281(1-3):140–150, 1997.

[30] T. Hansson, C. Oostenbrink, and W. F. van Gunsteren. Molecular dynamics simulations. *Current Opinion in Structural Biology*, 12(2):190–196, 2002.

[31] C. Hardin, T. V. Pogorelov, and Z. Luthey-Schulten. Ab initio protein structure prediction. *Current Opinion in Structural Biology*, 12(2):176–181, 2002.

[32] A. Heger and L. Holm. Exhaustive enumeration of protein domain families. *Journal of Molecular Biology*, 328(3):749–767, 2003.

[33] D. T. Jones, W. R. Taylor, and J. M. Thornton. A new approach for protein fold recognition. *Nature Structual Biology*, 358:86–89, 1992.

[34] L. Kalé, R. Skeel, M. Bhandarkar, R. Brunner, A. Gursoy, N. Krawetz, J. Phillips, A. Shinozaki, K. Varadarajan, and K. Schulten. NAMD2: Greater scalability for parallel molecular dynamics. *Journal of Computational Physics*, 151:283–312, 1999.

[35] M. Karplus. Molecular dynamics of biological macromolecules: A brief history and perspective. *Biopolymers*, 68:350–358, 2003.

[36] L. N. Kinch, J. O. Wrabl, S. S. Krishna, I. Majumdar, R. I. Sadreyev, Y. Qi, J. Pei, H. Cheng, and N. V. Grishin. CASP5 assessment of fold recognition target preditions. *Proteins: Stucture, Function, and Bioinformatics*, 53(S6):395–409, 2003.

[37] S. Kirkpatrick, C. D. Gelatt Jr., and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

[38] E. Krieger, S. B. Nabuurs, and G. Vriend. Homology modeling. *Methods of Biochemical Analysis*, 44:509–523, 2003.

[39] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman. The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *Journal of Computational Chemistry*, 13(8):1011–1021, 1993.

[40] S. M. Larson, C. D. Snow, M. R. Shirts, and V. S. Pande. Folding@Home and Genome@Home: Using distributed computing to tackle previously intractable problems in computational biology. In R. Grant, editor, *Computational Genomics*. Horizon Press, 2002.

[41] E. Lattman. The state of the protein structure initiative. *Proteins: Stucture, Function, and Bioinformatics*, 51(4):611–615, 2004.

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

[43] J. Lee. New monte carlo algorithm: Entropic sampling. *Physical Review Letters*, 71(2):211–214, 1993.

[44] J. Lee, H. A. Scheraga, and S. Rackovsky. New optimization method for conformational energy calculations of polypeptides: Conformational space annealing. *Journal of Computational Chemistry*, 18(9):1222–1232, 1997.

[45] C. Levinthal. Are there pathways for protein folding? *Journal de Chimie Physique*, 65(1):44–45, 1968.

[46] M. Levitt. Molcular dynamics of native protein—I. Computer simulation of trajectories. *Journal of Molecular Biology*, 168:595–620, 1983.

[47] M. Levitt. Protein folding by restraining energy minimization and molecular dynamics. *Journal of Molecular Biology*, 170:723–764, 1983.

[48] R. Linding, L. J. Jensen, F. Diella, P. Bork, T. J. Gibson, and R. B. Russell. Protein disorder prediction: Implications for structural proteomics. *Structure*, 11:1453–1459, 2003.

[49] 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*, 21:1087–1092, 1954.

[50] J. Moult. A decade of CASP: Progress, bottleneck and prognosis in protein structure prediciton. *Current Opinion in Structural Biology*, 15(3):285–289, 2005.

[51] J. Moult, K. Fidelis, A. Zemla, and T. Hubbard. Critical assessment of methods of protein structure prediction (CASP): round IV. *Proteins: Stucture, Function, and Bioinformatics*, 45(Suppl. 5):2–7, 2001.

[52] J. Moult, K. Fidelis, A. Zemla, and T. Hubbard. Critical assessment of methods of protein structure prediction (CASP)–round V. *Proteins: Stucture, Function, and Bioinformatics*, 53(Suppl. 6):334–339, 2003.

[53] J. Moult, T. Hubbard, K. Fidelis, and J. T. Pedersen. Critical assessment of methods of protein structure prediction (CASP): round III. *Proteins: Stucture, Function, and Bioinformatics*, 37(Suppl. 3):2–6, 1999.

[54] NIGMS/NIH Protein Structure Initiative (PSI). Better tools and better knowledge for Structural Genomics. http://www.nigms.nih.gov/psi/.

[55] NIGMS/NIH Protein Structure Initiative (PSI). Report on the NIGMS workshop on high accuracy comparative modeling. http://www.nigms.nih.gov/psi/reports/comparative_ modeling.html.

[56] NIGMS/NIH Protein Structure Initiative (PSI). The shapes of life: NIGMS project yields more than 1,000 protein structures. http://www.nigms.nih.gov/news/releases/021005.html. News Release, February 10, 2005.

[57] Protein Data Bank. http://www.pdb.org.

[58] Y. M. Rhee and V. S. Pande. Multiplexed-replica exchange molecular dynamics method for protein folding prediction. *Biophysical Journal*, 84:775–786, 2003.

[59] C. A. Rohl, C. E. M. Strauss, K. M. S. Misura, and D. Baker. Protein structure prediction using Rosetta. *Methods in Enzymology*, 383:66–93, 2004.

[60] M. Saar-Tsechansky and F. Provost. Active sampling for class probability esitmation and ranking. *Machine Learning*, 54(2):153–178, 2004.

[61] O. Sasson, A. Vaaknin, H. Fleischer, E. Portugaly, Y. Bilu, N. Linial, and M. Linial. ProtoNet: Hierarchical classification of the protein space. *Nucleic Acids Research*, 31(1):348–352, 2003.

[62] J. M. Sauder, J. W. Arthur, and R. L. Dunbrack Jr. Large-scale comparison of protein sequence alignment algorithms with structure alignments. *Proteins: Stucture, Function, and Bioinformatics*, 40(1):6–22, 2000.

[63] L. K. Saul and S. T. Roweis. Think globally, fit locally: Unsupervised learning of low dimensional manifolds. *Journal of Machine Learning Research*, 4:119–155, 2003.

[64] M. Shirts and V. S. Pande. Screen savers of the world unite. *Science*, 290(5498):1903–1904, 2000.

[65] N. Siew and D. Fischer. Analysis of singleton ORFans in fully sequenced microbial genomes. *Proteins: Stucture, Function, and Bioinformatics*, 53:241–251, 2003.

[66] K. T. Simons, C. Kooperberg, E. Huang, and D. Baker. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *Journal of Molecular Biology*, 268(1):209–225, 1997.

[67] Ø. Skare, E. Bølviken, and L. Holden. Improved sampling-importance resampling and reduced bias imprtance sampling. *Scandinavian Journal of Statistics*, 30(4):719–737, 2003.

[68] J. Skolnick, Y. Zhange, A. K. Arakaki, A. Kolinski, M. Boniecki, A. Szilágyi, and D. Kihara. TOUCHSTONE: A unified appoach to protein structure prediction. *Proteins: Stucture, Function, and Bioinformatics*, 53(S6):469–479, 2003.

[69] C. D. Snow, E. J. Sorin, Y. M. Rhee, and V. S. Pande. How well can simpulation predict protein folding kinetics and thermodynamics? *Annual Reviews Biophysics and Biomolecular Structure*, 34:43–69, 2005.

[70] R. H. Swendsen and J.-S. Wang. Replica Monte Carlo simulation of spin-glasses. *Physical Review Letters*, 57(21):2607–2609, 1986.

[71] Theoretical and Computational Biophysics Group. NIH resource for macromolecular modeling and bioinformatics – NAMD performance. http://www.ks.uiuc.edu/Research/namd/performance.html.

[72] G. L. Thomas, R. B. Sessions, and M. J. Parker. Density guided importance sampling: Application to a reduced model of protein folding. *Bioinformatics*, 21(12):2839–2843, 2005.

[73] A. Tramontano and V. Morea. Assessment of homology-based predictions in CASP5. *Proteins: Stucture, Function, and Bioinformatics*, 53(S6):352–368, 2003.

[74] A. Šali. 100,000 protein structures for the biologist. *Nature Structual Biology*, 5(12):1019–1020, 1998.

[75] A. Šali, L. Potterton, F. Yuan, H. van Vlijmen, and M. Karplus. Evaluation of comparative protein modeling by MODELLER. *Proteins: Stucture, Function, and Genetics*, 23(3):318–326, 1995.

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

[77] D. Vitkup, E. Melamud, J. Moult, and C. Sander. Completeness in structural genomics. *Nature Structual Biology*, 8(6):559–566, 2001.

[78] S. Vucetic, C. J. Brown, A. K. Dunker, and Z. Obradovic. Flavors of protein disorder. *Proteins: Stucture, Function, and Bioinformatics*, 52:573–584, 2003.

[79] P. E. Wright and H. J. Dyson. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *Journal of Molecular Biology*, 293(2):321–331, 1999.

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

[81] K. Wüthrich. Biological crystallography. *Acta Cristallographica Section D*, 51:249, 1995.

[82] H. Xu and B. J. Berne. Multicanonical jump walking: A method for efficiently sampling rough energy landscapes. *Journal of Chemical Physics*, 110(21):10299–10306, 1999.

[83] J. Xu and M. Li. RAPTOR: Optimal protein threading by linear programming. *Journal of Bioinformatics and Computational Biology*, 1(1):95–117, 2003.

[84] C. Zhang and S.-H. Kim. Overview of structural genomics: From structure to function. *Current Opinion in Chemical Biology*, 7:28–32, 2003.

[85] Y. Zhang, D. Kihara, and J. Skolnick. Local energy landscape flattening: Parallel hyperbolic Monte Carlo sampling of protein folding. *Proteins: Stucture, Function, and Bioinformatics*, 48(2):192–201, 2002.

[86] Y. Zhang, A. Kolinski, and J. Skolnick. TOUCHSTONE II: A new approach to ab initio protein structure prediction. *Biophysical Journal*, 85:1145–1164, 2003.

[87] Y. ZHang and J. Skolnick. Automated structure prediction of weakly homologous proteins on a genomic scale. *Proceedings of the National Academy of Sciences*, 101(20):7594–7599, 2004.

[88] Y. Zhang and J. Skolnick. The protein structure prediction problem could be solved using the current PDB library. *Proceedings of the National Academy of Sciences*, 102(4):1029–1034, 2005.

[89] R. Zhoug and B. J. Berne. Smart walking: A new method for Boltzmann sampling of protein conformations. *Journal of Chemical Physics*, 107(21):9185–9196, 1997.