ADAPTIVE BALANCING OF EXPLOITATION WITH EXPLORATION TO IMPROVE PROTEIN STRUCTURE PREDICTION

A Dissertation Presented

by

TJ BRUNETTE

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2011

Department of Computer Science

ADAPTIVE BALANCING OF EXPLOITATION WITH EXPLORATION TO IMPROVE PROTEIN STRUCTURE PREDICTION

A Dissertation Presented

by

TJ BRUNETTE

Approved as to style and content by:

Oliver Brock, Chair

Lila Gierasch, Member

David Kulp, Member

Ileana Streinu, Member

Andrew G. Barto, Department Chair
Department of Computer Science

I want to thank my advisor, family and friends for making University of Massachusetts a fulfilling and enjoyable place to study. Without their support, guidance and friendship this dissertation would not have been possible.

Foremost, I want to thank my advisor Oliver Brock. Oliver has consistently given good guidance and helped mold me into a much stronger researcher. I wish him the best of luck at his new lab in Berlin.

In addition, I want to thank my secondary advisor Lila Gierasch. Lila welcomed me into her lab and throughout the years has given me valuable advice about protein folding and the academic community.

Many others faculty members have contributed to molding my research. Much thanks goes to my committee members Ileana Streinu and David Kulp. Ileana helped to teach me how computational geometry can be used in protein folding. David was a big help in setting up the compbio computer cluster, without which this dissertation would never exist. Additional faculty members I would like to thank include Erik Learned-Miller, Rod Grupen, and Robin Popplestone for the valuable advice I've received over the years.

In addition to the faculty, I want to thank the computer science staff for making everything run smoothly. Specific thanks to Gary Rehorka and Valerie Caro for keeping all the computers running, and Leeanne Leclerc and Priscilla Scott for keeping me funded and on-track to graduate.

I would also like to thank my fellow graduate students, Dubi Katz, Jackie Feild, Filip Jagodzinski, Ines Putz, Nasir Mahmood and everyone else who passed through the Robotics and Biology laboratory over the years. Thanks for all the editing, and talks you helped with. In addition I want to thank my lab mates from the Gierasch lab for helping me to understand proteins better, especially, Beena Krishnan, Rob Smock, Ken Rotondi, and Joanna Swain.

I could never have graduated without great friends pushing me along. I have many fond memories of all my housemates at 92 cowls. Thanks guys. Lunch would have been boring without the gang. My softball team the Truculent Turkeys made summers exciting. Individual thanks goes to Brian Jordan, Hossein Baghdadi, Alex de Geofroy, Sarah Osentoski, Gene Novark, Elizabeth Russell, Ashvin Shah, Aaron St. John, Audrey St. John and Matt Rattigan for great conversations and advice over the years.

Most importantly, I want to thank my family and girlfriend. My mother, Bev Brunette, started my interest in proteins way back in high school as my biology teacher. My father, Tom Brunette, has always been there with a supportive e-mail to help me along. Much thanks goes to my wonderful, patient girlfriend Sarah Shepard, for her motivation, encouragement and help editing this thesis.

To everyone I know at University of Massachusetts: thanks for the memories, you made this Ph.D. a worthwhile and fun experience.

Finally, I am grateful for the support of the funding agencies. They really did make this work possible. This work was supported in part by the National Institutes of Health (NIH) under grant NIGMS 1R01GM076706 and by National Science Foundation (NSF) under grants CNS 0551500 and CCF 0622115.

ABSTRACT

ADAPTIVE BALANCING OF EXPLOITATION WITH EXPLORATION TO IMPROVE PROTEIN STRUCTURE PREDICTION

MAY 2011

TJ BRUNETTE

B.S., STATE UNIVERSITY OF NEW YORK AT GENESEO

M.S., UNIVERSITY OF MASSACHUSSETTS AMHERST

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Oliver Brock

The most significant impediment for protein structure prediction is the inadequacy of conformation space search. Conformation space is too large and the energy landscape too rugged for existing search methods to consistently find near-optimal minima. Conformation space search methods thus have to focus exploration on a small fraction of the search space. The ability to choose appropriate regions, i.e. regions that are highly likely to contain the native state, critically impacts the effectiveness of search. To make the choice of where to explore requires information, with higher quality information resulting in better choices. Most current search methods are designed to work in as many domains as possible, which leads to less accurate information because of the need for generality. However, most domains provide unique, and accurate information. To best utilize domain specific information search needs to be customized for each domain. The first contribution of this thesis customizes

search for protein structure prediction, resulting in significantly more accurate protein structure predictions.

Unless information is perfect, mistakes will be made, and search will focus on regions that do not contain the native state. How search recovers from mistakes is critical to its effectiveness. To recover from mistakes, this thesis introduces the concept of adaptive balancing of exploitation with exploration. Adaptive balancing of exploitation with exploration allows search to use information only to the extent to which it guides exploration toward the native state. Existing methods of protein structure prediction rely on information from known proteins. Currently, this information is from either full-length proteins that share similar sequences, and hence have similar structures (homologs), or from short protein fragments. Homologs and fragments represent two extremes on the spectrum of information from known proteins. Significant additional information can be found between these extremes. However, current protein structure prediction methods are unable to use information between fragments and homologs because it is difficult to identify the correct information from the enormous amount of incorrect information. This thesis makes it possible to use information between homologs and fragments by adaptively balancing exploitation with exploration in response to an estimate of template protein quality. My results indicate that integrating the information between homologs and fragments significantly improves protein structure prediction accuracy, resulting in several proteins predicted with <1 $\mathring{A}$ RMSD resolution.

TABLE OF CONTENTS

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

CHAPTER 1

INTRODUCTION

Many important problems in science and engineering require high dimensional search. Some examples include wind turbine design, asset allocation, chess, structure learning of graphical models, and protein structure prediction. For search problems, each additional variable increases the number of possible solutions exponentially, leading to a huge search space after just a few variables. For example, in chess looking 7 steps into the future reveals approximately 6 billion possible chess board configurations. Only recently have the fastest computers in the world been able to address search to the scope needed for chess. The protein search problems addressed in this thesis are much larger than chess. Even in the foreseeable future, all the computers in the world would still be insufficient to explore every possible solution for problems that are as large as those discussed in this thesis.

Since it is computationally intractable to search everywhere, the only way to find a near optimal solution is to carefully choose where to explore. Certain areas of space are more likely to contain the solution (native state) and to be efficient search must focus exploration on these relevant regions. To identify which regions are most likely to contain the solution, information is required.

Most current approaches to search are applicable to all search domains and as a result must use information applicable to all domains. However, each domain provides unique sources of information that are more descriptive than the information applicable to all domains. Since information is key to choosing where to explore, it critical for search to be designed with domain specific information. The first con-

tribution of this thesis is to develop a search customized for the information sources available to protein structure prediction.

Unless information is perfect, search will eventually make a mistake and direct exploration into a region that does not contain the solution. The only way to identify when a mistake has been made is to find a new piece of information that refutes the inaccurate information that led to the mistake. For example, when searching for a global minimum in a function that contains many local minimum, a mistake has been made if search is trapped in a local minimum. Search makes these mistakes when it exploits inaccurate information. To acquire new information and move beyond the local minimum, search must become more exploratory in new regions of conformation space.

Information sources vary widely in quality. High quality information is less likely to cause search to make a mistake and therefore search behavior can be very exploitative. Lower quality information is more likely to cause search to make a mistake, which wastes resources in a region that does not contain the solution. To compensate for the increased risk incurred when using lower quality information, additional information must be gathered to identify mistakes. New information is gathered by search becoming more exploratory. The second contribution of this thesis is to adaptively balance exploitation and exploration in response to information quality.

1.1 Protein structure prediction

Fast and accurate determination of protein structure is one of the most important challenges in biology. Determining protein structure quickly and accurately will help researchers diagnose and cure diseases, design drugs with fewer side effects, aid nanotechnology manufacturing, and increase our understanding of molecular biology.

Current methods for protein structure determination are slow and difficult to apply on all proteins. Researchers presently rely on X-ray crystallography and nuclear mag-

netic resonance (NMR) spectroscopy to determine protein structures. These methods are labor- and cost-intensive. As a result, there are approximately 12 million protein sequences known, but only fifty-thousand structures. Protein sequence estimates are from the combined size of the global ocean survey (GOS) and non redundant protein (NR) databases, while the structure count is from the protein database (PDB) [55]. In response to this imbalance, the government has spent over 270 million dollars on the Protein Structure Initiative. This investment has resulted in 1100 new structures, 700 of which are unique. At that rate it would cost approximately 3 trillion dollars to determine the structure of all known proteins [46].

Computational protein structure prediction has the potential to help close the gap between the number of known sequences and structures. Although methods have progressed substantially over the years, the problem of protein structure prediction is far from solved [8, 71, 53, 12, 76, 41, 29, 21]. The main barrier to improved protein structure prediction is the inadequacy of conformation space search techniques [9]. Due to the vast size of conformation space and the ruggedness of the protein energy landscape, existing search methods fail to find the native state in all but the smallest proteins. For large proteins, existing search methods are only effective if search is constrained to the region containing the native state. If the native state is in a different region of space, search fails. To enable general, accurate structure prediction, it is therefore of paramount importance to devise methods capable of harnessing all available information. Although there are many information sources specific to proteins, this thesis primarily utilizes two which are described below.

The first source of information is the score (energy) function which compares protein conformations. In most search domains there is only one score function, but for protein folding there are two. First search uses a coarse grained, backbone-only energy function, and then a very accurate all-atom energy function [59, 62, 56]. The coarse grained energy function is inaccurate, checking only for collisions. As search

progresses, the coarse grained energy function adds terms that consider secondary structure, the residue environment, and inter-residue pairing. Toward the end of search, an all-atom energy function is used, which measures protein energy after all side chains have been added. The all-atom energy is tuned using experimental structures deposited in the PDB. Since only native structures are deposited into the PDB, the all-atom energy function is only accurate when the structure is native-like. With native-like proteins occurring only after many search steps, the all-atom energy function can only be used at the end of search. Chapter 3 will describe how we improve search effectiveness by using information acquired from the all-atom energy function to evaluate non-native structures.

The second, and most effective protein-specific source of information is evolutionarily related proteins. When two proteins evolve from a common ancestor, they are likely to have similar native structures. Evidence suggests that the currently available evolutionary information could solve protein structure prediction. For example, no new protein motifs have occurred during the last several years of the Computational Assessment of Structure Prediction experiment (CASP) [15]. Additional evidence comes from Zhang and Skolnick who showed that 99.8% of single-fold proteins have a corresponding homolog within $6\mathring{A}$ RMSD, and 97% have a homolog under $4\mathring{A}$ RMSD [80]. Throughout this dissertation I will refer to homologs as sequence homologs or structural homologs. Sequence homologs can be found by methods that use the amino acid sequence to find homologs. Structural homologs are when a structural similarity exists but not detectable by sequence homology methods.

Despite the great potential of evolutionary information, protein folding remains unsolved because it is difficult to identify which information is relevant. There are thousands of proteins which could serve as the template protein, each with a vast number of ways to align to the target protein. The key to using potentially inaccurate template proteins is to adjust search behavior based on the quality of the template.

When a template protein is high quality, search should exploit that template protein and use it to guide exploration to the native state. When a template is lower quality there is a higher likelihood of making a mistake. The only way the mistake can be avoided is by gathering additional information to refute the incorrect template protein. In this way, search performance needs to adaptively balance exploitation of a template protein with exploration of new templates.

Current approaches with template proteins either use the whole structure (homology modeling) or short protein fragments (de novo). Homology modeling relies on the presence of a template protein that is aligned over most of the target protein. If the structure is poorly matched, homology modeling results in inaccurate structure prediction. De novo protein structure prediction combines a library of protein fragments with the use of an elaborate search procedure. Adequate exploration of conformation space requires a very large amount of space be visited. Thus, de novo methods only explore a sufficient amount of conformation space in the smallest proteins.

Homologs can be considered to provide very high quality information, resulting in a search that is almost pure exploitation. On the other hand, protein fragments offer very low quality information and result in a search that is almost pure exploration. Homologs and short fragments represent extremes on the spectrum of information that can be extracted from known protein structures. Currently, no method exists to use information between these two extremes due to the enormous variety of ways the information could be used, most of which are incorrect. Since information is critical to search performance, the ability to use information anywhere on the spectrum between homologs and fragments needs to be explored and remedied. This thesis makes it possible to use information between homologs and fragments by adaptively balancing exploitation with exploration in response to an estimate of template protein quality.

1.2 Overview of results

The first method developed in this thesis builds an approximate, partial model of the energy landscape using highly accurate information obtained from the all-atom energy function. I call this method model-based search. Model-based search aggregates information in a model as it progresses, and in turn uses information in the model to guide exploration toward the regions most likely to contain the native state. I validate model-based search by predicting the structure of 32 proteins, ranging in length from 49 to 213 amino acids. My results demonstrate that model-based search is more effective at finding low-energy conformations in high-dimensional conformation spaces than existing search methods. The reduction in energy translates into structure predictions of increased accuracy.

I further validated model-based search by taking part in the Critical Assessment of Techniques for Protein Structure Prediction experiment in 08. (CASP 8). Unlike most other methods taking part in CASP model-based search did not have the ability to retrieve homology information. This was a major handicap, as homology information, if available, renders conformational space search substantially easier. Due to this limitation of my first-ever entry into CASP, the method still did very well. In the free modeling competition the method was ranked **6th out of 74** in one evaluation scheme and **14th out of 69** in a second evaluation scheme. On one protein model-based search made the most accurate prediction.

The second method developed in this thesis uses evolutionary information anywhere on the spectrum between homologs and fragments. I refer to this method as balanced exploitation exploration template search (BEETS). Information between homologs and fragments is often wrong, so my method relies on adaptive balancing of exploitation with exploration. I validate BEETS by predicting the structure of 36 proteins, ranging in length from 54 to 139 amino acids. My results indicate that BEETS significantly improves the accuracy of protein structure prediction, predict-

ing several proteins with <1 $\mathring{A}$ RMSD resolution. BEETS will be validated in CASP during Summer 2010.

1.3 Thesis outline

The remainder of my thesis is organized as follows.

Chapter 2 reviews related work.

Chapter 3 describes using a model of the protein energy landscape to guide search; this method is called model-based search. Model-based search is much more effective than previous methods at modeling regions in the energy landscape. Enabled by accurate region modeling, my method assesses which region is more likely to contain the native state using information from the accurate all-atom energy function. Model-based search is shown to significantly improve the accuracy of protein structure prediction.

Chapter 4 integrates evolutionary information anywhere on the spectrum between homologs and protein fragments. Use of information between fragments and homologs is made possible by reasoning about information quality and adaptively balancing exploitation with exploration. This is the first successful use of information between homologs and fragments, resulting in significant improvements to protein structure prediction accuracy.

Finally, chapter 5 concludes this dissertation and suggests future sources of information that will further improve search.

1.4 Contributions

This thesis makes two contributions.

The first contribution is to develop a search customized for the information sources available to protein structure prediction.

The second contribution is to adaptively balance exploitation and exploration in response to information quality. This is the first time exploitation and exploration have been adaptively balanced during the search process.

CHAPTER 2

RELATED WORK

2.1 Introduction

The conformational space of proteins is too large to be searched exhaustively [32]. This is true even for small proteins. Conformation space search methods thus have to focus exploration on a small fraction of the search space. The ability to choose appropriate regions, i.e. regions that are highly likely to contain the native state, will critically impact the effectiveness of search. Information is required to make the choice of where to explore, with higher quality information resulting in better choices. Unless the information is perfect mistakes will be made. How search recovers from mistakes also critically effects search quality.

In this thesis I argue that search improves with additional information, and better ways to recover from mistakes. My specific insights are the following: additional information can best be acquired by specializing search for each individual domain, and mistakes can most effectively be recovered from by adaptively changing search behavior between exploitation and exploration. In this section I show support for my arguments by looking at how existing methods have improved search through partial realization of my insights.

2.2 Related work

The most basic approach for conformation space search is the Metropolis Monte Carlo method [39]. This method remembers only a single piece of information, namely the energy value of the current step. Based on this information, the next exploration

step is accepted if the new conformation is lower in energy, and if the energy increases, the new conformation is rejected with probability proportional to the increase in energy.

The Metropolis Monte Carlo method is susceptible to making a mistake and getting trapped in local minima. Much of the ongoing work on conformation space search aims to overcome this problem. To increase the chances of escaping small local minima, simulated annealing [28, 47], slowly transitions from exploitation to exploration during search. When search begins, little information is known, but each step in search contributes a small amount of information. As more information becomes known search performance improves by exploiting the newly available information. Eventually search converges on a local minima. To escape the local minima exploration must be increased. Simulated annealing methods increase exploitation by restarting search, thus discarding all information that has been acquired.

One approach to retain previously acquired information and escape local minima is to randomly transition between exploitation and exploration. When search is biased toward exploration a larger region of conformation is explored, while a bias toward exploitation favors exploring a smaller region in more detail. The goal of randomly transitioning between exploitation and exploration is search will hop from one region to another when exploration is favored, while focusing search on a single low energy region when exploitation is favored. The problem is that randomly transitioning between exploitation and exploration results in search exiting deep minima before they are fully explored, causing search to never find the best solutions. Examples include replica exchange [68, 58], jump walking [16], multi-canonical jump walking [74], and others Monte Carlo methods [36].

The balance between exploitation and exploration could more effectively be established if done in response to the effect exploitation or exploration has on search. Evaluating the effect of increased exploitation or exploration requires information.

The only source of information used by current approaches is an evaluation of whether search has converged. When search has converged there is no additional information in that region of space, so search needs to explore new regions to gather additional information. Methods that adjust the balance in response to convergence include basin hopping [34, 72], max-min ant farm [67], reactive tabu search [3] and genetic algorithms that have been combined with tabu search [1].

An additional way to prevent search from making mistakes is to reduce the number of possible mistakes. For Metropolis Monte Carlo (MC) methods the number of mistakes that search could possibly make is lowered by reducing the number and depth of local minima. To achieve this, one can smooth the protein's energy landscape [52, 77]. This will have the desired effect of making search more likely to escape local minima, but it will invariably introduce inaccuracies in the energy landscape. These inaccuracies are due to the merging or shifting of minima or may arise as a result of rank inversions [50]. In principle, smoothing is similar to simulated annealing methods: it makes it easier to overcome the energy barriers between local minima, in particular during the early phases of the search. This insight has been confirmed by experimental studies [19].

Current methods for protein structure prediction employ smoothing in conjunction with Metropolis Monte Carlo-based search methods. Smoothing can be achieved with multi-resolution energy functions. Early stages of the search are conducted in a simpler, backbone-only energy function. As search progresses, the energy function becomes increasingly accurate, until an all-atom energy function is used to evaluate the final decoys [59]. MC methods use the backbone-only energy function to assemble the majority of a protein's structure, and the all-atom energy function to make smaller structural changes and to evaluate prediction quality.

The combination of smoothing and Metropolis Monte Carlo-based search methods has proven very successful in practice and is currently the most widely used approach

to conformation space search [41]. However, this combination of the two methods also inherits their disadvantages: MC methods use a very limited amount of information (only the current energy value) to guide exploitation, which due to smoothing is likely to be inaccurate, leading to search exploring the wrong region. The method developed in chapter 3 of this thesis avoids the problems of smoothing by using the most accurate information available (obtained using the all-atom energy function) to select the region of conformation space to search.

Genetic algorithms [20] introduce the idea of maintaining multiple samples and exchanging information among them. This improves on the amount of information maintained by MC-based methods. Tabu search [17, 69, 49] maintains aggregate information about the entire history of the search to exclude so-called tabu regions from further exploration. These methods demonstrate that the information obtained during search can be beneficial in informing further exploration. Similar ideas can be found in conformation space annealing (CSA) [31] and conformation-family Monte Carlo (CFMC) [54], two conformation space search methods developed specifically for protein structure prediction. These methods monitor the state of multiple concurrent searches in order to ensure broad coverage of the search space.

Search is one of the foundational topics in the study of artificial intelligence (AI) [61]. It is thus not surprising that the idea of using information obtained during search to guide search has been studied extensively in AI. These methods aggregate information in what I will call a "model." They then use this model to select those regions of the search space for exploration that are most likely to contain the sought minimum. Two such methods are STAGE [7] and MIMIC [5]. These methods use a model to make predictions about regions of the search space, even regions that have not yet been explored. STAGE builds a problem specific model using either linear regression or least squares TD(λ). MIMIC uses pairwise conditional probabilities $P(Xi|Xj)$ and unconditional probabilities $P(Xi)$ to model the true joint distribution $P(X)$. Both

STAGE and MIMIC are successful in problems with a small number of variables with few inter-dependencies. Proteins, however, have hundreds of degrees of freedom with with complex inter-dependencies. MBS uses a model customized for the unique properties of proteins to better capture critical features of the energy landscape.

Another relevant area of research within artificial intelligence is active learning. In active learning, the goal is to learn a function from examples. The learner is able to interactively select training examples so as to maximize learning progress [37, 13]. Active learning requires a model to maintain information about the examples seen so far. But it also requires a strategy to select the best next example. Applied to conformation space search, such a strategy would redirect search from one region to another in response to the information obtained.

My thesis draws inspiration from AI search and active learning, applying the relevant concepts in the context of protein structure prediction. The most important distinctions between typical active learning domains and protein structure prediction is that information comes both from an energy function and from evolutionarily related proteins. Applying the principles of active learning to two sources of information leads to an approach that simultaneously models both information sources, and coordinates resources to efficiently acquire information from both sources.

Evolutionary information comes from either the whole protein (homology modeling) or protein fragments (de novo approaches). Homology modeling relies on sequence similarity between the protein to be predicted and a template protein to infer a common evolutionary origin, and hence a similar structure [30, 70, 25]. Use of a template constrains the degrees of freedom that are explored from all parts of a protein to only those parts where the sequence varies from the template [14, 35, 56]. Homology modeling focuses all computational resources on a very small region of space. If the native state is not in that region, search will fail to find a reasonable solution.

De novo prediction methods combine a library of protein fragments with the use of an elaborate search procedure. A protein fragment provides one possible conformation for a short amino acid sequence. For each sequence, hundreds of protein fragments are gathered from the protein data bank. These are assembled using search [11, 64, 26, 2, 18, 65, 78]. To adequately explore conformation space requires a vast amount of space be visited. Thus, de novo methods only explore a sufficient amount of conformation space for small proteins.

Current homology modeling techniques are only applied when a homolog can be accurately identified, otherwise search uses de novo protein structure prediction. However, there is significantly more information that falls in the range between homologs and fragments. Early work has blurred the distinction between homology modeling and de novo structure prediction by adjusting the composition and length of fragments [81].

The key to using homologs when they are difficult to identify is the development of a search strategy that can deal with the inevitable mistaken identifications. Unless the homolog can be perfectly identified, mistakes will be made, and search will focus on regions that do not contain the native state. How search recovers from mistakes is critical to its effectiveness. To recover from mistakes, my thesis introduces the concept of adaptive balancing of exploitation with exploration. Adaptive balancing of exploitation with exploration allows search to use information only to the extent to which information leads to the discovery of new low energy structures or exploration of new regions in conformation space.

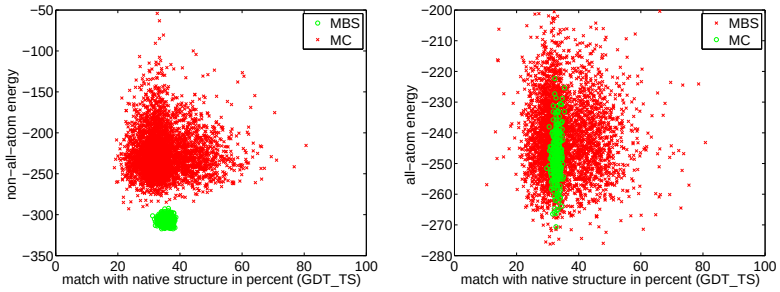
CHAPTER 3

GUIDING CONFORMATION SPACE SEARCH WITH AN ALL-ATOM ENERGY POTENTIAL

The challenge of search in high-dimensional conformation spaces is exacerbated by the fact that any search method—no matter how effective—can only achieve accurate results if it relies on accurate information. In the case of protein structure prediction, the most accurate information applicable to all proteins is captured by the need for the computationally expensive all-atom energy functions. Most existing protein structure prediction methods, however, rely on simplified, non-all-atom energy functions to alleviate the difficulties of conformation space search. I believe this inherently limits their ability to perform accurate structure prediction.

To illustrate the importance of accurate information for protein structure prediction, I predict the structure of retinoic acid binding protein (136aa) using two different prediction methods. Both prediction methods attempt to find low-energy conformations in an accurate, but *non*-all-atom energy function. The predictors differ in the conformation space search method they employ. The first predictor uses model-based search (MBS), my new conformation space search algorithm introduced in this chapter. The second predictor uses simulated annealing Monte Carlo search (MC). Figure 3.1(a) compares the resulting predictions, showing that MBS finds lower-energy structures in the non-all-atom energy landscape than MC.

I evaluate the predictions obtained by both algorithms using an all-atom energy function. The resulting scatter plot is shown in Figure 3.1(b). I see that MBS' predictions are energetically indistinguishable from those obtained using MC. Even though



(a) Assessment of structure predictions using an inaccurate, non-all-atom energy function

(b) Assessment of the structure predictions using an all-atom energy function

Figure 3.1. Discrepancies between the non-all-atom and the all-atom energy function illustrate that accurate conformation space search must rely on all-atom information. (a) Structures found by model-based search (MBS) in the non-all-atom energy function are lower in energy than those found by a Monte Carlo-based method (MC). (b) These structures become energetically indistinguishable when evaluated in the all-atom energy function.

MBS is able to find lower-energy structures in the non-all-atom energy landscape, indicating more effective conformation space search, this advantage vanishes in the all-atom landscape. This implies that more effective search alone will not necessarily lead to improved prediction accuracy. To take advantage of more effective conformation space search, it is necessary to search a more accurate energy landscape.

In this chapter, I present model-based search, a new conformation space search method for finding minima in protein energy landscapes. Model-based search combines highly effective conformation space search with the ability to perform search using highly accurate all-atom energy information. The improvements afforded by my approach are based on two main contributions. First, my method is more effective than previous methods at identifying and selecting the appropriate regions to focus resources. Second, enabled by the first contribution, my method is able to obtain high-quality all-atom information without incurring a significant performance penalty.

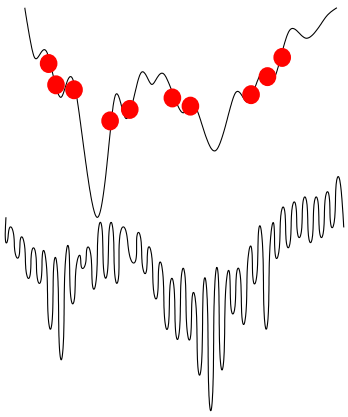
Experiments demonstrate that the combination of more effective conformation space search and highly accurate information results in the prediction of structures of lower energy than those predicted by one of the leading structure prediction protocols. I also show that this reduction in energy translates into more accurate structure predictions. Predictions for which reduced energy does not lead to improved prediction accuracy identify errors in the energy function and thus may lead to the improvement of these functions.

3.1 Model-Based Search

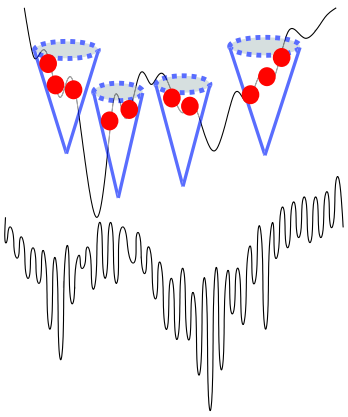
Effective conformation space search must guide exploration towards regions of conformation space likely to contain the global minimum. Consequently, the effectiveness of search is based on how accurately the relevant regions can be identified. The effectiveness of this identification, in turn, depends on the usage and accuracy of information.

I refer to the representation of relevant regions as a *model* of the energy landscape. At any point during the search, this model will represent an approximation to a small part of the energy landscape. The model contains important information that is leveraged by model-based search to guide exploration towards relevant regions of conformation space. Due to the central role of this model in making the conformation space search method accurate and efficient, I refer to my search method as *model-based search*.

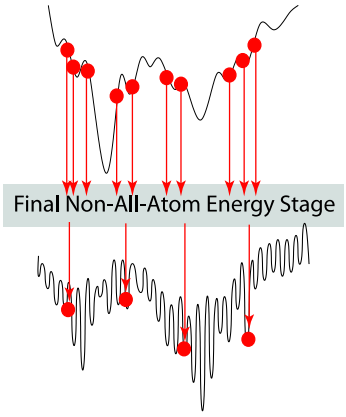
Model-based search incrementally refines an initial coarse model of conformation space by incorporating new information obtained during an ongoing search. Information quality is critical to direct resources toward the correct regions of space. The acquisition of high quality information is driven by three core algorithmic elements described below. Figure 3.2 illustrates the use of these algorithmic elements for a single iteration of model-based search.



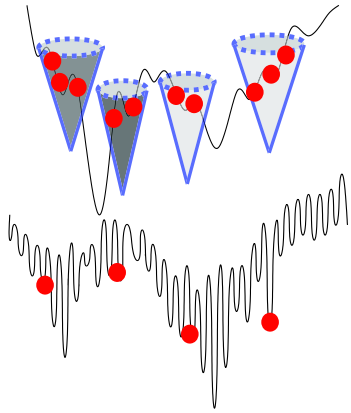
(a) Each stage of model-based search (MBS) uses a model consisting of 2,000 samples in the approximate energy function. The model in the first stage of MBS consists of samples generated by short Monte-Carlo runs. In subsequent stages of MBS, the model contains samples from the preceding stage.



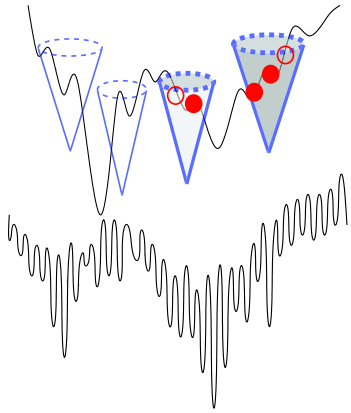
(b) MBS groups the samples into funnels that accurately capture coherent regions of the energy landscape. This expressive characterization of conformation space regions enables the computationally efficient assessment of the relevance of a region based on the all-atom energy function.



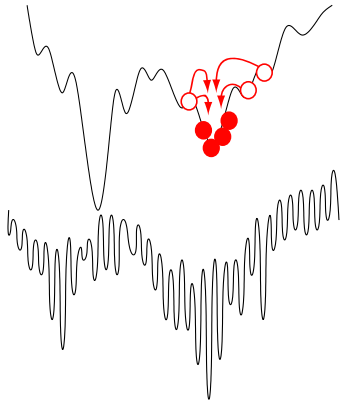
(c) To assess the relevance of a region as accurately as possible, I acquire information from the all-atom energy function. The samples of all funnels are evaluated with Monte Carlo trajectories to Rosetta's final non-all-atom stage. For each funnel, the lowest-energy trajectory is continued through Rosetta's all-atom energy function.



(d) The all-atom scores obtained in the previous step are used to determine the relevance of individual funnels. This information will guide the search towards the most promising regions of conformation space. In the image, region relevance is illustrated by the color of the funnel. Lighter funnels correspond to more relevant regions.



(e) Based on funnel relevance, the model is pruned by discarding entire funnels and most of the samples within the remaining funnel. I retain the best half of the funnels and the top four samples within each of the remaining funnels. Note that MBS discards funnels corresponding to low-energy regions in the approximate energy function; while keeping funnels with low-energy in the all-atom energy function.



(f) MBS replaces discarded samples using short Monte Carlo trajectories starting at the samples remaining in the model. Eighty percent of the starting locations for these trajectories are chosen evenly across funnels; twenty percent of the starting locations are chosen proportional to the relevance of the funnel.

Figure 3.2. Step-by-step illustration of a single stage of model-based search. Each image contains two energy landscapes. Rosetta's approximate energy function is shown on the top, Rosetta's all-atom energy function on the bottom. Note that the global minimum in the approximate energy landscape does not correspond to the global minimum in the all-atom energy landscape.

1. **Characterization of Regions as Funnels:** At the core of model-based search is a method to identify meaningful regions of conformation space. Instead of viewing samples in isolation, my method groups samples so as to capture the funnel-like quality of the landscape (see Figure 3.2(b)). Section 3.1.1 describes how funnels can be computed efficiently, while accurately capturing coherent regions of the energy landscape with similar biological characteristics.

2. **Assessment of Funnel Relevance:** To assess the relevance of a region as accurately as possible, I acquire accurate information about that region. Since my method of determining regions ensures that all samples in a region share biological characteristics, we can draw conclusions about the relevance of an entire region based on high-accuracy information about carefully chosen samples. Figures 3.2(c) and 3.2(d) illustrate the efficient acquisition of information and how that information is used to assess the relevance of a region.

3. **Coordination of computational resources:** Once meaningful regions have been identified and the quality of each region has been assessed, I use this information to distribute computational resources in accordance with this assessment. Figures 3.2(e) and 3.2(f) illustrate this process.

In this section, I present model-based search as a general optimization method for high-dimensional spaces, making as few domain-specific assumptions as possible. The optimization algorithm is applicable to problems that exhibit spatial coherence and global variation. Spatial coherence means that the quality of a specific point in the solution space reveals information about its immediate neighborhood. Global variation means that there are significant differences between "good" and "bad" solutions in the search space. Together with spatial coherence, this implies that there are "good" and "bad" regions of space. The assumptions of spatial coherence and global variance, I believe, are quite general and are shared by many real-world problems.

I make one domain-specific assumption, namely that I are searching a series of related energy functions. Search of conformation space begins in a computationally efficient, low-accuracy energy function and incrementally progresses to a computationally costly, high-energy function. This technique is commonly applied in protein structure prediction [6]. This assumption is only required for the part of my search method described in Section 3.1.2. The overall search method remains valid even if this assumption does not hold and only a single energy function is searched.

The following three sections provide detailed descriptions of the algorithmic elements; Section 3.2 augments the description provided below with details about the implementation.

3.1.1 Characterization of regions as funnels

The notion of a conformation space region permits us to reason about volumes of space as a single entity. This is more effective than reasoning about individual samples. To reason about an entire region in a meaningful way, however, the conformations in that region have to share some relevant property. Only then will it be possible to assess the relevance of a region as a whole.

Some existing clustering techniques used in CSA [31], CFMC [54], and SPICKER [79], incorporate high-dimensional spheres, or hyper-spheres, to describe regions of conformation space. Such a region is described by a point in conformation space (the center of the sphere) and a radius, usually given by the backbone RMSD in Ångstrom between two conformations. Such a hypersphere is a simple representation of conformation space volume but it is unlikely to exclusively capture parts of space that share a relevant property. The extent of a meaningful region will vary greatly along the different dimensions of the space. This holds true in particular in protein energy landscapes, in which the motion of some degrees of freedom can cause very large variations in energy, whereas other degrees of freedom can move significantly without a

major energetic effect. Consequently, a hypersphere will include regions with different properties, cause overlap between distinct regions, or even merge distinct regions of conformation space. Based on this inaccurate representation of conformation space regions, it is difficult to guide search effectively using conformation space techniques.

I propose the notion of funnels as a more accurate representation for conformation space regions. I know that the energy landscape of a protein contains many such funnels. The funnel shape implies that a Metropolis Monte Carlo run started at a point in a funnel has a higher probability of leading to the bottom of that funnel than of leaving it. I can thus view the entire funnel as the domain of attraction for the energetic minimum of the funnel. Hence, funnels represent a region of space in which all points share a property that is important for search: they can all be associated with the same local minimum in the energy landscape. Based on this well-established fact, I believe that funnels provide a characterization of conformation space regions appropriate for guiding search (see Figure 3.2(b)).

I identify funnels by exploiting the following simple observation: In low-energy regions of a funnel the spatial density of the samples resulting from Monte Carlo-runs will be high. As I approach the ridge of a funnel, the spatial density of samples decreases. The spatial density of samples obtained from Metropolis Monte Carlo thus captures the extent and energetic variation of the lower-energy region of the funnel. It is this lower-energy region that is most helpful for guiding search.

MBS identifies funnels using a heuristic clustering method. This method is computationally efficient and, more importantly, sensitive to density variations of samples in different dimensions of conformation space. Due to its ability to identify directional variations in sample density, my method can identify clusters of arbitrary shapes and varying local densities, even when they are close to each other. The method does not impose a particular representation for the funnel, such as a hypersphere, but lets the data determine the extent and shape of the conformation space region.

I now describe the details of the funnel finding algorithm. Starting with a set of conformation space samples, the lowest energy sample is selected as the root of a tree. I build a tree by adding samples to the tree in order of increasing distance to the root. A new sample will be connected to the closest node of the tree, as long as the distance between the new sample and the closest node does not exceed the average length of edges on the path between the node and the root of the tree by more than a constant factor. This insertion operation is illustrated in Figure 3.3. The tree-building algorithm terminates when all remaining samples are too far away from nodes in the tree to be added. The computed tree represents a funnel; the root of the tree is at the bottom of the funnel.

This procedure is repeated, starting with the lowest-energy sample among the remaining samples, until all samples have been processed into trees.

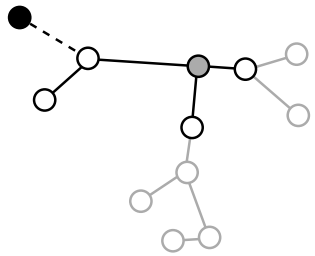


Figure 3.3. Building a funnel from samples: The lowest-energy sample (solid gray) is chosen as the root of a tree. Samples are inserted in order of increasing distance to the root. To insert a new sample (black), the distance between the new sample, the root, and all of its children is considered. If the root is closest to the new sample, the sample is inserted as a child of the root. Otherwise, the process recurses on the child closest to the new sample. When a leaf is reached, the new sample is inserted as its child. Following this procedure, many sub-trees (gray outline) remain unexamined during the insertion.

I now have obtained trees of samples, each of which represents a funnel. Each funnel captures a meaningful region of conformation space. The shape of this region is determined by the properties of the underlying energy landscape. The connectivity of the tree and the degree of its vertices capture additional information about proximity,

compactness, and transition characteristics between nearby samples. This illustrates the benefit of viewing samples in context rather than in isolation: the context reveals additional information highly relevant to my goal of effectively searching conformation space.

My clustering method achieves expected $O(n \log n)$ time complexity by sacrificing provable correctness of the clustering results. In my experience, the gain in efficiency by far outweighs the consequences of the slight inaccuracies in the resulting clusters.

3.1.2 Assessing funnel relevance

An accurate assessment of the relevance of a region is essential for effective conformation space search. Many of the funnels will prove to be irrelevant for search and should be discarded from my model. Among the remaining regions, I would like to allocate computational resources proportional to the estimated relevance of a region. Therefore, to guide conformation space search most effectively, we need a way to accurately evaluate the relevance of a conformation space region.

The tree-based representation of funnels enables an efficient and highly accurate assessment of region relevance. The tree representation of samples provides us with a simple way of determining the size, shape, and sample density of funnels. These properties of the region, as well as the energy values of individual samples, reveal information about the relevance of a funnel. They can be combined in a variety of ways to estimate relevance. In this section, I do not focus on how to combine all available information to assess region relevance but instead on how the accuracy of sample energy evaluation can be improved. The determination of region relevance based on all available information will be the subject of future work.

Model-based search enables the assessment of region relevance based on highly accurate all-atom energy evaluations. The algorithm achieves this by leveraging the funnel-based representation of the model. Regions in the model have been determined

based on the property that local searches from most conformations in a region will be investigating the same minima. I exploit this property to assess the relevance of an entire region by assessing the relevance of several conformations inside the region. Due to the aforementioned property of the region, the quality (energy) of these samples provides information about the relevance of the entire region (see Figures 3.2(c) and 3.2(d)).

Model-based search assesses the relevance of a region by determining the all-atom scores for the lowest-energy non-all-atom samples in the region. The all-atom score of a sample is determined by performing a Metropolis Monte Carlo run through increasingly accurate energy functions, including a final, highly accurate all-atom energy function (the details of this computation are described in Section 3.2). The best score of all evaluations determines the score of a region.

As the experiments presented in Section 3.3 will demonstrate, this procedure for assessing the relevance of a region greatly improves the accuracy and efficiency of conformation space search. The accuracy is improved because the assessment of relevance is based on the most accurate source of information available: an all-atom energy function. This accurate assessment of relevance would not be computationally feasible for all conformations generated during an entire conformation space search. By using a few costly all-atom computations to judge the relevance of entire regions of conformation space, however, the amortized computational cost is negligible. Information is leveraged very effectively to guide search towards important regions of conformation space.

3.1.3 Coordination of resources

Model-based search allocates computational resources to regions based on their estimated relevance. If the assessment of region relevance were perfect, only a single region should be explored further. No assessment of region relevance would lead to an

equal exploration of all regions. Model-based search attempts to find a middle-ground between these two extremes so as to guide search effectively while accounting for inaccuracies in assessment of region relevance by spreading computational resources.

Model-based search discards irrelevant regions and redundant samples to maintain computational efficiency (see Figure 3.2(e)). Available computational resources are divided into two parts. The first part is divided equally among all regions of the model. The second part of the computational resources is allotted to a region proportional to its estimated relevance. To replace discarded samples, model-based search initiates short Metropolis Monte Carlo trajectories from the samples remaining in the model. The resulting samples are added to the model (see Figure 3.2(f)).

3.2 Implementation

3.2.1 Integration with rosetta

The focus of my research is the development of effective conformation space search techniques. To leverage existing software infrastructure, I have integrated model-based search with Rosetta [6, 59], a leading method for protein structure prediction that has repeatedly performed well in the CASP competition [44, 42, 43]. My implementation replaces the simulated annealing Metropolis Monte Carlo search method implemented in Rosetta with model-based search, allowing us to rely on Rosetta's energy function, local search methods, and infrastructure for representing proteins, etc.

Due to my integration with Rosetta, model-based search inherits the following algorithmic features. Rosetta uses the fragment assembly approach to reduce the size of the search space. Initial backbone-only samples are generated by setting all ϕ and ψ angles of the backbone to zero. Local search for low-energy conformations is started from this point in conformation space. The local search, based on the Metropolis Monte Carlo method, progresses in a number of stages. As the search progresses

through the different stages, the move set changes, the number of local search steps are varied, and the accuracy of the energy function is increased. The initial move set replaces 9-mers of the backbone with candidate structures retrieved from the PDB. The move set then changes to 3-mers and finally to a full angle representation in later stages. The energy function progresses gradually from a coarse-grained low-resolution energy function that considers secondary structure, residue environment, and inter-residue pairing to a full-atom energy function that includes side chains and solvation effects. Additional details about the move sets, and energy functions can be found in the literature [6, 59].

3.2.2 Iterative model refinement

Each iteration of model-based search uses the same move set and energy function as the corresponding stage in Rosetta. Search begins with 2,000 extended structures. The first MBS stage occurs after an initial 4,000 Monte Carlo fragment insertions have been attemped for each sample. The remaining 32000 Monte Carlo steps inside Rosetta are divided into 13 stages based on when terms are introduced into the approximate energy function.

3.2.3 Characterization of regions as funnels

The tree-based algorithm for finding funnels described in Section 3.1.1 only relies on a single parameter: the constant factor that determines whether or not a node is added to the tree. In my implementation I empirically chose that factor to be 1.2. Hence, a node is added to the tree if its distance to the closest node in the tree is less than 1.2 times the average length of edges between the root of the tree and the closest node.

The implementation of the funnel-finding algorithm also terminates tree construction if more than 5% of all samples have been added to the tree. Furthermore, trees of less than 5 samples are merged with the closest funnel. Funnels that are too large are

not helpful in differentiating between different regions of conformation space. Funnels represented by too few samples arise when most funnels have been discovered. The few remaining samples could not be added to any of the previously found funnels. They are likely to be distributed over the entire conformation space and do not represent a meaningful funnel in the energy landscape.

3.2.4 Assessing funnel relevance

Model-based search assesses the relevance of a region by gathering information about what energy level is attainable by local searches started in that region. The exact procedure is described in Section 3.1.2 and illustrated in Figure 3.2(c). To determine an estimate of the attainable energy level, model-based search continues the local searches for all samples in a funnel to the final non-all-atom energy stage in Rosetta. The computational cost of doing this is small, as the energy evaluations in non-all-atom energy functions are computationally efficient. Among the resulting samples, the best five are selected. For each of these, model-based search computes a computationally expensive all-atom energy score after adding side-chains to the backbone. The best of these scores is used as the energy score for the entire funnel.

The searches performed during this evaluation are entirely local; they run through the energy functions associated with the remaining stages of Rosetta, without being influenced by model-based search. To leverage the information obtained during these local searches, we remember a trace of the search for the best 80 full-atom energy evaluations. A trace contains the conformation at the transition points between the different energy functions. Once model-based search has progressed to a particular stage, the model is augmented with the conformations at that stage from those 80 traces.

The current implementation of model-based search estimates region relevance exclusively based on the full-atom energy score. In future research, I will investigate how region relevance can be evaluated by metrics such as funnel size and density.

3.2.5 Coordination of resources

Resource allocation first occurs between funnels. The resources assigned to a funnel are then distributed among the samples within each funnel.

I begin by discarding 50% of the funnels in the model based on their relevance. Eighty percent of the computational resources are distributed evenly among the remaining funnels. The remaining 20% are distributed to funnels proportional to their relevance score. The increased emphasis on particular regions is amplified over multiple stages, increasing the focus on a consistently relevant region at an exponential rate.

Within each funnel I keep the four lowest-energy samples. Eighty percent of the computational resources assigned to a funnel are distributed evenly between these samples; the remaining 20% are distributed proportional to sample score.

3.3 Results and Discussion

In this section, I compare the effectiveness of model-based search (MBS) with that of simulated annealing Monte-Carlo search method (MC) implemented in Rosetta [6, 59]. By comparing with Rosetta, I achieve two objectives. First, since MBS uses the same energy function and local search as Rosetta, I am able to stage a fair test. Second, since the search method of Rosetta is highly optimized for protein structure prediction, I gain a realistic view of MBS's performance in this domain. Rosetta's performance in CASP indicates that the specific implementation of MC is equivalent in performance to other available search methods.

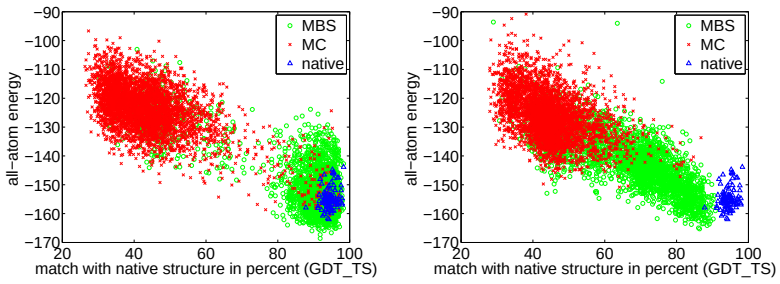
In my evaluation, MBS and MC rely on the same parameters wherever possible. MC and MBS go through a number of stages (see Section 3.2.1); in each stage they use the same move sets, number of local search steps, and energy function. But whereas in MC all samples traverse all stages and these traversals proceed independently of each other, MBS orchestrates these trajectories, stopping some and splitting some into multiple trajectories in later stages. MBS also generates search trajectories for the evaluation of region relevance (see Section 3.2.4). As a result, MBS generates about 3,000 decoys when 2,000 samples are used in each stage of the model. Given the computational overhead of model maintenance in MBS, the computation time required to compute 3,000 MBS samples approximately corresponds to the time required to generate 4,000 MC decoys. Consequently I compare MBS searches with a model size of 2,000 samples with MC searches generating 4,000 decoys.

I would like to emphasize that my experiments are exclusively intended to evaluate the effectiveness of search. The main criterion for the evaluation of my experimental results given in Table 3.1 must therefore be the energy of samples produced by the search. The energy of the native state and for samples produced by MC and MBS are given in the columns labeled E_{Native}, E_{MC}, and E_{MBS}, respectively.

In recent work, Bradley and Baker [8] make highly accurate structure predictions, using an order of magnitude more samples than I use in my experiments. my experiments thus do not give a representative view of the prediction quality obtained by Rosetta. my experiments demonstrate that MBS searches the energy function of Rosetta more effectively than the MC search implemented in Rosetta, given an equal amount of computational resources. The performance increase of MBS relative to MC will become more pronounced when the number of samples is increased, because MBS coordinates the search of conformation space whereas in MC all samples are treated independently.

For my experimental evaluation, I chose 32 proteins of varying size and secondary structure composition. These proteins were selected from recent CASP competitions, from experiments performed by Bradley and Baker [9], and from the PDB. The list of proteins is shown in Table 3.1. Search was conducted with two move sets: one excludes fragments from proteins homologous to the prediction target, the other one includes these fragments. Both move sets contain 200 fragments at each position, however, the homology move set contains fragments more structurally similar to the native structure of the prediction target.

Homology information in the fragment library simplifies the search problem because it introduces a structural bias towards homologous structures. Search using a homology move set thus leads to lower energy samples and more accurate structure predictions. Irrespective of the move set, MBS search outperforms MC search (see Figure 3.4).



(a) Search with homolog move set (74aa, PDB: 1dtj)

(b) Search with homolog-free move set (74aa, PDB: 1dtj)

Figure 3.4. Homology information in the move set greatly improves the effectiveness of search. With and without homology information, MBS finds lower-energy samples and more accurate predictions than MC.

To obtain native-like structures for comparison, I run 100 relaxations on the all-atom structures found in the PDB. I determine structural similarity between the native state and predictions using the total score of the global distance test (GDT_TS) [75]; the score is reported in percent with 100% being a complete match

between structures. When RMSD is reported, it refers to all-atom RMSD calculated by PyMOL given in Ångstrom. To assess the energy of points in conformation space, I use the unit-less number returned by Rosetta's all-atom energy function.

For the discussion of my results I divide the proteins into four categories. Category 1 contains proteins for which both MBS and MC make accurate predictions. The second category encompasses proteins for which MBS found lower-energy structures and made more accurate structure predictions. For proteins in category 3, MBS found structures that were lower in energy than the native state, pointing to inaccuracies in the energy function. Finally, category 4 contains proteins for which neither MBS nor MC can find structures comparable to the native state in terms of energy or structural similarity. In the remainder of this section, I discuss the findings for these four categories in detail.

Category 1: Adequate Conformation Space Search

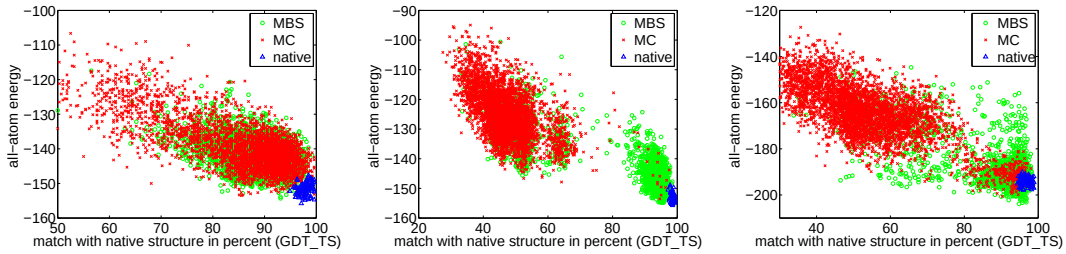
For the seven proteins in category 1 (see Table 3.1), model-based search (MBS) and Monte Carlo (MC) perform equivalently. Both find structures with an RMSD of less than 1.5Å from the native state. Proteins in this category are relatively small (less than 116 amino acids). It is plausible that the conformation spaces for these proteins are relatively easy to search. Obviously, if MC search finds the global minimum of the energy landscape, MBS cannot improve the result.

Figure 3.5 shows samples generated by MC and by MBS for three representative proteins from category 1. The scatter plots indicate that both MC and MBS find conformations in the bottom right of the graph, where the structural match with native structures is very high and the energy is low.

It should be noted that for one protein (434 repressor, PDB: 1r69) MC finds a lower energy samples than MBS (see Table 3.1).The lower energy of the MC sample can be attributed to the stochastic nature of the search.

Protein Attributes						Energy of Best Sample				Structural Match with Native				
						Homolog Move Set		Homolog Free Move Set		Homolog Move Set		Homolog Free Move Set		
PDB	L	%α	% β	# homologs	E_{Native}	E_{MC}	E_{MBS}	E_{MC}	E_{MBS}	GDT_TS$_{\text{MC}}$	GDT_TS$_{\text{MBS}}$	GDT_TS$_{\text{MC}}$	GDT_TS$_{\text{MBS}}$	cat.
1b72	49	69	0	3	-119	-117 (-111)	-119(-112)	-114(-109)	-116 (-110)	92 (90)	95 (85)	67 (70)	57 (66)	1 (4)
1shf	59	5	41	9	-143	-131 (-116)	-139 (-130)	-113 (-107)	-116 (-110)	89 (73)	92 (90)	55 (49)	48 (55)	2 (4)
2reb	60	62	20	1	-142	-146 (-143)	-146 (-144)	-144 (-136)	-146 (-141)	96 (93)	94 (94)	95 (85)	88 (90)	1 (1)
1r69	61	64	0	2	-148	-146 (-141)	-145 (-142)	-143 (-137)	-147 (-140)	96 (91)	91 (91)	74 (77)	87 (83)	1 (2)
1csp	67	4	54	1	-151	-143 (-132)	-147 (-140)	-138 (-126)	-137 (-128)	69 (56)	91 (75)	53 (50)	69 (53)	2 (4)
1di2	69	46	33	1	-156	-154 (-151)	-154 (-150)	-149 (-142)	-151 (-146)	86 (92)	93 (91)	64 (65)	66 (64)	1 (4)
1n0u	69	43	25	1	-155	-141 (-132)	-139 (-135)	-138 (-132)	-138 (-133)	64 (50)	47 (54)	68 (45)	41 (44)	4 (4)
1mla	70	34	37	1	-156	-154 (-142)	-154 (-150)	-143 (-137)	-141 (-137)	95 (60)	95 (95)	44 (47)	49 (47)	1 (4)
1af7	72	72	0	1	-171	-167 (-161)	-173 (-167)	-171 (-161)	-169 (-163)	38 (51)	38 (38)	38 (50)	39 (41)	3 (3)
1dcj	73	32	27	0	-163	-140 (-130)	-154 (-141)	-142 (-134)	-149 (-141)	64 (59)	71 (64)	67 (67)	68 (68)	2 (4)
1dtj	74	39	27	1	-162	-161 (-150)	-169 (-163)	-151 (-143)	-165 (-160)	95 (81)	92 (92)	64 (54)	86 (84)	2 (2)
1o2f	77	39	27	0	-177	-168 (-153)	-166 (-159)	-161 (-153)	-164 (-156)	40 (39)	40 (41)	43 (40)	39 (41)	4 (4)
1mky	81	32	25	0	-182	-171 (-153)	-168 (-162)	-166 (-154)	-167 (-159)	68 (45)	53 (54)	46 (45)	50 (49)	4 (4)
2hfq	83	30	29	0	-187	-179 (-168)	-185 (-178)	-176 (-166)	-182 (-174)	49 (40)	45 (41)	39 (40)	48 (47)	4 (4)
1tig	88	35	35	1	-198	-200 (-196)	-204 (-199)	-191 (-178)	-195 (-189)	95 (92)	95 (94)	54 (52)	52 (55)	1 (4)
1hhp	99	7	48	3	-216	-196 (-179)	-210 (-203)	-188 (-177)	-189 (-180)	80 (36)	84 (83)	20 (19)	18 (19)	2 (4)
2hg6	103	35	21	0	-211	-208 (-198)	-222 (-213)	-210 (-200)	-211 (-203)	23 (22)	21 (22)	20 (22)	21 (22)	3 (3)
1pva	109	57	0	12	-260	-246 (-237)	-262 (-257)	-246 (-234)	-247 (-237)	54 (39)	92 (88)	33 (29)	30 (30)	2 (4)
1elw	116	79	0	7	-302	-297 (-292)	-297 (-293)	-291 (-281)	-290 (-283)	88 (92)	97 (95)	57 (62)	55 (62)	1 (4)
1bm9	120	54	12	1	-260	-284 (-272)	-290 (-272)	-282 (-270)	-279 (-270)	27 (27)	28 (27)	29 (27)	20 (24)	3 (3)
2h5n	123	70	0	0	-291	-281 (-268)	-288 (-280)	-277 (-269)	-283 (-276)	26 (27)	33 (32)	28 (27)	28 (30)	4 (4)
1jb2	123	29	49	1	-275	-246 (-234)	-265 (-255)	-236 (-218)	-247 (-234)	82 (46)	79 (71)	31 (26)	46 (41)	2 (4)
8rat	124	21	33	6	-238	-218 (-201)	-227 (-219)	-211 (-194)	-223 (-210)	20 (24)	24 (23)	18 (19)	19 (19)	4 (4)
2j6a	135	33	16	1	-283	-257 (-243)	-261 (-254)	-258 (-241)	-257 (-246)	23 (20)	21 (19)	24 (20)	18 (21)	4 (4)
1cbr	136	13	57	7	-305	-276 (-266)	-290 (-280)	-277 (-263)	-281 (-271)	31 (36)	72 (56)	30 (42)	31 (31)	2 (4)
1aly	139	2	51	3	-223	-241 (-227)	-263 (-245)	-247 (-227)	-259 (-246)	17 (17)	18 (19)	12 (17)	18 (20)	3 (3)
1h3q	140	34	25	0	-311	-291 (-279)	-306 (-297)	-292 (-276)	-303 (-291)	27 (25)	26 (27)	29 (24)	30 (29)	4 (4)
1oo0	144	34	34	0	-340	-287 (-271)	-303 (-293)	-285 (-267)	-292 (-281)	24 (22)	18 (24)	21 (22)	29 (26)	4 (4)
1kd6	166	8	39	0	-290	-310 (-295)	-334 (-324)	-308 (-293)	-321 (-308)	15 (16)	16 (17)	15 (16)	18 (17)	3 (3)
1ad6	180	70	0	1	-403	-387 (-376)	-396 (-386)	-387 (-375)	-388 (-380)	33 (26)	34 (29)	24 (25)	24 (24)	4 (4)
1qdl	195	25	35	3	-464	-379 (-354)	-409 (-396)	-366 (-345)	-380 (-362)	50 (25)	48 (47)	18 (17)	23 (23)	2 (4)
3cla	213	30	29	1	-390	-395 (-374)	-433 (-412)	-399 (-376)	-412 (-396)	14 (17)	20 (19)	18 (17)	21 (19)	3 (3)
( ) = Top 100 structures														

Table 3.1. The 32 proteins used in our experiments ordered by increasing length; columns contain the PDB code, protein length in amino acids, percentage of α-helix and β-sheet calculated by DSSP [27], and the number of homologs in the homolog move set. The next five columns list the energy of the lowest sample obtained for the native state (E_{Native}), by using MC search (E_{MC}), and by using MBS (E_{MBS}) for both move sets, in parentheses is the average of the 100 lowest all-atom energy structures. The next four columns indicate the structural match between the lowest energy structure predictions by MC/MBS and the native state, measured as a percentage using GDT_TS [75]. The final column contains the category of the protein, referred to in our discussion. The protein category corresponding to the homolog free move set is given in parentheses.



(a) RNA binding protein A (69aa, PDB: 1di2)

(b) Malonyl-CoA ACP transacylase (70aa, PDB: 1mla)

(c) IF3-C (88aa, PDB: 1tig)

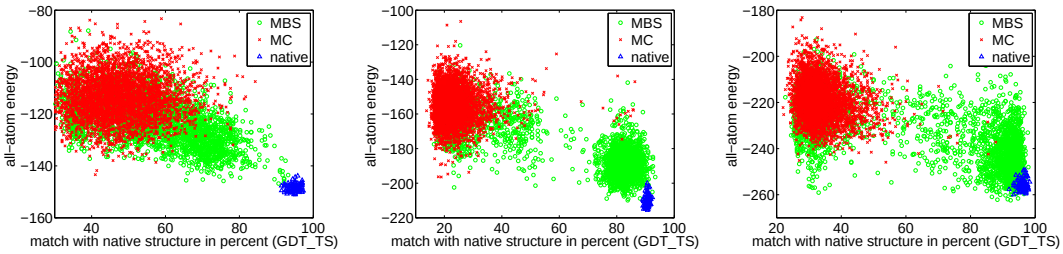
Figure 3.5. For proteins in category 1 both MC and MBS adequately search conformation space, resulting in near-native structure predictions. Each point in the scatter plots represents a conformation space sample. MC samples and native structures samples are drawn on top of MBS samples. These results were obtained using the homolog move set.

Category 2: Improved Conformation Space Search

Category 2 consists of proteins for which MBS searches conformation space more effectively than MC. When the homolog move set was used, nine of the 32 proteins fell into this category. Using the homolog-free move set, only two proteins fell into this category. For all proteins in this category, MBS finds lower-energy samples than MC; these samples correspond to higher-accuracy structure predictions. These proteins range in size between 59 and 195 amino acids.

The improvement of MBS over MC is illustrated in the scatter plots in Figure 3.6. Samples generated by MBS are lower in energy and in many cases overlap the energy of the relaxed native structure.

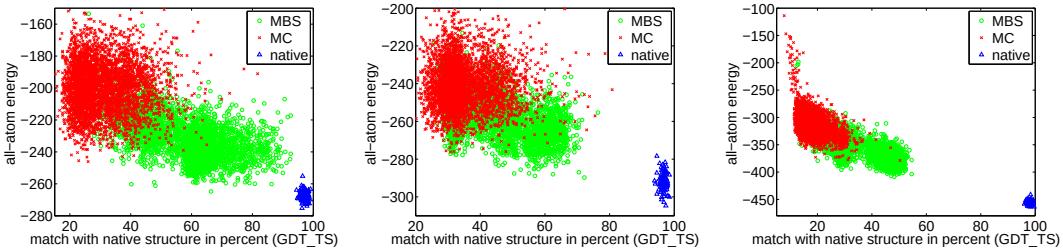
The lower-energy predictions generated by MBS result in more accurate structures. This is illustrated for two proteins using the homolog move set in Figure 3.7 and two proteins using the homolog-free move set in Figure 3.8. Note that the structure prediction shown in Figure 3.7(c) corresponds to the scatter plot shown in Figure 3.6(e). The lowest-energy samples found by MBS only achieve a GDT_TS of 72; nevertheless, with an RMSD of 2.7Å the prediction is quite accurate.



(a) Cold-shock protein (67aa, PDB: 1csp)

(b) Aspartyle protease (99aa, PDB: 1hhp)

(c) EF Hand parvalbumin (109aa, PDB: 1pva)



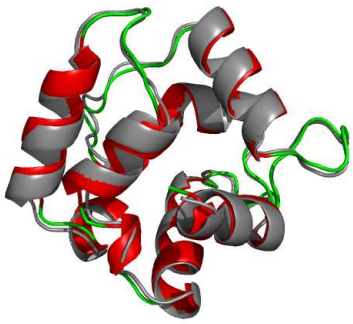
(d) NTF2 M84E mutant (123aa, PDB: 1jb2)

(e) Retinoic acid binding protein(136aa, PDB: 1cbr)

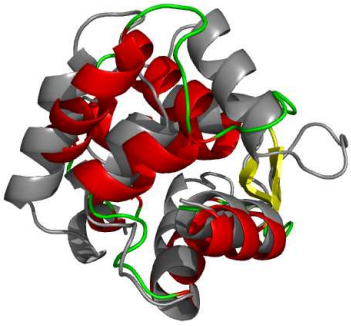
(f) Anthanilate synthase(195aa, PDB: 1qdl)

Figure 3.6. For the proteins of category 2, MBS outperforms MC. MBS finds lower-energy samples and these samples match the structure of the native protein more closely than the samples obtained by MC. These results were obtained using the homolog move set.

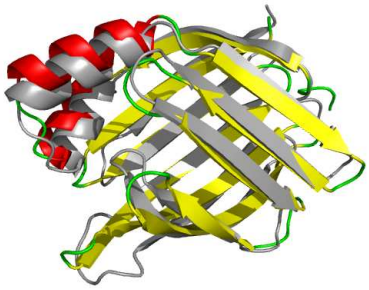
The only protein in category 2 that is not accurately predicted is Anthanilate Synthase (PDB: 1qdl). With a length of 195 amino acids, this is the second-largest protein in my test set. The scatter plot in Figure 3.6(f) shows that the samples generated by MBS seem to lie on a trajectory towards the native state but get stuck before reaching it. These results indicate either that conformation space search still remains inadequate for proteins of this length or that the energy function is inaccurate.



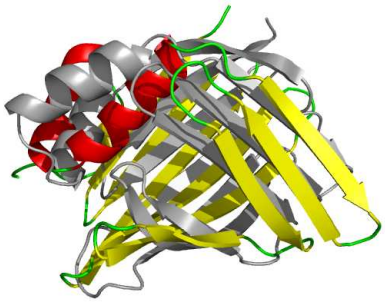
(a) MBS: 1.1Å RMSD, 92.4 GDT_TS



(b) MC: 5.3Å RMSD, 53.8 GDT_TS

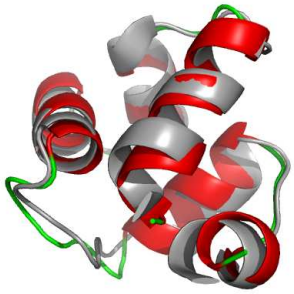


(c) MBS:2.7Å RMSD, 72.2 GDT_TS

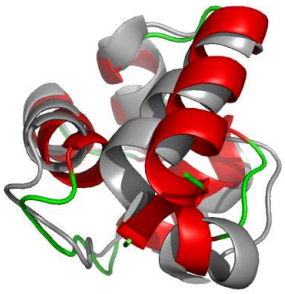


(d) MC: 6.7Å RMSD, 30.8 GDT_TS

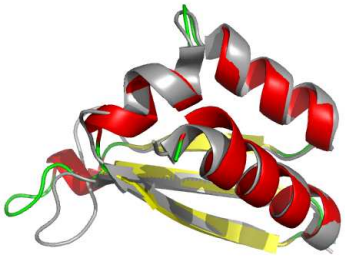
Figure 3.7. Predicted structures for proteins from category 2 using the homolog move set (color), superimposed on native structures from the PDB (gray): EF Hand parvalbumin (109aa, PDB: 1pva) predicted with MBS (a) and with MC (b); retinoic acid binding protein (136aa, PDB: 1cbr) predicted with MBS (c) and with MC (d).



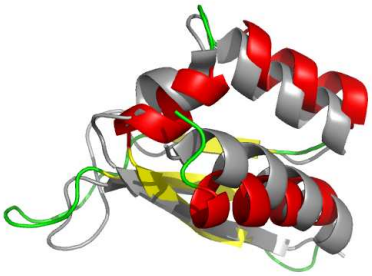
(a) MBS: 1.5Å RMSD, 86.8 GDT_TS



(b) MC: 2.3Å RMSD, 74.1 GDT_TS



(c) MBS:1.4Å RMSD, 86.4 GDT_TS

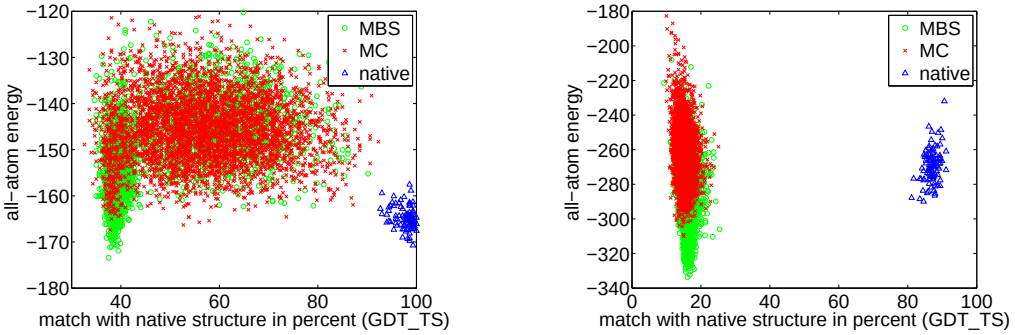


(d) MC: 3.6Å RMSD, 64.1 GDT_TS

Figure 3.8. Predicted structures for proteins from category 2 using the homolog move set (color), superimposed on native structures from the PDB (gray): 434 Repressor (61aa, PDB: 1r69) predicted with MBS (a) and with MC (b); KH domain of Nova-2 (74aa, PDB: 1dtj) predicted with MBS (c) and with MC (d).

Category 3: Inaccurate Energy Function

For six proteins model-based search finds conformations with lower all-atom energy than that of the native state, regardless of the move set. The scatter plots shown in Figure 3.9 illustrate this for two of the six proteins. In the case of Cher domain 1 (Figure 3.9(a)), both MC and MBS find samples with a high GDT_TS. However, samples in the highest-density region exhibit little structural similarity to the native state. These samples have lower energy than the native protein. Figure 3.9(b) illustrates this phenomenon even more strikingly: both search methods and move sets find conformations considerably lower in energy than the native structures.



(a) Cher domain 1 (72aa, PDB: 1af7)

(b) Pore-forming cytolysin equinatoxin (166aa, PDB: 1kd6)

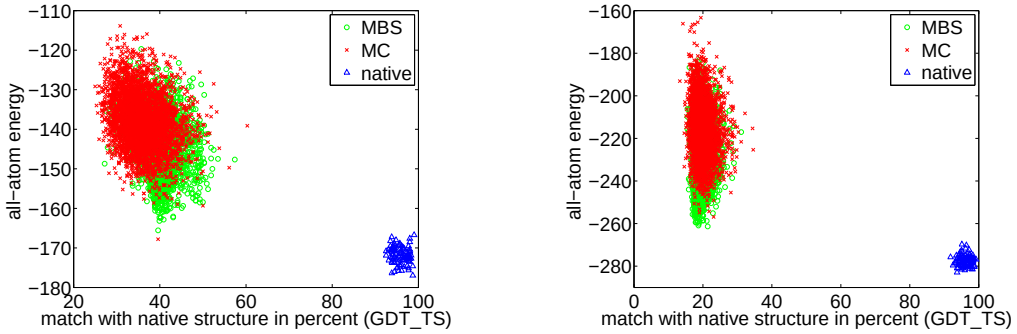
Figure 3.9. For proteins in category 3 MBS finds conformations with lower energy than the native state, pointing to inaccuracies in the energy function. These results were obtained using the homolog move set.

These results obtained for proteins in category 3 show that MBS searches conformation space more effectively than MC. However, for proteins in this category, the reduced energy of samples does not result in accurate predictions. This is a consequence of inaccuracies in the energy function. An inaccurate energy function guides search towards wrong regions of conformation space. No matter how much search is improved, it will not be able to compensate for these inaccuracies.

Model-based search may serve as a tool to improve energy functions. Once inaccuracies are identified, using the results of accurate conformation space search, it may be possible to identify and correct inaccurate components of the energy function.

Category 4: Inadequate Conformation Space Search

This last category of proteins is the most interesting one. For all proteins in this category neither MC nor MBS adequately searches the conformation space. Using the homolog move set, ten proteins of varying sizes (from 69 to 180 amino acids) fall into this category. Using the homology-free move set, 24 of the 32 proteins are in this category. This large number indicates that search becomes very difficult when the information contained in homologous fragments is not available to the search.



(a) Glucose permease IIBC (77aa, PDB: 1o2f)

(b) Zinc Finger Protein (135aa, PDB: 2j6a)

Figure 3.10. For proteins in category 4 neither MBS nor MC search conformation space adequately. These results were obtained using the homolog move set.

Relative to the results for proteins in category 2, the structural match with the native state is very poor. A comparison of the scatter plots in Figures 3.6(f) and 3.9 also reveals a qualitatively different behavior of search between categories 2 and 4, i.e. between successful search and unsuccessful search. The samples generated for category 4 proteins do not form a trajectory towards the native state whereas those

for category 2 do. This difference is particularly apparent in Figure 3.10(b), where search seems to be unable to access large regions of the conformation space.

Category 4 seems to point to a fundamental problem with model-based search. If none of the samples in the initial model of MBS are close to native state, the search conducted by MBS will focus on regions not containing the native state of the protein. Overcoming this problem seems to be the key to further improvements of conformational space search based on MBS.

A previous study [9] also achieved poor prediction quality for three proteins from category 4, even though an order of magnitude more all-atom samples were used. This indicates that a mere intensification of sampling does not lead to a discovery of the conformation space region containing the native state. Also, since the size of category 4 proteins varies significantly, I believe that the size of the conformation space is not the main source of this problem either.

I have two hypotheses that may explain the problem encountered by MBS. The first hypothesis states that proteins in category 4 have energy landscapes with a narrow funnel leading to the native state. The second hypothesis states that inaccurate intermediate energy functions may steer search away from the region containing the native structure.

The narrow funnel hypothesis explains why category 4 contains short as well as longer proteins. Already for small proteins, the conformation space is too large for search to accidentally discover a small region that represents the entrance to the funnel, unless the energy landscape contains large regions that slope towards it.

The narrow funnel hypothesis emphasizes the importance of understanding residual native structure present in the denatured states of proteins. Biological proteins exhibit residual structure as a consequence of interactions among side-chains in close proximity along the backbone. In contrast, MBS has to discover this structure by random assembly of fragments, a proposition of vanishingly small probability. This

probability is reduced even further when homologs are excluded from the move set, explaining the large increase in the number of category 4 proteins when homologs are removed from the move set.

It should be noted that for such narrow funnels, MC-based search with random restarts may in some cases have a higher probability of discovering the entrance to the funnel. I observe this in three of the ten proteins in category 4 for experiments using the homolog move set (elongation factor 2, glucose permease IIBC, and enga protein).

A second hypothesis is also consistent with my observations. To find the entrance to the folding funnel in the all-atom energy function, the energy function of stage i must lead samples into the correct funnel of the energy function at stage $i + 1$. This may not hold for proteins in category 4: assume that search at stage i, MBS identifies the correct minimum of the energy function. If local search in the energy function at stage $i + 1$ does not lead to the global minimum when started from the minimum of stage i, search will be guided away from the native structure and is unlikely to recover from it, no matter whether MC or MBS is used as the search strategy. Therefore, my second hypothesis states that for category 4 proteins the global minima in consecutive energy functions are shifted, preventing search from identifying the correct folding funnel.

This second hypothesis, if true, may be an indication that conformation space search is no longer the most pressing problem in protein structure prediction. It may be equally important to leverage the capabilities provided by MBS to further improve the accuracy of the approximate energy functions.

3.4 Results of model-based search on CASP 8

I validated model-based search(MBS) by taking part in the Critical Assessment of Techniques for Protein Structure Prediction experiment. To enable my participation I

Protein Category	Grishin Lab Analysis (servers)	Baker Lab Analysis (servers)
FM or FR_A-NF (corrected)	**6/74**	**14/69**
FM or FR_A-NF	6/74	28/69
FR (corrected)	42/74	57/70
FR	46/74	57/70
CM-hard	64/74	67/72
CM-medium	69/74	69/73
CM-easy	69/74	69/72
All-proteins	69/74	N/A

Table 3.2. My performance in CASP 8 based on two different evaluations, detailed by prediction categories. Categories are listed from hard (top) to easy (bottom): FM = free modeling, FR_A-NF = fold recognition new fold, FR = fold recognition, CM = comparative modeling. The most relevant category for MBS is FM or FR_A-NF; these are the most difficult predictions for which no homology information was available. Only when no homology information was available could my method be fairly compared to other, since my CASP entry did not have the capability to incorporate homology information. During CASP, I experienced a power outage and did not submit predictions for one of the target. The lines with the annotation "corrected" show the results of my server had that missing protein been submitted. The rankings of the Baker and Grishin laboratories differ because they classified proteins differently.

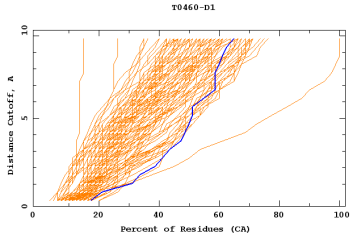
developed a protein structure prediction server. My server receives e.mails containing an amino acid sequence, initiate my MBS-based prediction algorithms, select the best decoys and return them. However, unlike most other methods taking part in CASP my entry did not have the ability to retrieve homology information. This was a major handicap, as homology information, if available, renders conformational space search substantially easier. Due to this limitation of my first-ever entry into CASP, I can only compare my results for free modeling targets (FM) in a meaningful way.

For free modeling targets, my server was ranked **6th out of 74** in one evaluation scheme and **14th out of 69** in a different evaluation scheme. Rankings for all prediction categories are shown in Table 3.2. My goal was to have a single top-10 prediction. I have exceeded this goal by far, having developed a **top-10 prediction server for free modeling targets**. In addition this work received the **best poster award**.

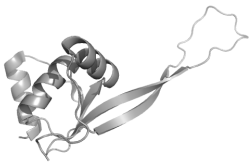
The eighth Critical Assessment of Techniques for Protein Structure Prediction (CASP8) was held in the summer of 2008. CASP8 required the prediction of 128 pro-

teins, which for analysis purposes are split into domains. There were 4 free modeling domains, 26 fold recognition domains, 36 difficult comparative modeling domains, 73 medium difficulty comparative modeling domains and 26 easy comparative modeling domains. Free modeling domains correspond to unique folds which have not been previously seen. Fold recognition targets are proteins for which a fold exists in the SCOP or CATH structural classification of protein databases, but the sequence is not close enough to identify homologs [45, 48]. Comparative modeling targets are proteins where a homolog can be identified from sequence.

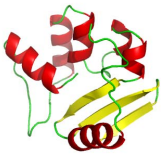
I would like to highlight five of the most accurate prediction results obtained with MBS. Two proteins were in the free modeling class (T460, T465, see Figure 3.11), two proteins in the fold recognition class (T478_1, T482, see Figure3.12), and—quite surprisingly—one protein in the easy comparative modeling class (T499). For this protein, I had the most accurate prediction, even when compared to homology modeling approaches (see Figure 3.13).



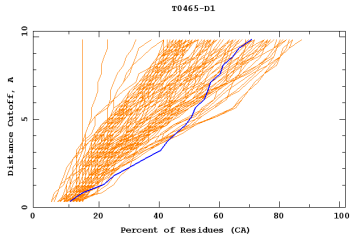
(a) T460-Hubbard



(b) T460-native

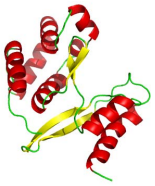


(c) T460-prediction



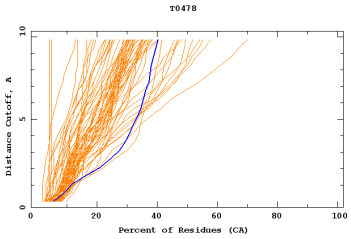
(d) T465-Hubbard

(e) T465-native

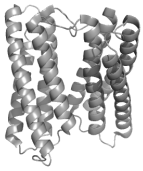


(f) T465-prediction

Figure 3.11. The Hubbard plots are of the CASP predictions by model-based search (MBS) (solid) and all other groups (dashed). Small slopes correspond to better predictions. In gray are the native structures of the proteins. In color are the predictions. The proteins in this group are free modeling proteins.

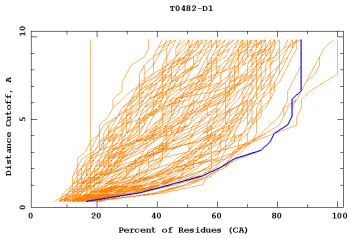


(a) T478-Hubbard



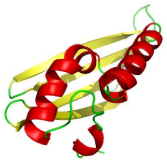
(b) T478-native

(c) T478-prediction



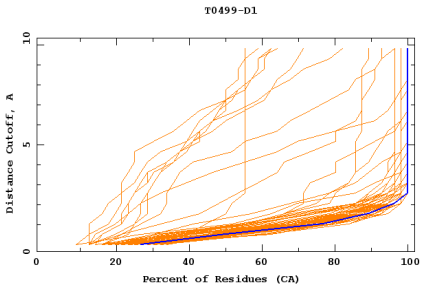
(d) T482-Hubbard

(e) T482-native

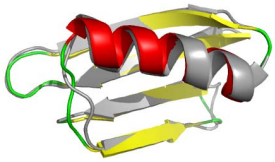


(f) T482-prediction

Figure 3.12. The Hubbard plots are of the CASP predictions by model-based search (MBS) (solid) and all other groups (dashed). Small slopes correspond to better predictions. In Gray are the native structures of the proteins. In color are the predictions. The proteins in this group are fold recognition targets.



(a) T499-Hubbard



(b) T499-predicted

Figure 3.13. For this easy template-based modeling protein, MBS obtained the most accurate prediction, outperforming all other servers that relied on homology information that was highly accurate for this protein.

CHAPTER 4

GUIDING CONFORMATIONAL SPACE SEARCH WITH STRUCTURAL INFORMATION FROM THE PROTEIN DATA BANK

In most search domains there exists a vast amount of information capable of directing search toward the solution, and this information is quickly growing. With proteins for instance, there are currently 8 million non-redundant protein sequences and that number is doubling every 28 months [33]. Similarly, the number of protein structures has doubled between 2000 and 2010 and is expected to triple before 2014. Since information is critical to search performance, more effective use of information from amino acid sequences and protein structures would significantly improve protein structure prediction.

When applicable, known protein structure is the most significant source of information for protein folding. This information is so promising that evidence suggests, if fully utilized it would come close to solving protein structure prediction. For example, no new protein motifs have occurred during the last several years of the Computational Assessment of Structure Prediction experiment (CASP) [15], and Zhang has shown that 99.8% of single-fold proteins have a corresponding homolog within $6\mathring{A}$ RMSD, and 97% have a homolog under $4\mathring{A}$ RMSD [80].

Current methods identify relevant template proteins by matching amino acid sequences, and predicted secondary structure. These methods fail to identify relevant templates when sequence similarity is low. However, structural similarity is conserved 3-10 times more than sequence similarity [40, 22], so if structural similarity could be used to identify homologs, more templates would be found. To identify proteins with

structural similarity I propose to search the protein data bank(PDB) with structural features from decoys produced in an initial round of structure prediction. Many of the structural features found within these will be incorrect, however, within a collection of decoys, some features from decoys have been correctly predicted [4].

Using decoy substructure to identify template proteins will incorrectly identify many template proteins. To deal with the incorrectly identified templates I have developed a new strategy, that seeks to use template proteins only to the extent they are useful in search. When the template protein is high quality, exploiting it will likely direct search toward the native structure. Exploitation of an incorrect template protein would likely lead search in a region that does not contain the native state. Therefore, when information is likely incorrect, exploration needs to be increased to decrease the risk of search becoming trapped. Adaptively balancing exploitation with exploration in response to information quality allows the template protein to be used to the extent the information appears accurate. This is the first time exploitation and exploration have been adaptively balanced during the search process.

In this chapter, I present my new method to locate and use information from template proteins. The improvements afforded by my approach are based on two main contributions. First, my method can identify and assess the quality of template proteins. Second, enabled by the first contribution, my method adaptively balances exploitation with exploration, allowing the template protein to be used only to the extent current information suggests the template protein is useful.

Experimental evidence demonstrates that my approach identifies the correct homolog from a vast number of incorrect homologs, resulting in accurate protein structure predictions. I remind the reader that existing homology modeling methods identify the homolog and alignment from sequence information while this method discovers the homolog and alignment using only the search process.

4.1 Adaptive balancing of exploration with exploitation

In protein structure prediction it is computationally intractable to try all possible template proteins and alignments. Therefore, structure prediction methods must carefully choose which template proteins to use. However, template identification is imperfect.

The only way to identify an incorrect template is to find a different template that generates a lower energy structure. However, most templates will produce structures with higher energy resulting in wasted resources. I introduce adaptive balancing of exploitation with exploitation as the key to controlling how the template protein is selected. Due to the central role of balancing exploitation with exploration I refer to this method as Balanced Exploitation Exploration Template Search or BEETS. BEETS incrementally refines an estimate of how useful the template protein is to search, and based on this assessment changes the balance between exploitation and exploration.

Template proteins vary widely in their usefulness to search. I will refer to template protein usefulness to search as quality of the template protein. High quality templates are less likely to cause search to make a mistake and therefore search behavior can be very exploitative. Lower quality template proteins are more likely to cause search to make a mistake. To compensate for the increased risk incurred when using lower quality templates, search behavior must become more exploratory.

In this section, whenever possible I describe adaptive balancing of exploitation with exploration as a general search procedure. Adaptive balancing will likely be applicable to other search domains where there exists a significant number of information sources, and these sources vary widely in quality. Balanced exploitation and exploration is driven by four algorithmic elements described below.

1. **Acquisition of relevant information** Information capable of directing search must first be gathered and this information will be unique to each domain. For

protein structure prediction I use information from template proteins. Section 4.1.1 and Figure 4.1 describe how template proteins are acquired.

2. **Assessing the quality of information** At the core of balancing exploitation with exploration is an assessment of information quality. My method evaluates information quality with three methods: structural match between a target and template protein, change in RMSD, and change in energy. Section 4.1.2 and Figure 4.3 describes these metrics in more detail.

3. **Adaptive balancing of exploitation with exploration** Once relevant information has been gathered, and information quality assessed, search selects an information source and decides how strongly to exploit the information in that source. For protein structure prediction a template protein is chosen and the balance between exploitation and exploration is set by choosing the number of residues aligned between the target and template proteins. This is described in section 4.1.3.

4. **Integration with model-based search** Adaptive balancing of exploitation with exploration (BEETS) controls how a single sample explores conformation space. To explore multiple regions, BEETS is integrated with model-based search (MBS) from chapter 3. Section 4.1.4 and Figure 4.4 describes the integration.

The following four sections provide descriptions of the algorithmic elements; Section 4.2 augments the description provided below with a brief list of implementation details.

4.1.1 Acquisition of relevant information

Since information is key to choosing where to explore, it is critical to utilize the most useful domain specific information. For proteins the most useful information

comes from known protein structures stored in the PDB. This information exists because multiple proteins are likely to have similar structure due to evolution. Current methods identify relevant template proteins by matching amino acid sequences and predicted secondary structure. This is a very effective when two proteins are closely related but fails when they are more distantly related. As proteins evolve mutations occur that change amino acids. Over time the amino acids become increasingly different, but the structure remains similar because evolutionary fitness puts pressure on how proteins function, and function is caused by structure. Research has shown that structure is 3-10 times more conserved than sequence [22].

As a result of structure being more conserved than sequence, searching the PDB with structure should identify significantly more information. Searching the PDB with structure is difficult because the only structures available are from an initial round of structure prediction (decoys), and unfortunately, most of the decoys are incorrect. However, research indicates that within a collection of decoys some substructures from the decoys have been correctly predicted [4]. Instead of trying to find the correct substructure, my method uses all of the decoy substructures to search the PDB resulting in both correct and incorrect templates. Additional template proteins are acquired by searching the PDB with full length sequence and predicted secondary structure using the method hh_search [66]. The implementation details for are described in section 4.2.1 and illustrated in Figure 4.1.

Two collections of template proteins are acquired. One with homologs, one without. Homologs are identified as any protein that has a hh_search probability of $>$ than 20%.

4.1.2 Assessing the quality of information

Most template proteins will be irrelevant and should be ignored. Therefore, accurate template quality assessment is essential. In this section I describe three methods

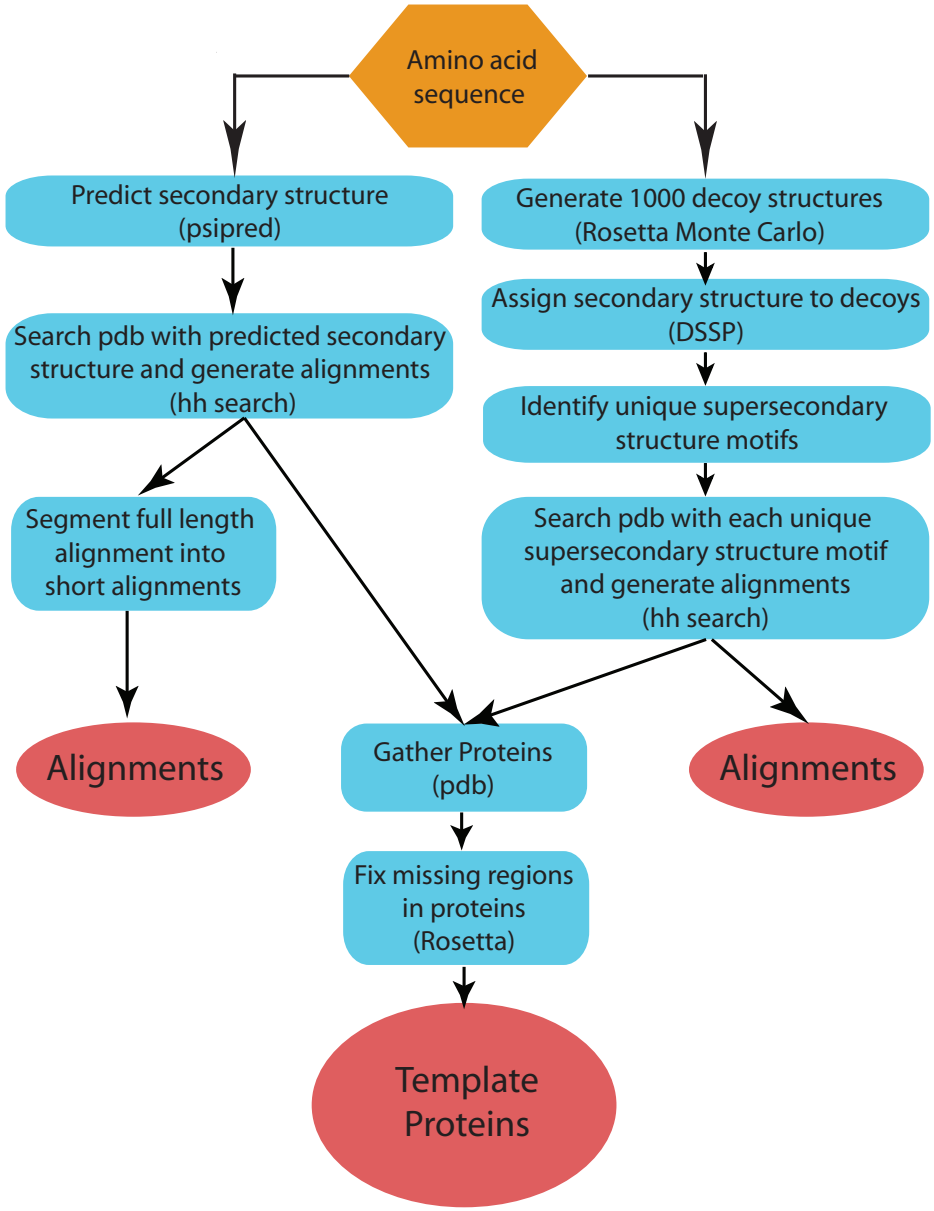


Figure 4.1. Step-by-step illustration of how template proteins are gathered and prepared. The left path corresponds to full length homologs identified with sequence and predicted secondary structure. The right path corresponds to identification of super-secondary motifs based on structural features found in decoys predicted during the search process. I gather two template collections: One with homologs and one without. I distinguish homologs as having a probability $> 20\%$ of being a homolog according to hh_search.

to evaluate template quality: the size of the match between template and target proteins, the effect exploiting the template protein has on structure, and the effect exploiting the template protein has on energy. These information sources are illustrated in Figure 4.3. It should be noted that the effect exploiting an information source has on structure (convergence) and energy (score) are general properties that would be applicable to domains in addition to protein structure prediction.

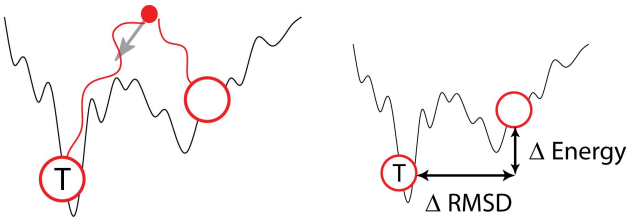
1. **Match size** My hypothesis states that the presence of a matching substructure in two proteins makes it more likely that residues neighboring the match also have similar structures. The longer the match, the more likely the two proteins share a very distant evolutionary relationship. If this is true then all residues near the match are more likely to have also share a distant evolutionary relationship and hence similar structure. Based on this hypothesis I can infer template quality based on match length.

   To identify match size I find the number of residues that fall under a RMSD cutoff. These residues can be discontinuous. The algorithm to compute this match is described in Section 4.2.2.

2. **Energy and structure change** Each template protein guides exploration into a different region of conformation space. Therefore, templates can be evaluated based on where they lead search. When exploiting a template protein produces lower energy structures, the template is believed to be higher quality. When exploiting a template protein produces a small change in the number of matched residues, search has converged. When search has converged no additional information is available. See Figure 4.2.

4.1.3 Adaptive balancing exploitation with exploration

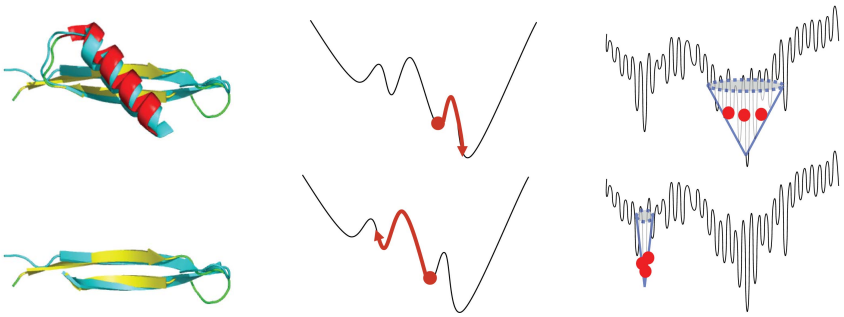
When the template protein has been assessed as high quality, search should be more exploitative of that template. When template proteins are assessed as low qual-



(a) Each source of information (template protein) guides exploration into a different region of conformation space. This image shows that when a template protein(grey arrow) is used, search is guided to the region labeled with a T (template). When no information is used, search leads to the unlabeled region.

(b) The quality of an information source (template protein) can be evaluated based on the difference between where exploration leads with and without that information. For proteins quality of information is assessed by measuring the change in RMSD and energy. Large RMSD and large decrease in energy correspond to high quality information. A large change in RMSD indicates search has not yet converged, and the decrease in energy suggests that the information from the template protein was valuable.

Figure 4.2. Evaluation of change in energy and RMSD



(a) Match size (b) Change in Energy (c) Change in RMSD

Figure 4.3. Shown in this image are the three sources of information used to evaluate template quality. The top images show information that favors exploitation of the template protein fragment. The bottom images show information favoring exploration. Image (a) shows the maximal match between a decoy and template protein. Larger matches increases the likelihood that parts of the template protein near the match contain useful information. Image (b) shows the effect exploiting a template protein has on the search process. When exploitation of a template protein results in lower energy solutions, the template protein is more likely accurate. Image (c) shows decoy protein structures represented as red dots clustered into two regions. When exploitation of a template protein results in structures that do not change RMSD significantly, search has converged. When search has converged, more information is only gathered by exploring a new region of conformation space.

ity, search should be more exploratory to continue looking for additional information. For template proteins, changing the balance between exploitation and exploration is accomplished by shifting the number of residues aligned between the target and template proteins. This is accomplished in two steps described below.

1. **Choice of which template** The primary information source used when choosing the template is the structural match between the decoy protein and the template protein. The longer the match, the more likely that the neighboring residues will be useful to guide search to relevant regions. I found that taking into consideration the number of secondary structures matched in addition to the number of residues improved search performance. The following example shows why. Long α helices are a basic building block of proteins and have likely evolved multiple times. Therefore, a long α helix will match many residues, however, this feature would have little benefit to identification of a structurally similar template. Utilizing secondary structure match in addition to residue match helps my method focus on features that would likely have evolved together.

2. **Adjusting the alignment between target and template proteins** When exploitation is favored, I insert the region that has been previously determined to match structurally and add an addition number of residues. It should be noted that this growth adds consecutive residues, but as the results will indicate, this may not be the best idea due to the increased number of insertions and deletions in loops. Future work will seek to improve how exploitation is increased. When exploration is chosen, either the entire match is removed or a portion thereof.

4.1.4 Integration with model-based search

Model-based search incrementally refines an initial coarse model of conformation space by incorporating new information obtained during an ongoing search. This information is used to allocate computational resources to regions based on their estimated relevance. As such, model-based search controls which regions are being investigated by search.

Balanced exploitation exploration template search (BEETS) incrementally refines a model of template protein quality by incorporating new information obtained during ongoing search. This assessment of template protein quality is used to choose how to allocate resources to template proteins. As such, BEETS controls how search explores an individual region.

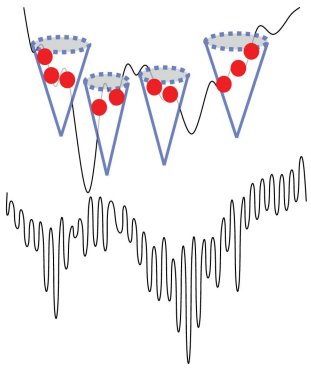
Figure 4.4 illustrates a single iteration of how model-based search and BEETS are integrated. Section 4.2.4 gives the implementation details.

4.2 Implementation

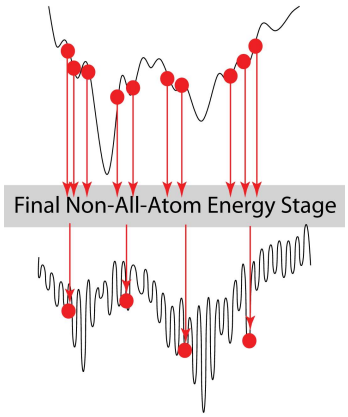
4.2.1 Acquisition of templates

The goal of template acquisition is to get all templates that may be useful to search, while discarding templates that have no possibility of being useful. The process of balancing exploitation with exploration will then select the templates that most effectively steer search toward low energy states. Templates are acquired by both the top current method, HH-search, and a new method that searches the PDB based on structural features found in decoys returned from an initial round of structure prediction.

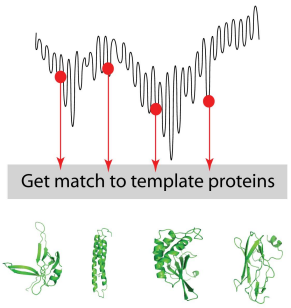
HH-search acquires templates by searching the PDB with sequence and predicted secondary structure using [66]. HH-search was the top performer in the CASP 9 TBM results. I collect all templates that have greater than 2% HH-search probability of being a match to the native structure.



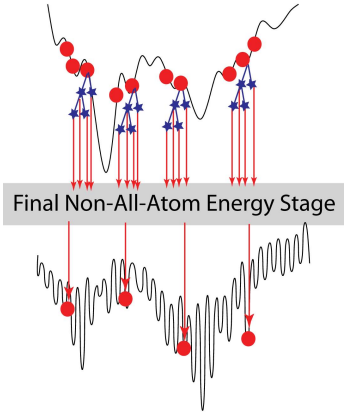
(a) Each step of model-based search (MBS) uses a model consisting of 2000 samples in the approximate energy function of Rosetta. MBS groups these samples into funnels that accurately capture coherent regions in the energy landscape. This expressive characterization of conformation space enables computationally effective assessment of region relevance.



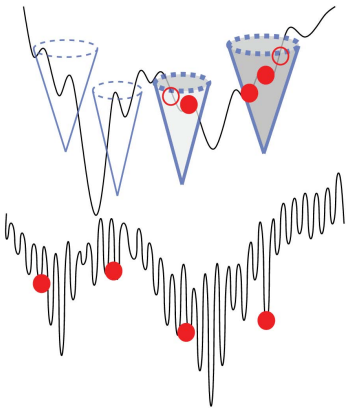
(b) To assess the relevance of a region as accurately as possible, information is acquired from the all-atom energy function. All samples are evaluated with Monte Carlo trajectories to Rosetta's final non-all-atom stage. For each funnel, the lowest-energy trajectory is continued through Rosetta's all-atom energy function.



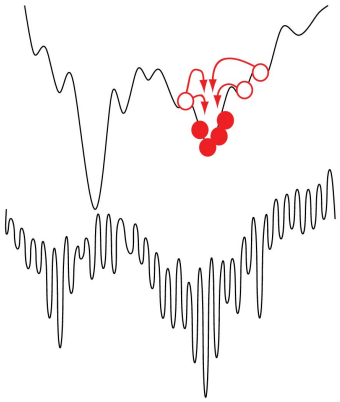
(c) There are a vast number of template proteins, and possible ways to align the template protein to the target protein. To assess the quality of a template protein my method determines the match between a decoy and each template protein. To conserve computational resources, structural match is only computed for the lowest all-atom energy decoy in each funnel.



(d) From the lowest energy samples in each region, BEETS explores the possible ways to use substructures of template proteins. Each star represents a change in the alignment between template protein(s) and target protein but no structural change to the target protein. There are 3 stage stages of template alignment change (BEETS) for every 1 stage of structure change (MBS).



(e) Based on funnel relevance assessed by all-atom information, the model is pruned by discarding entire funnels and most of the samples within the remaining funnels. Search retains the best half of the funnels and the top four samples within each of the remaining funnels. Note MBS discards funnels corresponding according to the all-atom energy score not the approximate energy function.



(f) MBS replaces discarded samples using short Monte Carlo trajectories starting at the samples remaining in the model. Eighty percent of the starting locations for these trajectories are chosen evenly across funnels; twenty percent of the starting locations are chosen proportional to the relevance of the funnel.

Figure 4.4. Step-by-step illustration of a single stage of integration between model-based search and balanced exploitation with exploration template search. Most images contain two energy landscapes. Rosetta's approximate energy function is shown on top, Rosetta's all-atom energy function on the bottom. Note that the minimum in the approximate energy landscape does not correspond to the minimum in the all-atom energy landscape.

Additional templates were gathered by searching the PDB with unique super-secondary motifs collected from an initial round of protein structure prediction. I define super-secondary motifs as having a α helix or β sheet - turn - α helix or β sheet morphology. Since the super-secondary motifs have a different structure than the predicted secondary structure, using motifs to search the PDB generates many additional templates. The top 30 templates for each super secondary structure motif are gathered if they have a HH-search probability of greater than 2%.

HH-search takes about 20 minutes to execute so it would be time prohibitive to run HH-search on all super-secondary motifs in a decoy set. Instead, only unique motifs are used to search the PDB. HH-search represents secondary structure information as a string stored in the DSSP format [27]. Therefore, I store motifs as DSSP strings. DSSP is an algorithm that converts the secondary structure of each residue to an 8 character string. There are 3 characters for helix and turn and 2 characters for sheet. To determine the uniqueness of a motif I measure the cost to converting one motif string into another. I define the cost function as follows. Converting a helix to a β-sheet is likely to represent different motifs so the cost is high (1). A loop is more likely to change into a helix or sheet later in search so the cost is lower (.5), and the conversion within types of helix, sheet or turn happens often so the cost is low (.1). If the characters are the same the cost is 0. Strings with a lower cost of conversion are clustered together using greedy agglomerative clustering [23]. This result is 100-300 motifs for 1000 decoys.

HH-search uses PDB 70 database that was built in January 2010. The PDB 70 database removes homologs structure that have greater than 70% sequence identify.

After the alignments are acquired Rosetta is used to fix missing regions in the template proteins. This is because many template proteins have missing regions. Once all missing regions have been fixed the structure is minimized in the all-atom energy function. Minimizing the energy of the template protein allows fragments

from the template protein to be used in search. For time consideration protein size is limited to 800 residues for fixing of loops and 1500 residues for the all-atom relax.

4.2.2 Assessing the quality of information

1. Structural match between decoy and template protein The goal is to find the longest match since longer matches indicate a more useful template. The match procedure starts at the middle residue of the alignment returned by hh search in the preceding step. The match is extended by one residue at a time in the direction that raises the RMSD the least. The growth stops when the RMSD is above 2.0. When the growth along the backbone is stopped the search continues to grow by adding the closest additional 3 residue matches between decoy and template. This repeats until every residue in the decoy has either been checked or eliminated. Adding the additional 3 residue matches allows the match to grow across loops. This is important because the most structural change happens in loops. For time purpose the match is only calculated for the lowest energy sample in each funnel because all structures within a funnel have the similar structure.

2&3. Energy change and RMSD change. Change in energy and change in RMSD is determined by comparing the different structure and energy produced by Monte Carlo trajectories produced with and without the the bias from the template protein. These short Monte Carlo trajectories conclude in Rosetta's final non-all-atom stage.

4.2.3 Adaptive balancing exploitation with exploration

For template proteins, changing the balance between exploitation and exploration occurs by changing the number of residues aligned, between the target and template proteins. The more residues aligned the more information is being used from the

template and the more exploitative the system is. The change in template alignments is made in three stages. The implementation details of these 3 stage are described below.

1. **Choose to increase exploitation or exploration** The change in RMSD and energy caused with the last alignment change, is compared to the highest and lowest changes to RMSD in energy. If equal to the highest change in RMSD and largest decrease in energy, exploitation is 100% likely because the template is likely useful. If equal to the smallest change in RMSD and largest increase in energy search is 50% likely to be exploitative and 50% likely to be more exploratory. Between these two values the likelihood of being exploitative is linearly interpolated. It should be noted that many samples are discarded during the MBS phase of search. These would have been the samples that would have strongly favored exploration.

2. **Choice of which template** The template is randomly chosen based on the structural match. Each template is scored between 0 and 1 then the template is randomly selected based on the score. 85% of the score comes from the number of template residues aligned. This biases search to use a template that is already aligned. 10% of the score comes from the number of secondary structure elements matched between template and target proteins and 5% of the score is given to the number of residues matched. The score from template matching biases search to use templates more similar to the structure currently in the decoy.

3. **Adjusting the alignment between target and template protein** The balance between exploitation and exploration is reflected in the number of residues aligned between target and template protein.

- If exploitation is chosen residues can be added to an existing alignment or an additional alignment to a template protein can be added. 70% percent of the time adding residues to an alignment is chosen and 30% percent of the time a new alignment is chosen. It should be noted that these alignments are a result of the match procedure, not the alignments identified using sequence information. I believe search would improved if sequence alignment information were used, but I wanted to prove that BEETS was able to build an alignment using only structural information.
- If exploration is chosen than the template can either be removed or the number of residues in an existing alignment can be reduced. 70% percent of the time the alignment size shrinks and 30% percent of the time the alignment is removed.

4.2.4 Integration with model-based search

Three stages of template substructure search (BEETS) occur for each stage of model-based search (MBS). For implementation details of model-based search the reader should see section 3.2.1.

4.3 Results

In this section, I compare the effectiveness of balanced exploitation exploration template search (BEETS) with model-based search (MBS) and simulated annealing Monte-Carlo (MC) implemented in Rosetta [6, 59]. To establish a fair test all three methods rely on the same parameters whenever possible. The only differences are described below.

All of the methods go through a series of stages (see Section 4.2.4); in each stage they use the same move sets, number of local search steps, and energy function. In MC all samples traverse all stages and these traversals proceed independently of

each other. MBS and BEETS however, orchestrate these trajectories, stopping some and splitting others into multiple trajectories. MBS uses the split off trajectories to evaluate region relevance (see Section 4.1.4). BEETS uses these search trajectories for both evaluation of region relevance and template assessment. (see Section 4.1.4). Given the computational overhead of model maintenance in MBS, the computation time required to compute 3,000 full-atom decoys with MBS and BEETS corresponds to generating 4,000 MC full-atom decoys. Consequently I compare MBS and BEETS searches with 3,000 full-atom decoys with MC searches generating 4,000 decoys. In addition, native-like structures are obtained for comparison by running 100 relaxations on the all-atom structures.

To evaluate the results I compare the all-atom energy, RMSD and GDT_TS [75]. The GDT_TS score is reported in percentage with 100% being a complete match between structures. When RMSD is reported, it refers to RMS calculated by Rosetta given in angstroms. To assess the energy of points in conformation space, I use the unit-less number returned by Rosetta's all-atom energy function.

For experimental evaluation, 36 proteins were chosen of varying size and secondary structure composition. These proteins were selected from recent CASP competitions, and from experiments performed by Blum and Baker [4]. The list of proteins is shown in Table 4.4.

Search was conducted both with homology information and without. Homology information was incorporated into the move set and the fragment collection. Homology information simplifies the search problem because it introduces a structural bias towards homologous structures. Homologs were excluded by removing templates that exceed 20% probability of being a true positive homolog as measured by hh-search [66]. The hh-search probability score incorporates both structure and sequence information. Templates that score below 20% are considered unlikely to be homologs. It should be noted that even with homolog information most templates are

not homologs (see tables 4.4 and 4.4.) This stands in contrast to existing homology modeling where most templates would be homologs.

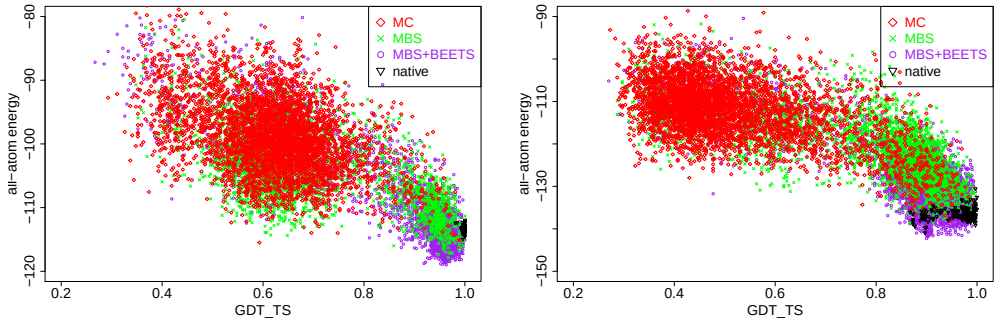
The results have been divide into five categories. Category 1 contains proteins for which both MBS and BEETS make accurate predictions. Category 2 encompasses proteins for which BEETS improves upon the predictions made by MBS. Category 3, contains proteins where structures are found that are lower in energy than the native state, pointing to inaccuracies in the energy function. Category 4 contains proteins for which neither BEETS nor MBS can find structures comparable to the native state in terms of energy or structural similarity. Category 5 contains proteins where MBS outperforms BEETS.

Category 1: adequate conformation space search

For the three proteins in category 1 (see Table 4.4), both MBS and BEETS accurately predict protein structure. Structures are predicted with a GDT_TS of greater than 0.90. Obviously, if MBS finds the global minimum of the energy landscape, BEETS cannot improve the results. All three proteins in this category have homologs in the move set and template library and are relatively small (less than 105 amino acids).

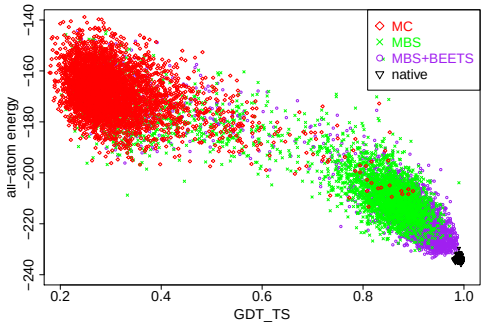
Figure 4.5 shows relaxed native structures, and decoys generated by BEETS, MBS, MC and native for all proteins in category 1. The scatter plots indicate that both MBS and BEETS find conformations in the bottom right of the graph, where the structural match with native structures is very high and the energy is low.

One interesting protein is 1di2. In chapter three, 1di2 was classified in category 1 because both MC and MBS find the native state. Now 1di2 is classified in category 2, indicating it has become more difficult to solve. Between chapter 3 and 4 significant changes have occurred in Rosetta including a move between Rosetta2 and Rosetta3, changes to the fragment set, and changes in the energy function. Rosetta 3 was



(a) 54aa PDB: 5cro

(b) 60aa PDB: 2reb



(c) 105aa PDB: 1aiu

Figure 4.5. For proteins in category 1 both MBS and BEETS adequately search conformation space. Each point in the scatter plots represents a structure prediction. Native is drawn on top, followed by MC, followed by MBS and finally MBS+BEETS.

a complete rewrite of Rosetta 2, which resulted in more object oriented code. In process of making the code object oriented the energy function was slightly changed. For most proteins the new energy function improved predictions. For this reason results in chapter 3 and 4 can not be directly compared. The CASP results from chapter 3 were produced using Rosetta 3.

Category 2: BEETS improves conformation space search

Category 2 consists of proteins for which BEETS searches conformation space more effectively than MBS. When homologous templates were used, 20 of the 36 proteins fell into this category. Using no homologs, 8 of the 36 proteins fell into this category. For all proteins in this category, BEETS finds structures greater than 0.05 GDT_TS more accurately than MBS. These proteins range in size between 54 and 128 amino acids. The improvement of BEETS over MBS is illustrated in the scatter plots in Figures 4.6, 4.7 and 4.8. The scatter plots indicate that BEETS can use information from structural homologs to guide search.

Category 3: Inaccurate energy function

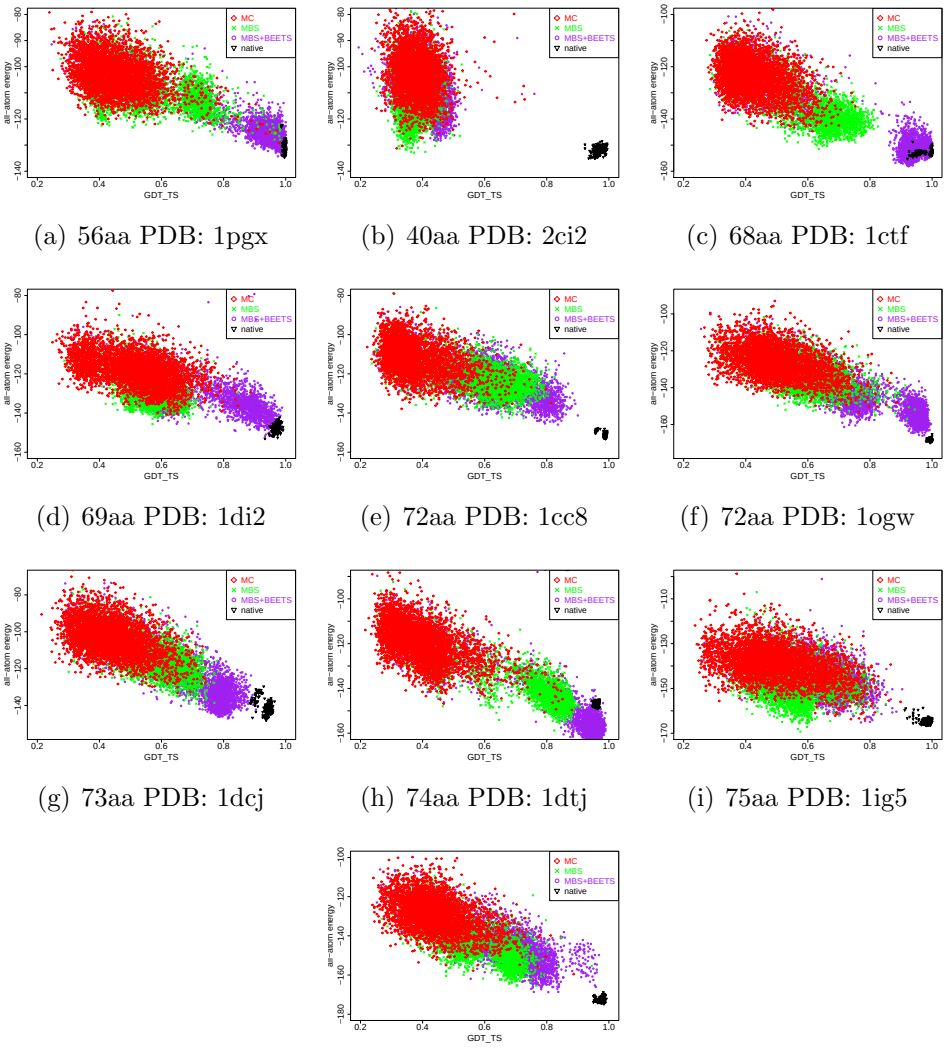
For two proteins all predictions were further than 0.8 GDT_TS away from native and within 0.1 full-atom energy of native. These two proteins are shown in Figure 4.9.

For proteins in this category, reduced energy does not result in accurate predictions. This is a consequence of inaccuracies in the energy function. An inaccurate energy function guides search towards wrong regions of conformation space. No matter how much search is improved, it will be unable able recover from the inaccurate energy function.

Improving the energy function is a challenging task. Work at the Baker lab has shown that it is possible to find structures with lower energy than native for nearly all proteins (unpublished). In this unpublished work all terms in the Rosetta energy function can be reduced below that of native. So fixing the energy function is not a matter of reweighting values. Additionally, much effort has been spent trying to improve the energy by making it more theoretically correct but these improvements have not improved benchmark sets.

One might think different energy functions such as AMBER [51] or CHARMM [10] might more accurately identify the native state. But these methods are no better at ranking Rosetta models than Rosetta (unpublished). In the refinement competition during CASP, better energy functions should result in more accurate structures. However, in CASP 9 Rosetta won that category.

It is likely some undiscovered features are missing in the energy function. Some possibilities include a dipole action in the helices, modeling the system entropy, side chain interaction with water, or some force that is currently modeled only at the quantum level.



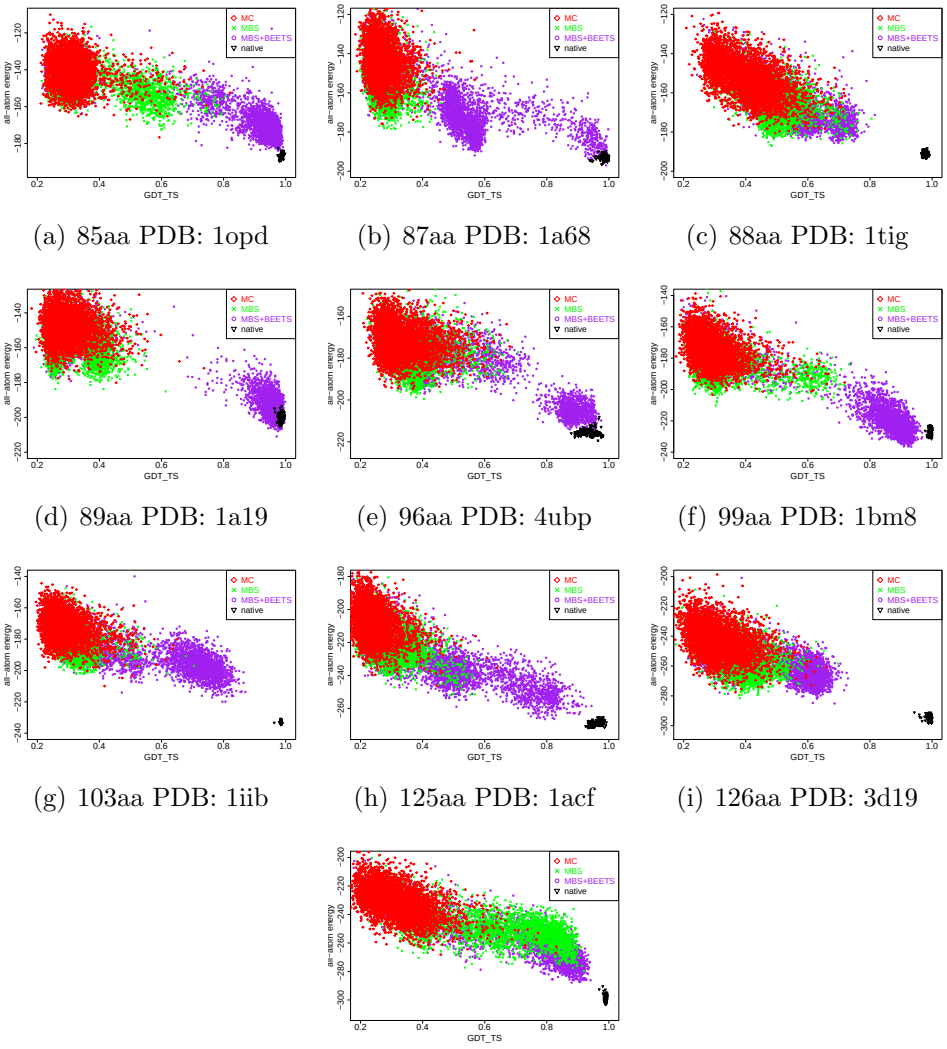
(a) 56aa PDB: 1pgx (b) 40aa PDB: 2ci2 (c) 68aa PDB: 1ctf

(d) 69aa PDB: 1di2 (e) 72aa PDB: 1cc8 (f) 72aa PDB: 1ogw

(g) 73aa PDB: 1dcj (h) 74aa PDB: 1dtj (i) 75aa PDB: 1ig5

(j) 76aa PDB: 1ubi

Figure 4.6. For proteins in category 2 BEETS finds samples at least 0.05 GDT_TS lower than MBS. This plot represents the smallest 10 of 22 proteins in category 2 when homologs exist in the fragment and template set. Each point in the scatter plots represents a structure prediction. Native is drawn on top, followed by MC, followed by MBS and finally MBS+BEETS.



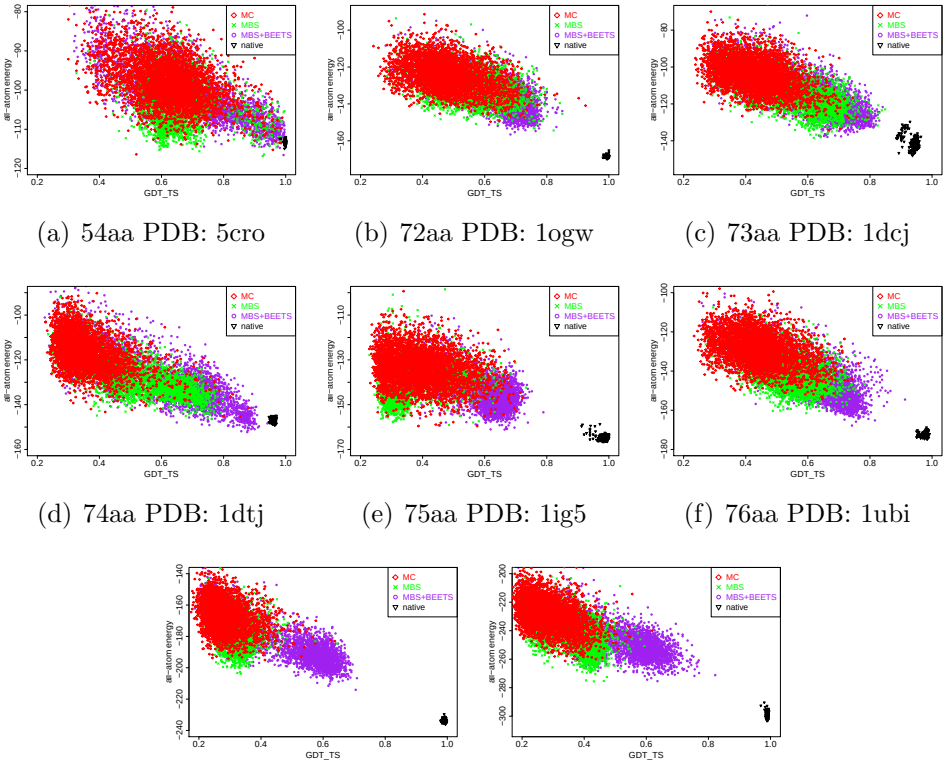
(a) 85aa PDB: 1opd (b) 87aa PDB: 1a68 (c) 88aa PDB: 1tig

(d) 89aa PDB: 1a19 (e) 96aa PDB: 4ubp (f) 99aa PDB: 1bm8

(g) 103aa PDB: 1iib (h) 125aa PDB: 1acf (i) 126aa PDB: 3d19

(j) 128aa PDB: 2chf

Figure 4.7. For proteins in category 2 BEETS finds samples at least 0.05 GDT_TS lower than MBS. This plot shows the larger proteins in category 2. For these proteins homologs exist in the fragment and template set. Each point in the scatter plots represents a structure prediction. Native is drawn on top, followed by MC, followed by MBS and finally MBS+BEETS.

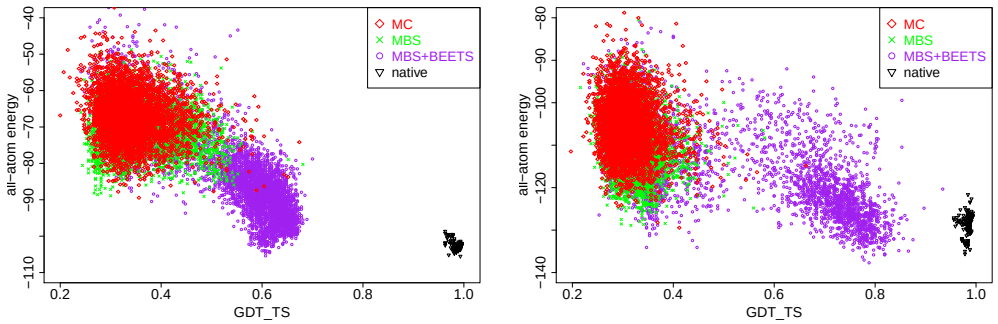


(a) 54aa PDB: 5cro (b) 72aa PDB: 1ogw (c) 73aa PDB: 1dcj

(d) 74aa PDB: 1dtj (e) 75aa PDB: 1ig5 (f) 76aa PDB: 1ubi

(g) 105aa PDB: 1aiu (h) 128aa PDB: 2chf

Figure 4.8. When homology is removed BEETS still produces more accurate structure predictions in many cases. These scatter plots show Category 2 proteins with no sequence based homology is used. The median GDT_TS of the top 1% of energy samples is predicted at least 0.05 GDT_TS lower by MBS. Each point in the scatter plots represents a structure prediction. Native is drawn on top, followed by MC, followed by MBS and finally MBS+BEETS.



(a) 54aa PDB: 1bq9 (b) 64aa PDB: 1scj

Figure 4.9. For proteins in category 3, structures with either lower energy from the native state or with very close to lower energy are found far from the native state. These low energy but not accurate predictions point to inaccuracies in the energy function. Each point in the scatter plots represents a structure prediction. Native is drawn on top, followed by MC, followed by MBS and finally MBS+BEETS.

Category 4: Inadequate conformation space search

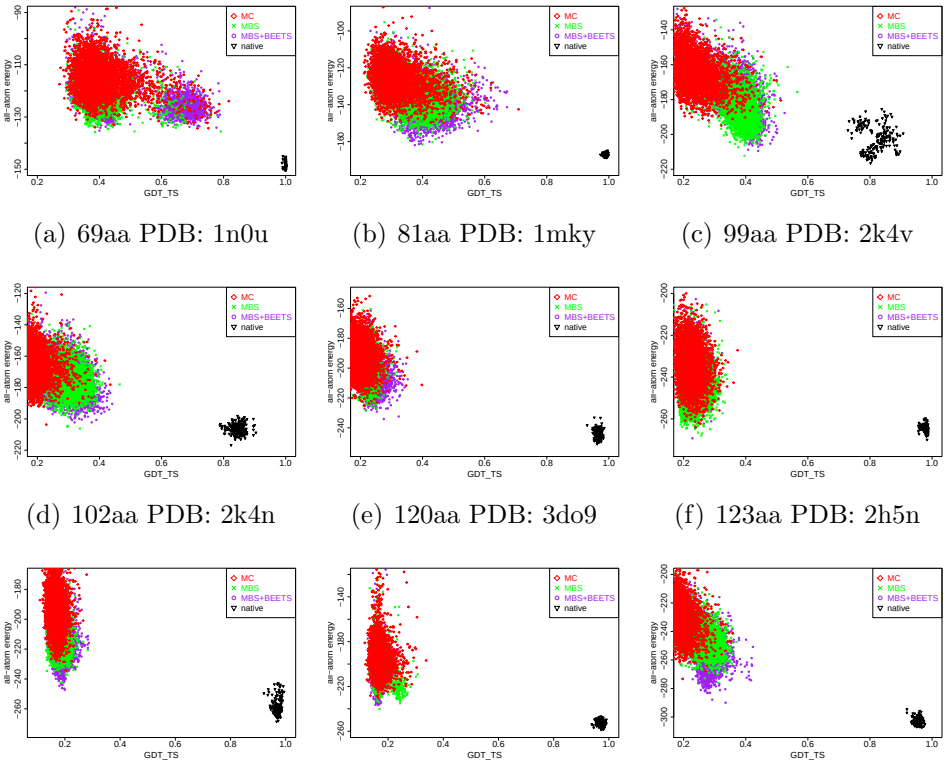
Category 4 contains all proteins where none of the methods adequately search conformation space. I define not working as having less than 0.6 GDT_TS. With homologous information present, nine proteins fall into this category without homology information 22 of the 36 proteins fall into this category.

It should be noted that all proteins with homologs present could be accurately predicted by homology modeling based system. These proteins with homologs present are shown in Figure 4.10.

A comparison of the scatter plots in Figures 4.5 and 4.10 reveals a qualitatively different behavior of search between categories 1 and 4, i.e. between successful search and unsuccessful search. The samples generated for category 4 proteins do not form a trajectory towards the native state whereas those for category 1 and 2 do.

Category 5: BEETS does not improve conformation space search

Category 5 contains all proteins where BEETS does not significantly improve conformation space search. All five of these cases exist when search does not use ho-

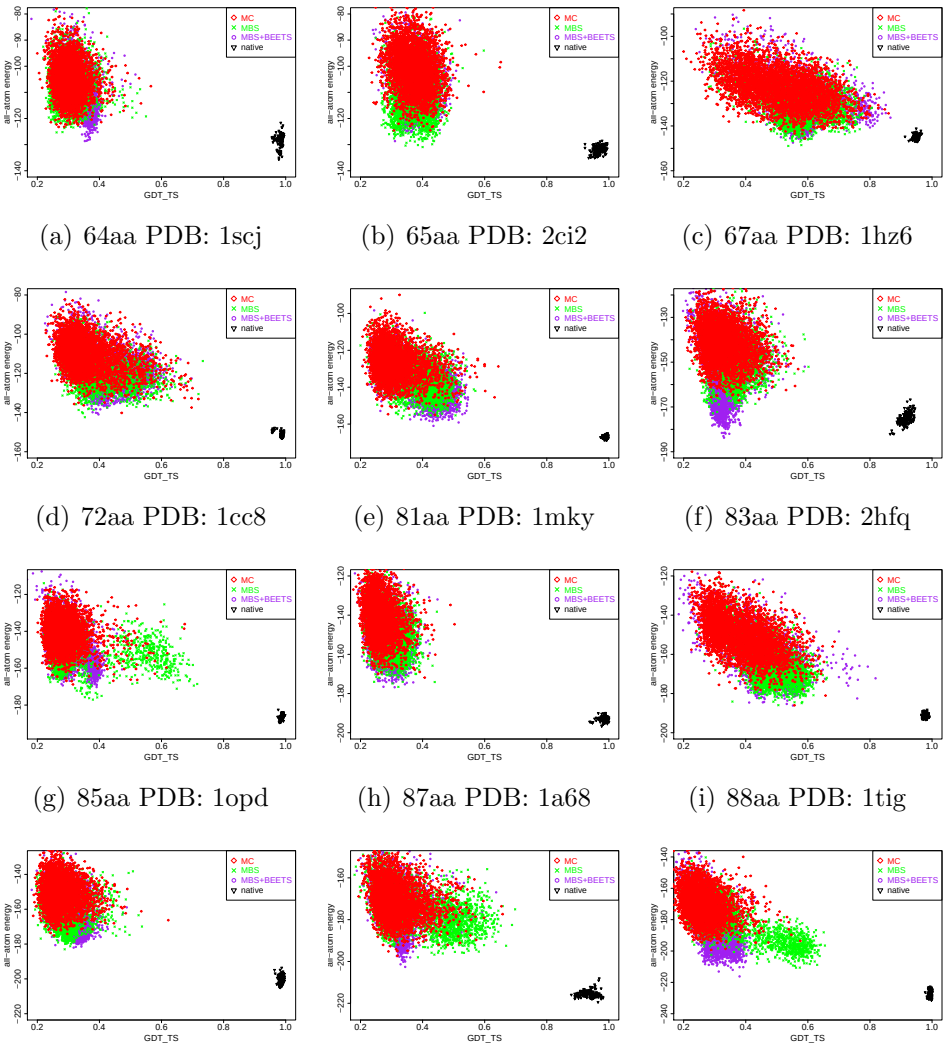


(a) 69aa PDB: 1n0u (b) 81aa PDB: 1mky (c) 99aa PDB: 2k4v

(d) 102aa PDB: 2k4n (e) 120aa PDB: 3do9 (f) 123aa PDB: 2h5n

(g) 127aa PDB: 3d0j (h) 135aa PDB: 2j6a (i) 139aa PDB: 3dnx

Figure 4.10. For proteins in category 4, none of the search methods adequately sample conformation space. The scatter plots in this graph are from the cases where homology information was included.



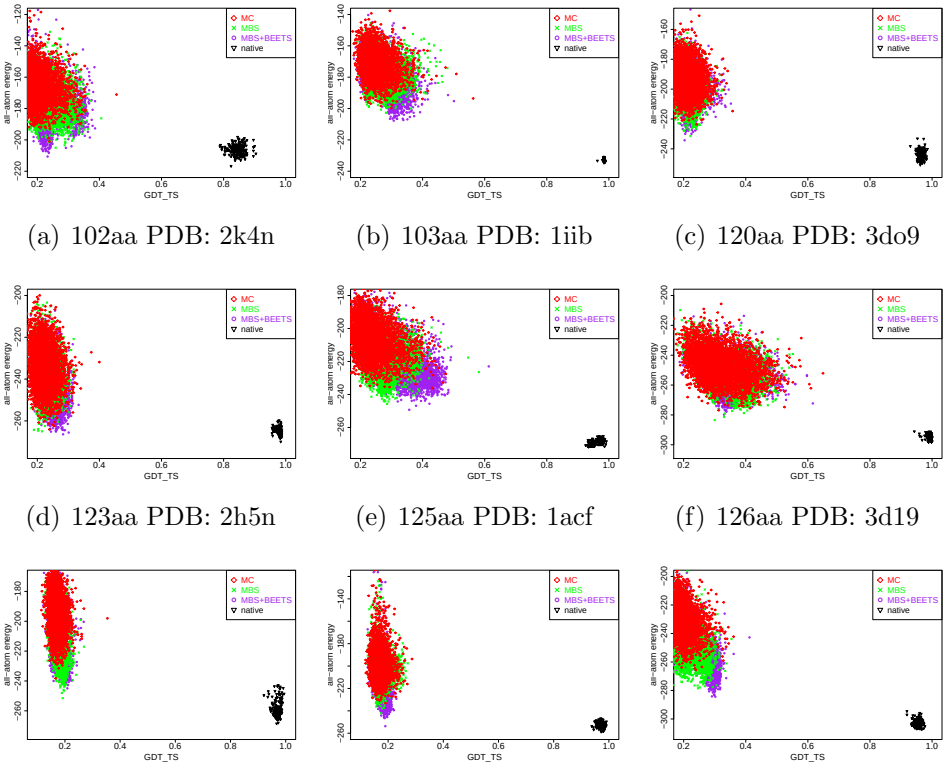
(a) 64aa PDB: 1scj (b) 65aa PDB: 2ci2 (c) 67aa PDB: 1hz6

(d) 72aa PDB: 1cc8 (e) 81aa PDB: 1mky (f) 83aa PDB: 2hfq

(g) 85aa PDB: 1opd (h) 87aa PDB: 1a68 (i) 88aa PDB: 1tig

(j) 89aa PDB: 1a19 (k) 96aa PDB: 4ubp (l) 99aa PDB: 1bm8

Figure 4.11. For proteins in category 4 none of the search methods adequately sample conformation space. The scatter plots in this graph are from the cases where no homology information was included. This Figure continues in Figure 4.12.



(a) 102aa PDB: 2k4n (b) 103aa PDB: 1iib (c) 120aa PDB: 3do9

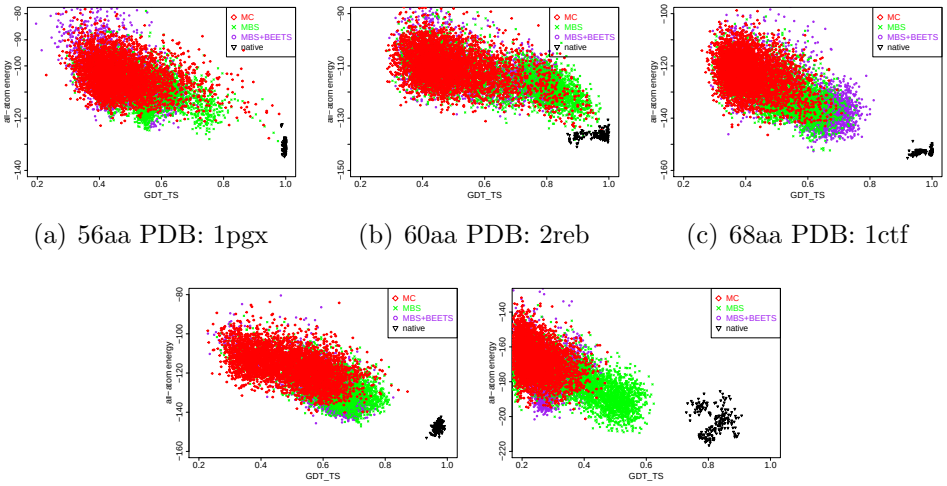
(d) 123aa PDB: 2h5n (e) 125aa PDB: 1acf (f) 126aa PDB: 3d19

(g) 127aa PDB: 3d0j (h) 135aa PDB: 2j6a (i) 139aa PDB: 3dnx

Figure 4.12. For proteins in category 4 none of the search methods adequately sample conformation space. The scatter plots in this graph are from the cases where no homology information was included.

mology. When no homolog exists, BEETS spends computational resources exploiting structures from templates that do not lead search toward native. While investigating these incorrect homologs, computational resources are wasted. So it is expected that MBS would outperform BEETS for some proteins.

Also, if the energy funnel is very narrow, MC-based search with random restarts may in some cases have a higher probability of discovering the entrance to the funnel. This was observed for proteins 1pgx and 2reb.



(a) 56aa PDB: 1pgx (b) 60aa PDB: 2reb (c) 68aa PDB: 1ctf

(d) 69aa PDB: 1di2 (e) 99aa PDB: 2k4v

Figure 4.13. For proteins in category 5 BEETS does not improve MBS and in several cases makes search worse. Examples of category 5 proteins only exist when no homology information exists.

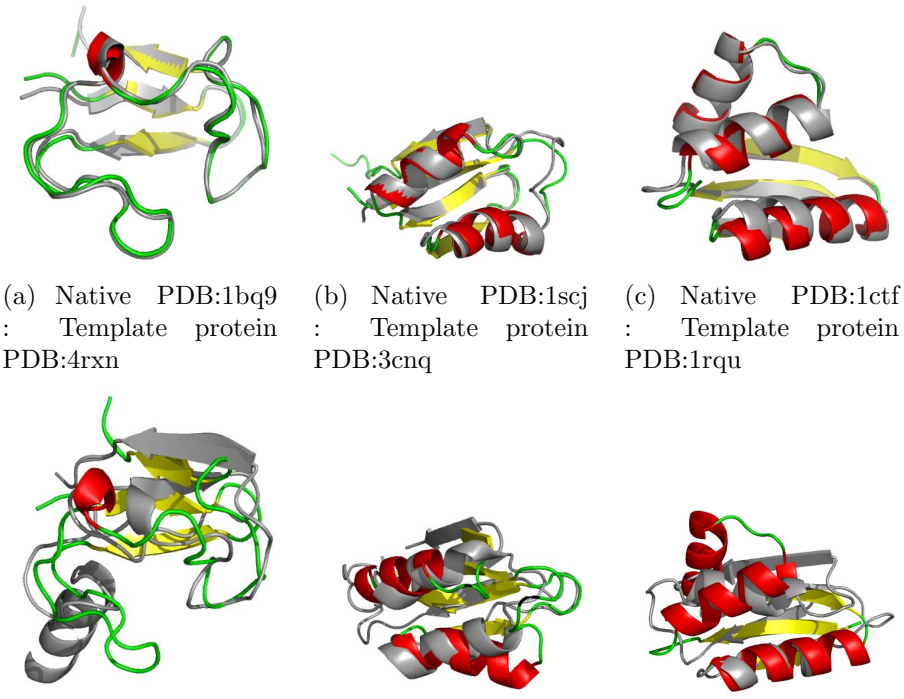
4.4 Analysis of results

The role homologs play in search

In this section I explore how the match between the template protein and target protein effects search performance. To determine the match, I use the structural matching procedure described in section 4.1.2 that identifies structural homologs. In some cases this matching procedure finds homologs in the homology-free data set.

I will refer to the structurally matched templates in the homology-free data set as structural homologs. Proteins identified as homologs by hh-search will be referred to as sequence homologs.

In the template collection with homologs there exists a sequence homolog for 28 of the 36 proteins. In the template collection without homologs there was a structural homolog for 11 proteins shown in Figure 4.15. Sequence homologs match native much better than structural homologs. See Figure 4.14.



(a) Native PDB:1bq9 : Template protein PDB:4rxn

(b) Native PDB:1scj : Template protein PDB:3cnq

(c) Native PDB:1ctf : Template protein PDB:1rqu

(d) Native 1BQ9 : Template protein 3EB5

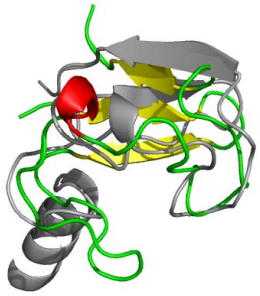
(e) Native 1SCJ : Template protein 1KN6

(f) Native 1CTF : Template protein 1CPZ

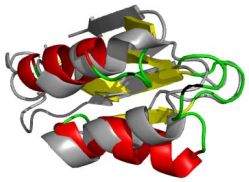
Figure 4.14. Alignment between target and top homolog protein (color) and native state (gray). The top row of images are sequence homologs and the bottom row are the corresponding structural homologs.

Homologs

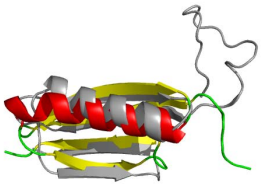
The 28 proteins with homologs in the template collection are shown in table 4.4.



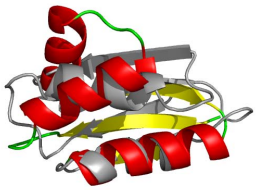
(a) Native 1BQ9 : Template protein 3EB5 (85% residues matched)



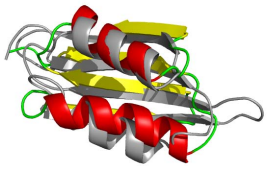
(b) Native 1SCJ : Template protein 1KN6 (94% residues matched)



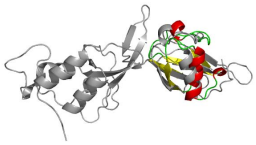
(c) Native 1HZ6 : Template protein 2RPV (88% residues matched)



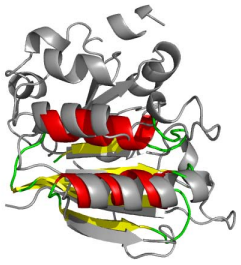
(d) Native 1CTF : Template protein 1CPZ (85% residues matched)



(e) Native 1CC8 : Template protein 2RAQ (100% residues matched)

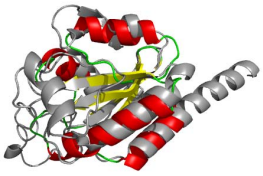


(f) Native 1OGW : Template protein 3JYU (100% residues matched)

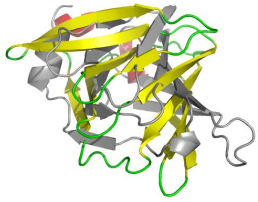


(g) Native 1DCJ : Template protein 2FCA (100% residues matched)

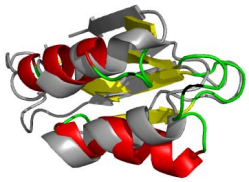
(h) Native 1UBI : Template protein 3JYU (100% residues matched)



(i) Native 1AIU : Template protein 2R37 (91% residues matched)



(j) Native 3D0J : Template protein 2F4P (86% residues matched)



(k) Native 2CHF : Template protein 3EZ6 (84% residues matched)

Figure 4.15. Alignment between structural homolog (color) and native state (gray). Structural homologs were identified with the matching procedure, described in section 4.2.2.

When homologs exist the results show that MBS+BEETS searches conformation space more efficiently then MBS or MC. These target proteins show improvement in energy as shown in Figure 4.19(a). The lower-energy predictions generated by MBS+BEETS result in more accurate structures averaging close to 2 $\mathring{A}$ RMSD with several proteins predicted lower than 1$\mathring{A}$ RMSD. This is illustrated for three proteins in Figure 4.16.

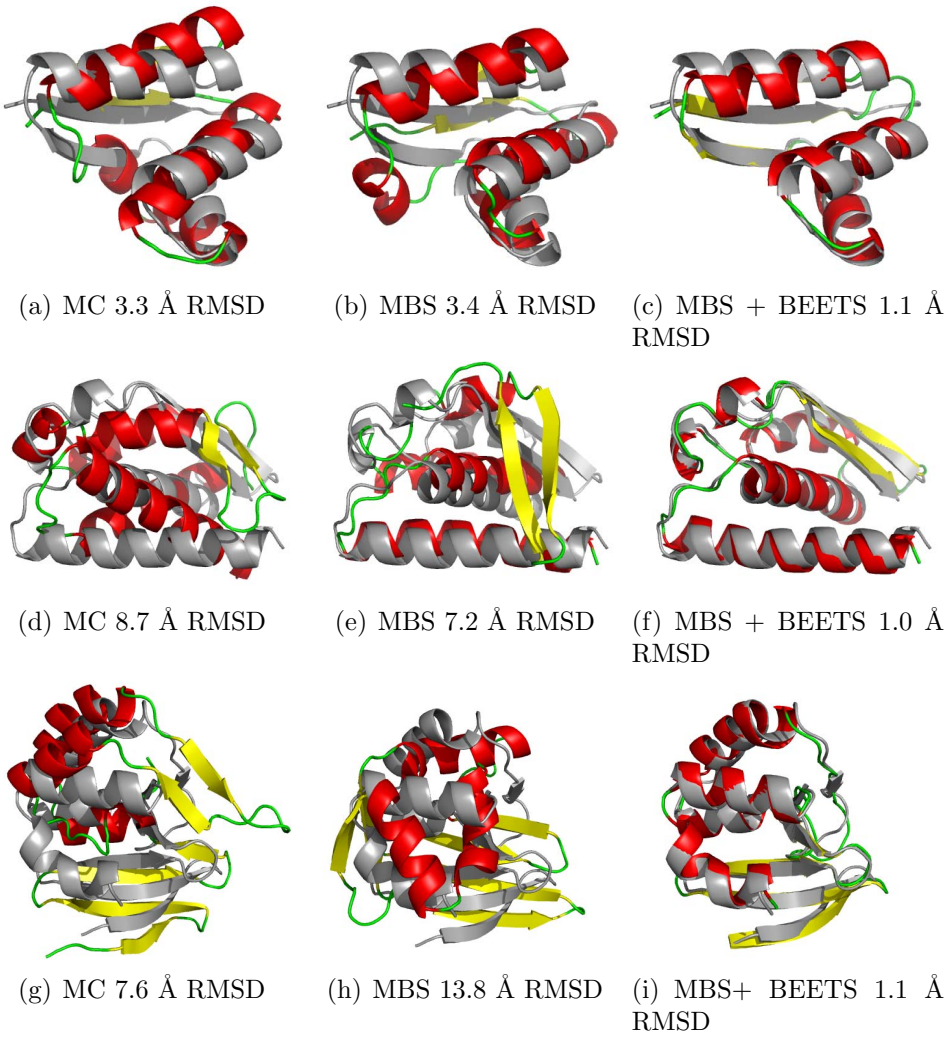
I would expect to see more improvement in RMSD correlate with higher percentage of homologs in the template collection. This correlation should occur because BEETS would be more likely to sample a homolog each time a template change occurred. However, my results show the opposite is true, as seen in Figure 4.17. I hypothesize that this is because proteins with many homologs also have homologous fragments making them easier to predict. There is less room for improvement in easier to predict proteins.

Structural Homologs

When sequence homologs are removed 11 of 36 proteins are found to have structural homologs. (These 11 proteins are shown in Figure 4.15.) Note that structural homologs are more structurally divergent than sequence identified homologs making it more difficult to exploit the relevant information from template protein.

The results obtained for proteins with structural homologs show that MBS+BEETS searches conformation space more effectively than MBS or MC as seen by the decrease in energy, shown in Figure 4.19(b). This reduction in energy corresponds to better RMSD for 9 of the 11 proteins. See Figure 4.4. However, the reduction in RMSD is lower than for proteins with sequence homologs. Three example protein predictions are shown in Figure 4.18.

The difference in RMSD accuracy points to a fundamental flaw in how the method is currently implemented. Currently, exploitation is increased by adding 15 additional



(a) MC 3.3 Å RMSD (b) MBS 3.4 Å RMSD (c) MBS + BEETS 1.1 Å RMSD

(d) MC 8.7 Å RMSD (e) MBS 7.2 Å RMSD (f) MBS + BEETS 1.0 Å RMSD

(g) MC 7.6 Å RMSD (h) MBS 13.8 Å RMSD (i) MBS+ BEETS 1.1 Å RMSD

Figure 4.16. Predicted structures for proteins from the homolog data set (color), superimposed on native structures from the PDB (gray): Ribosomal Protein L7/L12(68aa, PDB:1ctf) predicted with MC (a), MBS (b) and with MBS + BEETS (c); Bacillus Pasteurii Urease (96aa, PDB:4ubp) predicted with MC (d), MBS (e) and with MBS + BEETS (f); DNA-binding domain of MBP1(99aa, PDB:1bm8) predicted with MC (g), MBS (h) and with MBS + BEETS (i). For all three proteins $< 1\%$ of the template proteins in the template collection used in BEETS were homologs.

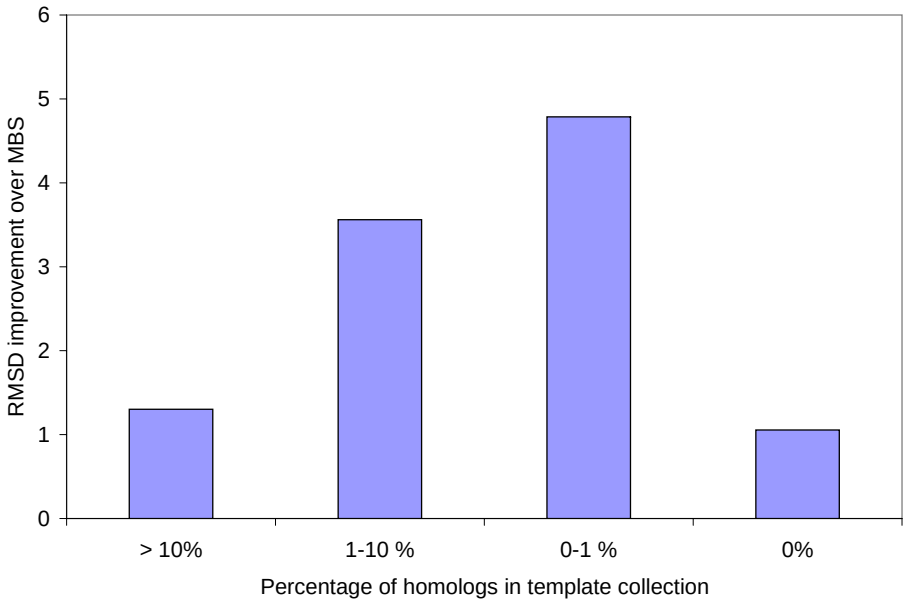


Figure 4.17. RMSD improvement of MBS+BEETS over MBS (higher bars are better). Improvement is seen in all cases with the most significant improvement where there are the fewest homologs in the template collection.

residues to the end of the match. For proteins with structural homologs this increase in exploitation will often occur in loop regions that have significantly changed structure. This problem can easily be fixed by more intelligent ways of increasing exploitation.

I would address this problem by incorporating sequence information into the alignment process. For template based modeling (TBM) evidence suggests sequence based alignment scoring methods can select the best alignment more accurately than energy function based alignment methods. My evidence for this is that HH-search outperformed Rosetta in CASP 9. Therefore a method like BEETS should start from a conformation that uses the best alignment generated from hh-search and only where the alignment has low confidence should search techniques take over.

For free modeling (FM) (or de novo) we should also incorporate sequence information. Secondary structure predictors such as psipred are 80% accurate [38]. Therefore

we should increase exploitation in ways that match the decoy secondary structure to the psipred predicted secondary structure. We should also use predicted secondary structure to throw out templates that can not possibly match the template. For example if a template is all helix, but our target is known to have a β-sheet.

No homolog

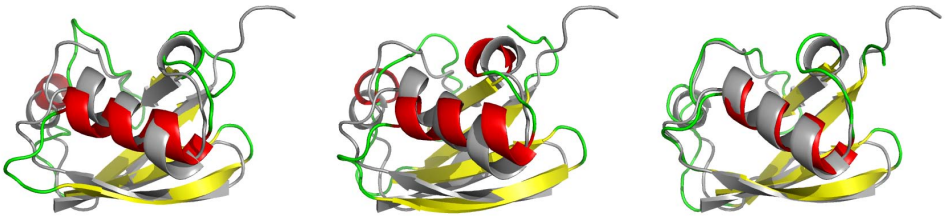
When no homolog is present, MBS+BEETS does not improve prediction accuracy. This need for a structural homolog illustrates the major importance homologs have on improving search accuracy.

For proteins where structural matches do not exist, the two fundamental problems described in chapter 3 still cause failure in search. I repeat these hypotheses below and adjust them with new insights that have occurred over the previous several years.

The first hypothesis states that unsolved proteins in category 3 have energy landscapes with a narrow funnel leading to the native state. The second hypothesis states that inaccurate intermediate energy functions may steer search away from the region containing the native structure.

1. The narrow funnel hypothesis explains why category 3 contains short as well as longer proteins. For small proteins, the conformation space is already too large for search to accidentally discover a small region that represents the entrance to the funnel, unless the energy landscape contains large regions that slope towards it. This narrow funnel hypothesis could easily be caused by a move set of fragments that is less representative of the native structure.

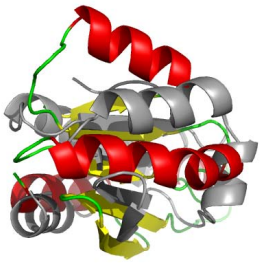
   The narrow funnel hypothesis emphasizes the importance of understanding residual native structure present in the denatured states of proteins. Biological proteins exhibit residual structure as a consequence of interactions among side-chains in close proximity along the backbone. In contrast, MBS and BEETS



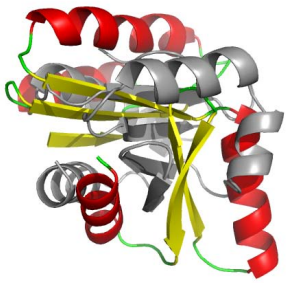
(a) MC 3.5 Å RMSD

(b) MBS 3.6 Å RMSD

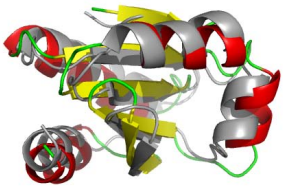
(c) MBS + BEETS 2.6 Å RMSD



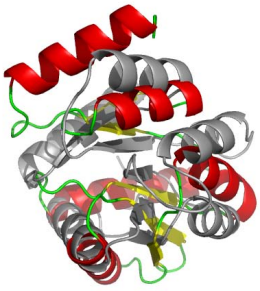
(d) MC 10.7 Å RMSD



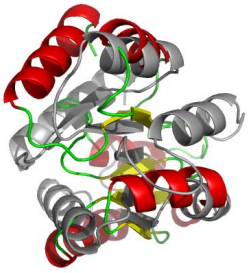
(e) MBS 11.0 Å RMSD



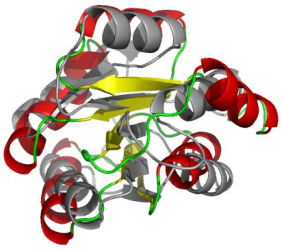
(f) MBS + BEETS 2.7 Å RMSD



(g) MC 10.9 Å RMSD



(h) MBS 10.7 Å RMSD



(i) MBS + BEETS 6.2 Å RMSD

Figure 4.18. Predicted structures for proteins by MBS+BEETS with no sequence homologs present(color), superimposed on native structures from the PDB (gray):Synthetic Ubiquitin protein(76aa, PDB:1ubi) predicted with MC (a), MBS (b) and with MBS + BEETS (c); Human thioredoxin homodimers(105aa, PDB:1aiu) predicted with MC (d), MBS (e) and with MBS + BEETS (f); Mg(2+)-bound form of CheY(128aa, PDB:2chf) predicted with MC (g), MBS (h) and with MBS + BEETS (i).

has to discover this structure by random assembly of fragments, a proposition of vanishingly small probability.

It should be noted that for such narrow funnels, MC-based search with random restarts may in some cases have a higher probability of discovering the entrance to the funnel.

2. A second hypothesis is also consistent with my observations. To find the entrance to the folding funnel in the all-atom energy function, the energy function of stage i must lead samples into the correct funnel of the energy function at stage $i+1$. This may not hold for proteins in category 4. Assume that searching at stage i, MBS identifies the correct minimum of the energy function. If local search in the energy function at stage $i+1$ does not lead to the global minimum when started from the minimum of stage i, search will be guided away from the native structure and is unlikely to recover from it, no matter whether MC or MBS is used as the search strategy. Therefore, my second hypothesis states that for difficult to solve proteins the global minima in consecutive energy functions are shifted, preventing search from identifying the correct folding funnel.

   This second hypothesis, if true, may be an indication that it is time to concentrate effort on improving the low-resolution energy function, or improving how folding follows a specific pathway.

Prognosis for the future

As the protein data bank grows it will become increasingly likely that all the information necessary to solve protein structure prediction will be available. However, the ability to locate and use the relevant template structures still remains a major obstacle to solving protein structure prediction.

In this chapter, I presented a novel way to locate and use relevant template proteins from the protein data bank. I showed that the key to using additional infor-

mation from the PDB is adaptively adjusting how search balances exploitation with exploration of the information in template proteins. Other search methods assume the template protein can be correctly identified, but my approach assumes the identification will have errors and develops a search method that can recover from those mistakes.

I demonstrated that my new method was better at identifying useful information than existing methods. This additional information led to a reduction in energy and improved prediction accuracy. The improvements in prediction accuracy were shown to be strongly linked to the presence of a homolog. The experimental results discussed in this chapter suggest that further improvement to protein structure prediction requires continued improvements to how homologs are located and used, in addition to additional improvements to conformation space search.

		1% RMSD (GDT_TS)				Median RMSD (GDT_TS) of 1% energy					% structural matches
protein	L	MC	MBS	MBS+BEETS	Pos.	MC	MBS	MBS + BEETS	Pos.	Native	in template collection
1bq9A	54	4.55(.56)	4.44(.55)	4.67(.68)	1.95(.83)	9.09(.36)	9.77(.32)	6.10(.61)	2.29(.79)	0.66(.98)	1.9%
5croA	54	1.16(.94)	0.70(.99)	0.68(.98)	0.66(.99)	3.04(.72)	1.07(.97)	1.28(.96)	0.83(.98)	0.57(1.0)	2.7%
1pgx	56	2.14(.82)	1.25(.94)	0.52(1.0)	0.82(.97)	4.05(.70)	1.83(.86)	0.63(.99)	0.90(.95)	0.56(.99)	4.2%
2reb	60	1.07(.95)	0.72(.99)	0.67(.99)	0.89(.98)	1.35(.89)	1.21(.92)	1.18(.91)	1.29(.92)	1.28(.90)	2.6%
1scjB	64	5.57(.47)	5.97(.48)	2.21(.85)	2.10(.82)	9.82(.32)	9.80(.34)	2.48(.78)	2.85(.73)	0.75(.98)	5.3%
2ci2I	65	5.40(.55)	6.29(.49)	5.66(.53)	2.22(.83)	9.17(.39)	10.67(.36)	6.68(.45)	2.78(.78)	1.03(.95)	6.7%
1hz6A	67	2.78(.76)	2.50(.80)	2.61(.78)	2.07(.88)	3.87(.58)	4.33(.56)	3.66(.57)	2.60(.77)	1.42(.95)	1.7%
1ctf	68	3.17(.67)	2.59(.80)	0.62(1.0)	2.23(.80)	4.02(.58)	3.55(.65)	1.37(.92)	2.75(.69)	1.13(.94)	0.98%
1di2	69	2.55(.78)	3.10(.72)	0.94(.97)	1.05(.96)	4.37(.63)	4.17(.59)	1.25(.94)	1.32(.92)	0.97(.96)	14.9%
1n0u	69	2.82(.74)	3.20(.68)	2.62(.75)	1.14(.94)	7.05(.46)	6.86(.44)	6.77(.45)	1.46(.91)	0.50(1.0)	6.2%
1cc8A	72	3.52(.69)	2.32(.79)	2.10(.86)	1.71(.89)	6.47(.51)	5.06(.67)	2.55(.80)	2.07(.83)	0.67(.99)	4.2%
1ogw	72	2.31(.79)	1.63(.87)	0.69(.99)	0.73(.98)	2.83(.70)	2.79(.70)	0.91(.96)	0.85(.97)	0.59(.99)	49.1%
1dcj	73	3.22(.70)	2.54(.75)	1.63(.88)	1.67(.86)	4.49(.63)	3.26(.70)	2.28(.80)	2.48(.75)	1.01(.94)	2.7%
1dtj	74	2.98(.79)	2.47(.89)	0.76(.99)	1.37(.92)	4.85(.59)	3.29(.85)	1.11(.95)	1.82(.90)	0.98(.96)	25.4%
1ig5A	75	2.20(.80)	2.20(.80)	1.97(.73)	1.46(.91)	2.80(.71)	6.78(.58)	2.59(.73)	2.30(.81)	0.66(.98)	6.8%
1ubi	76	3.04(.72)	2.59(.78)	1.76(.94)	1.58(.96)	4.80(.56)	3.12(.69)	2.94(.79)	1.90(.95)	0.83(.98)	43.8%
1mkyA	81	4.01(.56)	4.08(.54)	3.58(.62)	2.26(.78)	6.45(.37)	6.24(.39)	5.33(.39)	2.85(.68)	0.66(.99)	0.83%
2hfq	83	5.44(.53)	6.47(.51)	7.46(.51)	2.21(.79)	11.13(.37)	11.47(.35)	10.96(.41)	3.53(.59)	1.60(.87)	0%
1opd	85	4.75(.65)	3.16(.75)	0.75(.98)	1.44(.91)	11.19(.34)	4.60(.60)	0.87(.97)	2.06(.84)	0.75(.98)	6.7%
1a68	87	6.58(.45)	8.14(.42)	0.78(.99)	1.39(.88)	12.07(.29)	9.45(.30)	0.87(.97)	2.20(.81)	0.74(.98)	4.7%
1tig	88	3.39(.67)	3.06(.73)	2.54(.76)	1.63(.90)	5.25(.52)	11.12(.50)	2.97(.80)	2.19(.85)	0.72(.98)	1.3%
1a19A	89	4.51(.52)	4.29(.51)	0.69(.99)	1.46(.91)	11.21(.29)	9.84(.39)	0.82(.97)	1.70(.87)	0.71(.98)	1.2%
4ubpA	96	3.97(.59)	3.35(.63)	1.10(.96)	1.77(.85)	9.23(.33)	8.15(.37)	1.48(.94)	2.20(.80)	0.94(.97)	0.78%
1bm8	99	4.36(.58)	3.50(.68)	1.08(.95)	1.90(.85)	10.51(.34)	7.83(.48)	1.37(.92)	2.99(.75)	0.57(1.0)	0.67%
2k4v	99	7.45(.43)	6.57(.48)	7.85(.48)	3.62(.64)	11.87(.29)	14.05(.40)	13.31(.43)	5.96(.52)	2.25(.80)	0%
2k4n	102	10.96(.35)	11.50(.40)	10.15(.42)	4.45(.64)	16.65(.20)	16.70(.30)	13.77(.32)	9.45(.52)	2.37(.84)	0%
1iibA	103	4.95(.53)	5.64(.54)	2.21(.84)	1.41(.91)	10.86(.36)	10.27(.35)	2.79(.80)	1.69(.87)	0.71(.98)	0.47%
1aiu	105	2.32(.79)	1.02(.95)	0.80(.98)	1.20(.92)	2.69(.76)	1.39(.91)	1.07(.95)	1.25(.91)	0.63(.99)	19.0%
3do9(D1)	120	10.39(.31)	10.95(.28)	8.25(.33)	3.04(.68)	14.86(.22)	12.75(.23)	12.94(.24)	3.99(.57)	1.03(.96)	0%
2h5n	123	10.01(.32)	10.61(.32)	11.04(.29)	2.80(.72)	13.46(.24)	13.25(.23)	12.66(.25)	7.00(.37)	0.83(.98)	0%
1acf	125	5.77(.46)	5.03(.53)	1.56(.89)	1.64(.88)	12.04(.30)	5.56(.49)	2.50(.79)	1.83(.86)	1.01(.94)	5.8%
3d19(D1)	126	4.19(.56)	4.36(.58)	2.83(.69)	2.02(.83)	8.90(.38)	8.24(.40)	3.22(.63)	2.99(.69)	0.60(1.0)	0.46%
3d0j	127	11.76(.24)	11.48(.24)	9.44(.26)	4.27(.52)	16.06(.17)	14.37(.18)	14.07(.19)	7.20(.33)	0.84(.97)	0%
2chf	128	3.49(.65)	1.56(.89)	1.35(.93)	1.56(.88)	9.86(.43)	1.88(.86)	1.55(.90)	1.81(.83)	0.66(.99)	19.8%
2j6a	135	11.71(.27)	12.96(.27)	13.59(.21)	4.67(.50)	16.32(.18)	17.63(.23)	17.06(.16)	17.36(.21)	0.79(.97)	0%
3dnx	139	11.35(.31)	11.01(.35)	10.10(.38)	3.39(.67)	15.74(.21)	15.14(.29)	12.15(.28)	4.63(.55)	1.15(.96)	0%

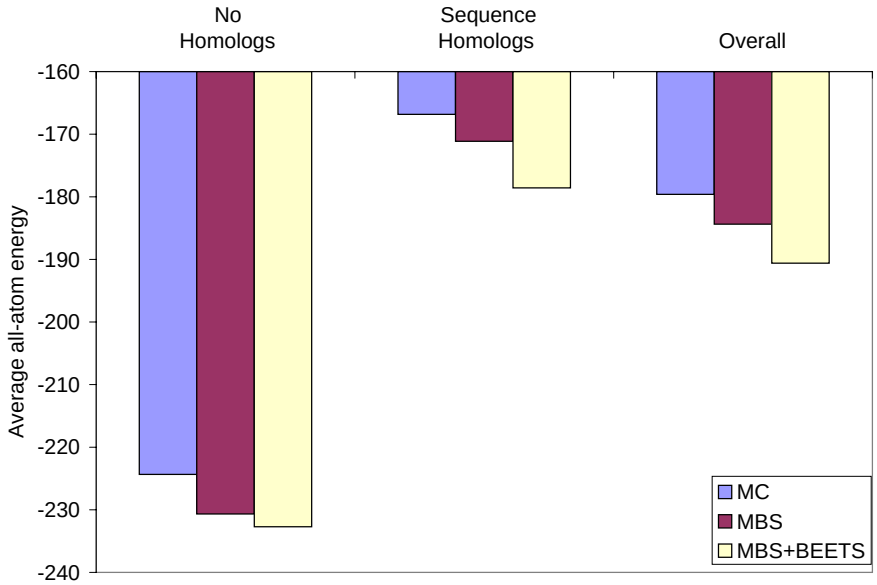
Table 4.1. The move set and template collection with homologs is used by the search methods to predict the structure of 36 proteins. The 36 proteins are shown by PDB code and chain. The next 4 columns give the average RMSD of the top 1% samples by RMSD, with the average GDT_TS of the top 1% of samples by GDT_TS shown in parenthesis. The next 5 columns give the median RMSD, and median GDT_TS in parenthesis, of the top 1% samples by energy. The final column indicates the percentage of template proteins with 80% residues matched within 2.0 $\mathring{A}$ RMSD in the template collection used by BEETS.

		1% RMSD(GDT_TS)				Median RMSD(GDT_TS) of 1% energy					% homologs
protein	L	MC	MBS	MBS + BEETS	Pos.	MC	MBS	MBS + BEETS	Pos.	Native	template collection
1bq9A	54	4.59(.55)	4.30(.60)	5.78(.48)	1.90(.84)	9.08(.35)	5.56(.53)	7.37(.34)	2.11(.79)	0.66(.98)	0.68%
5croA	54	1.16(.94)	0.84(.97)	0.72(.99)	0.71(.99)	2.95(.71)	4.02(.67)	1.06(.95)	1.07(.97)	0.57(1.0)	0%
1pgx	56	1.87(.85)	1.82(.84)	4.25(.68)	0.87(.96)	3.92(.60)	3.01(.71)	5.35(.57)	0.88(.96)	0.56(.99)	0%
2reb	60	1.35(.91)	1.02(.95)	1.59(.86)	0.93(.98)	1.84(.85)	1.35(.89)	2.39(.81)	1.21(.92)	1.28(.90)	0%
1scjB	64	6.06(.45)	5.64(.47)	6.59(.43)	2.35(.78)	10.56(.31)	11.41(.30)	9.79(.37)	2.80(.71)	0.75(.98)	1.4%
2ci2I	65	5.74(.51)	6.11(.48)	6.25(.48)	2.54(.80)	9.23(.38)	8.62(.40)	9.62(.36)	2.80(.77)	1.03(.95)	0%
1hz6A	67	2.60(.78)	2.68(.76)	2.34(.82)	1.97(.89)	3.81(.61)	3.76(.57)	3.59(.58)	2.42(.79)	1.42(.95)	3.8%
1ctf	68	3.23(.66)	2.93(.73)	2.51(.77)	2.07(.83)	4.18(.57)	3.93(.63)	3.41(.66)	3.20(.68)	1.13(.94)	0.84%
1di2	69	2.81(.76)	2.32(.80)	2.82(.76)	1.34(.95)	4.95(.59)	3.04(.71)	3.57(.69)	1.76(.91)	0.97(.96)	0%
1n0u	69	2.57(.76)	2.68(.75)	2.45(.71)	1.07(.96)	3.60(.57)	3.27(.68)	3.14(.71)	1.51(.93)	0.50(1.0)	0%
1cc8A	72	3.74(.65)	3.62(.64)	3.79(.61)	1.72(.88)	6.18(.44)	7.64(.38)	7.16(.40)	2.25(.81)	0.67(.99)	0.59%
1ogw	72	2.54(.74)	2.35(.78)	2.19(.80)	0.79(.98)	3.23(.64)	3.21(.67)	2.77(.72)	0.94(.96)	0.59(.99)	10.1%
1dcj	73	3.28(.69)	2.40(.79)	2.11(.81)	1.79(.85)	5.86(.59)	4.27(.64)	3.07(.70)	2.25(.80)	1.01(.94)	4.9%
1dtj	74	3.52(.73)	3.86(.77)	1.90(.90)	1.50(.92)	5.22(.54)	4.37(.73)	2.40(.88)	2.24(.86)	0.98(.96)	0%
1ig5A	75	3.13(.67)	3.52(.64)	2.60(.73)	1.84(.87)	6.74(.46)	9.38(.32)	3.21(.63)	2.58(.74)	0.66(.98)	0%
1ubi	76	3.13(.69)	2.87(.73)	2.37(.83)	1.54(.95)	4.68(.53)	3.69(.60)	3.11(.74)	1.96(.93)	0.83(.98)	10.2%
1mkyA	81	4.23(.54)	4.25(.52)	4.15(.53)	2.43(.76)	6.56(.35)	5.31(.42)	5.05(.44)	3.26(.62)	0.66(.99)	0%
2hfq	83	5.93(.51)	4.72(.52)	6.85(.50)	2.36(.76)	10.99(.35)	11.28(.35)	10.34(.33)	2.70(.72)	1.60(.87)	0%
1opd	85	5.46(.53)	3.78(.67)	6.44(.42)	1.79(.87)	10.9(.30)	5.41(.37)	8.60(.38)	2.33(.81)	0.75(.98)	0%
1a68	87	7.40(.41)	8.34(.40)	9.09(.39)	2.56(.74)	12.08(.29)	12.97(.32)	10.05(.30)	3.1(.67)	0.74(.98)	0%
1tig	88	3.96(.62)	3.58(.63)	2.97(.71)	2.03(.85)	6.49(.48)	11.24(.54)	6.48(.52)	2.41(.79)	0.72(.98)	0%
1a19A	89	5.77(.44)	4.57(.47)	6.50(.41)	1.79(.85)	12.17(.27)	9.68(.30)	10.87(.33)	2.10(.80)	0.71(.98)	0%
4ubpA	96	4.48(.54)	3.25(.63)	4.83(.52)	2.07(.84)	9.61(.34)	4.70(.47)	12.73(.34)	2.94(.68)	0.94(.97)	0%
1bm8	99	5.74(.46)	4.29(.62)	7.77(.41)	2.17(.83)	11.98(.29)	5.18(.56)	11.04(.33)	2.58(.78)	0.57(1.0)	0%
2k4v	99	7.01(.43)	4.13(.60)	5.66(.44)	3.63(.63)	11.59(.28)	5.17(.53)	12.70(.28)	5.44(.46)	2.25(.80)	0%
2k4n	102	10.78(.34)	12.13(.35)	11.69(.37)	4.80(.58)	15.73(.22)	17.08(.26)	15.36(.23)	8.63(.47)	2.37(.84)	0%
1iibA	103	6.94(.39)	6.50(.42)	8.03(.41)	1.72(.86)	12.64(.27)	11.35(.28)	11.37(.32)	2.72(.75)	0.71(.98)	0%
1aiu	105	5.19(.50)	5.90(.46)	4.19(.68)	1.72(.85)	11.40(.33)	11.51(.33)	4.68(.63)	2.14(.81)	0.63(.99)	6.8%
3do9(D1)	120	9.36(.30)	10.12(.29)	10.87(.32)	3.26(.65)	14.66(.22)	14.65(.23)	14.56(.22)	4.51(.58)	1.03(.96)	0%
2h5n	123	10.04(.32)	10.53(.30)	10.95(.31)	2.73(.73)	13.46(.24)	14.11(.22)	12.29(.26)	7.00(.38)	0.83(.98)	0%
1acf	125	6.58(.42)	6.67(.41)	6.85(.48)	1.92(.84)	11.98(.28)	11.40(.28)	11.20(.38)	2.42(.76)	1.01(.94)	0%
3d19(D1)	126	4.50(.53)	5.01(.50)	3.94(.53)	2.09(.82)	9.05(.38)	10.92(.39)	8.91(.34)	2.87(.72)	0.60(1.0)	0%
3d0j	127	11.27(.24)	11.04(.23)	10.74(.23)	4.82(.47)	15.31(.18)	15.32(.19)	12.72(.21)	14.76(.19)	0.84(.97)	1.9%
2chf	128	5.09(.49)	4.82(.53)	3.54(.74)	2.02(.83)	11.68(.36)	9.50(.42)	4.15(.64)	2.77(.73)	0.66(.99)	0.35%
2j6a	135	11.87(.25)	12.64(.25)	12.68(.22)	4.72(.48)	16.89(.18)	15.36(.17)	15.28(.19)	13.94(.19)	0.79(.97)	0%
3dnx	139	10.96(.31)	11.62(.31)	12.11(.33)	3.79(.65)	15.49(.20)	16.88(.26)	15.30(.30)	5.22(.47)	1.15(.96)	0%

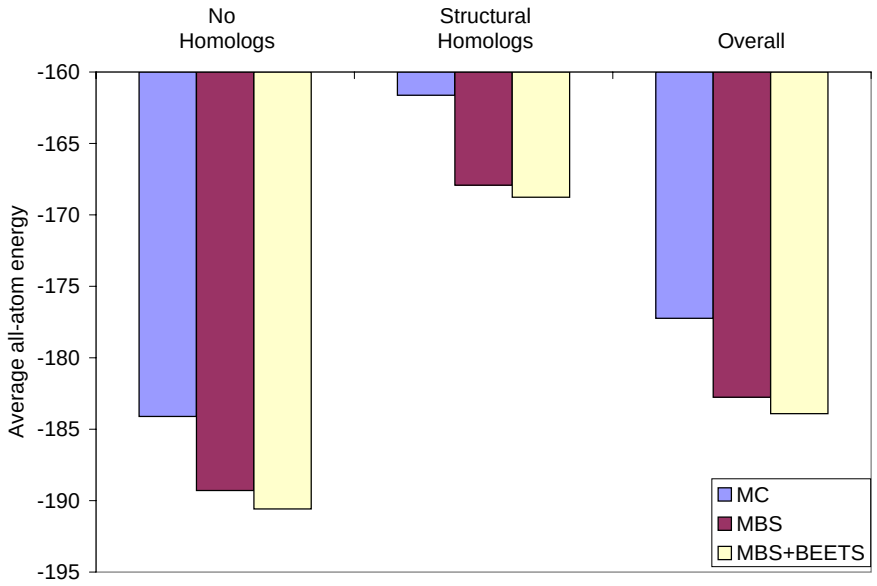
Table 4.2. The homolog free move set and template collection is used by the search methods to predict the structure of 36 proteins. The proteins are shown by PDB code and chain. The next 4 columns give the average RMSD of the top 1% samples by RMSD, with the average GDT_TS of the top 1% of samples by GDT_TS shown in parenthesis. The next 5 columns give the median RMSD, and median GDT_TS in parenthesis, of the top 1% samples by energy. The final column indicates the percentage of template proteins with 80% residuse matched within 3.5 $\mathring{A}$ RMSD in the template collection used by BEETS.

		Homolog move set				Homolog Free move set				
protein	L	MC	MBS	MBS + BEETS	Pos.	MC	MBS	MBS + BEETS	Pos.	Native
1bq9A	54	-89.4(-84.1)	-89.1(-86.2)	-105.4(-102.5)	-99.5(-94.7)	-89.7(-83.9)	-92.9(-87.3)	-90.0(-83.5)	-103.4(-94.8)	-105.5(-105.1)
5croA	54	-115.5(-112.2)	-117.9(-116.1)	-119.0(-118.3)	-117.3(-116.1)	-116.4(-111.6)	-116.8(-114.5)	-116.6(-113.6)	-118.5(-115.6)	-114.9(-114.6)
1pgx	56	-129.2(-120.5)	-127.9(-124.2)	-134.3(-133.2)	-131.3(-128.8)	-127.7(-119.3)	-128.8(-124.3)	-124.5(-121.6)	-130.9(-126.8)	-134.7(-134.4)
2reb	60	-138.0(-133.3)	-138.6(-136.6)	-142.2(-140.9)	-137.2(-134.8)	-134.3(-129.2)	-138.3(-132.5)	-129.3(-126.3)	-136.6(-132.7)	-140.9(-140.1)
1scjB	64	-129.4(-121.5)	-128.9(-125.8)	-137.7(-134.3)	-123.1(-117.9)	-123.2(-120.7)	-125.3(-122.0)	-128.6(-125.4)	-117.0(-112.8)	-135.6(-135.1)
2ci2I	65	-131.2(-123.6)	-132.9(-129.9)	-128.3(-126.6)	-121.8(-119.3)	-124.6(-120.3)	-130.9(-127.9)	-128.7(-124.3)	-117.3(-113.9)	-135.1(-135.0)
1hz6A	67	-144.2(-140.7)	-150.1(-146.4)	-145.5(-142.5)	-140.4(-138.2)	-145.6(-141.0)	-148.9(-145.4)	-147.4(-144.5)	-143.8(-140.7)	-147.0(-146.9)
1ctf	68	-144.7(-141.4)	-151.5(-149.0)	-158.1(-157.3)	-142.3(-140.1)	-146.4(-141.0)	-152.3(-147.8)	-152.7(-148.0)	-145.1(-151.0)	-155.2(-154.8)
1di2	69	-140.8(-137.1)	-145.6(-141.3)	-151.9(-148.3)	-149.4(-145.5)	-140.4(-135.8)	-147.2(-143.2)	-145.6(-143.0)	-144.8(-140.9)	-154.1(-152.5)
1n0u	69	-134.3(-130.3)	-135.8(-133.5)	-135.2(-132.9)	-139.7(-136.4)	-132.7(-130.4)	-140.4(-138.0)	-139.9(-137.5)	-139.2(-136.9)	-150.3(-150.3)
1cc8A	72	-137.9(-132.8)	-142.7(-138.6)	-145.9(-142.9)	-142.2(-135.5)	-140.2(-132.3)	-142.4(-136.4)	-141.0(-136.6)	-140.5(-135.1)	-153.1(-153.1)
1ogw	72	-153.2(-146.8)	-157.0(-151.3)	-166.1(-163.4)	-162.9(-160.0)	-149.7(-143.6)	-154.4(-148.8)	-156.6(-152.5)	-160.1(-156.3)	-169.9(-169.6)
1dcj	73	-133.2(-127.0)	-140.1(-135.1)	-147.0(-145.1)	-135.8(-130.1)	-134.3(-125.0)	-138.4(-134.1)	-137.9(-135.0)	-134.8(-129.9)	-148.0(-147.6)
1dtj	74	-150.7(-142.9)	-158.8(-153.9)	-163.9(-162.7)	-151.2(-148.0)	-145.3(-138.3)	-146.5(-143.4)	-152.3(-149.5)	-149.3(-146.7)	-149.2(-148.9)
1ig5A	75	-163.7(-158.4)	-169.3(-164.7)	-165.2(-161.5)	-164.5(-162.5)	-159.5(-154.2)	-159.1(-155.7)	-161.1(-158.1)	-163.3(-160.0)	-166.6(-166.5)
1ubi	76	-159.6(-150.6)	-165.5(-162.9)	-168.8(-165.4)	-165.5(-162.3)	-154.2(-149.0)	-161.5(-158.3)	-167.8(-162.9)	-165.9(-159.1)	-174.8(-174.6)
1mkyA	81	-150.1(-146.7)	-161.5(-153.7)	-162.0(-158.0)	-157.6(-151.6)	-155.0(-147.7)	-156.9(-153.6)	-161.0(-156.8)	-154.3(-150.7)	-168.7(-168.7)
2hfq	83	-169.2(-163.0)	-172.0(-168.2)	-172.7(-168.8)	-162.2(-161.1)	-167.1(-162.2)	-172.8(-168.6)	-183.7(-179.0)	-166.8(-158.9)	-182.1(-181.9)
1opd	85	-176.6(-161.6)	-172.5(-167.9)	-183.8(-181.9)	-167.7(-165.5)	-168.8(-162.1)	-177.0(-171.4)	-170.2(-167.3)	-174.6(-164.5)	-189.5(-189.4)
1a68	87	-175.8(-168.3)	-178.5(-175.4)	-197.0(-192.0)	-181.8(-177.6)	-171.9(-166.1)	-176.6(-171.0)	-176.8(-173.2)	-171.8(-164.6)	-196.6(-196.2)
1tig	88	-180.6(-176.1)	-187.0(-182.8)	-185.6(-183.7)	-185.4(-180.8)	-186.0(-175.3)	-185.9(-182.1)	-186.1(-181.8)	-179.5(-175.8)	-193.3(-193.3)
1a19A	89	-180.0(-171.2)	-185.1(-178.6)	-211.0(-205.7)	-192.4(-186.1)	-175.1(-170.4)	-183.0(-178.8)	-182.1(-179.4)	-183.5(-178.5)	-205.0(-204.8)
4ubpA	96	-194.7(-190.9)	-199.5(-195.8)	-217.1(-213.3)	-202.0(-193.6)	-197.3(-190.6)	-198.0(-193.9)	-202.6(-197.2)	-188.9(-185.9)	-219.4(-218.4)
1bm8	99	-203.3(-196.9)	-210.1(-204.1)	-236.4(-234.4)	-205.8(-202.7)	-199.3(-192.8)	-210.6(-206.4)	-216.2(-208.3)	-205.5(-198.3)	-231.5(-231.0)
2k4v	99	-194.9(-187.7)	-212.2(-204.0)	-209.5(-203.9)	-188.2(-182.3)	-201.3(-191.9)	-209.4(-206.3)	-201.6(-196.9)	-185.9(-176.2)	-216.5(-215.3)
2k4n	102	-203.6(-189.7)	-206.2(-195.8)	-202.3(-197.9)	-203.4(-186.6)	-201.0(-190.3)	-205.0(-198.5)	-210.6(-204.6)	-189.7(-182.4)	-216.7(-214.2)
1iibA	103	-210.0(-195.7)	-202.3(-200.1)	-219.9(-214.6)	-215.9(-207.7)	-199.8(-193.7)	-202.7(-197.7)	-207.7(-204.4)	-205.9(-199.7)	-234.6(-234.3)
1aiu	105	-213.3(-200.4)	-230.5(-226.4)	-233.4(-231.5)	-229.4(-220.0)	-195.1(-190.2)	-203.9(-198.8)	-214.1(-206.9)	-207.9(-201.7)	-236.1(-235.6)
3do9(D1)	120	-225.1(-218.0)	-228.5(-222.6)	-234.1(-225.8)	-217.2(-201.6)	-223.4(-216.9)	-231.4(-224.2)	-234.3(-225.9)	-213.1(-203.1)	-250.7(-250.0)
2h5n	123	-264.1(-257.6)	-268.0(-263.0)	-269.3(-258.4)	-252.1(-240.1)	-264.1(-257.6)	-265.0(-259.2)	-266.4(-263.0)	-252.1(-240.8)	-269.9(-268.7)
1acf	125	-246.3(-235.3)	-251.9(-244.8)	-266.5(-263.1)	-254.1(-243.6)	-236.1(-231.4)	-245.8(-241.0)	-249.3(-242.0)	-239.8(-231.6)	-271.7(-271.6)
3d19(D1)	126	-277.5(-270.5)	-281.6(-276.5)	-285.1(-278.8)	-273.6(-269.4)	-276.9(-271.4)	-283.5(-275.3)	-279.8(-274.9)	-270.8(-265.4)	-298.3(-298.1)
3d0j	127	-234.5(-224.3)	-242.4(-233.1)	-247.2(-240.6)	-193.5(-187.2)	-238.2(-226.1)	-251.5(-242.1)	-244.9(-238.9)	-196.5(-184.9)	-268.5(-267.5)
2chf	128	-268.2(-260.0)	-279.7(-273.8)	-287.8(-285.3)	-272.5(-266.3)	-261.3(-256.1)	-275.6(-267.8)	-275.4(-270.0)	-260.8(-256.2)	-303.2(-302.9)
2j6a	135	-230.1(-222.4)	-240.1(-230.4)	-236.6(-232.6)	-203.1(-200.8)	-233.1(-225.3)	-242.2(-232.0)	-253.7(-242.9)	-208.1(-200.2)	-258.6(-258.3)
3dnx	139	-273.3(-259.0)	-275.9(-266.0)	-290.0(-279.7)	-241.6(-233.9)	-265.6(-258.2)	-278.5(-273.7)	-284.4(-279.4)	-236.9(-231.9)	-307.2(-306.9)

Table 4.3. The 36 proteins shown by PDB code and chain. The next 4 columns give the lowest energy found for the homolog move set. The 4 columns on the right side of the table give the lowest energy found for the homolog free move set. The average energy of the top 1% of samples is shown in parenthesis.

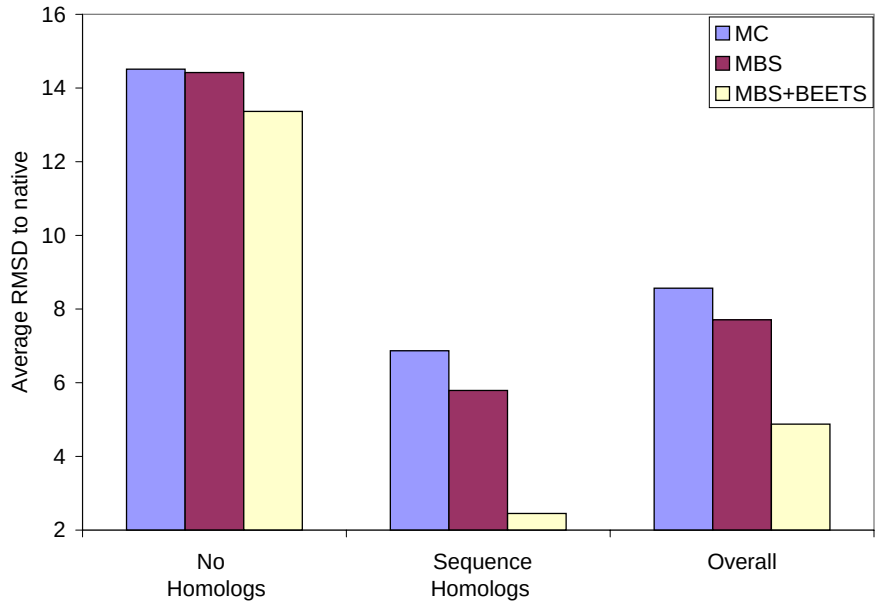


(a) Homology included data set, lower is better

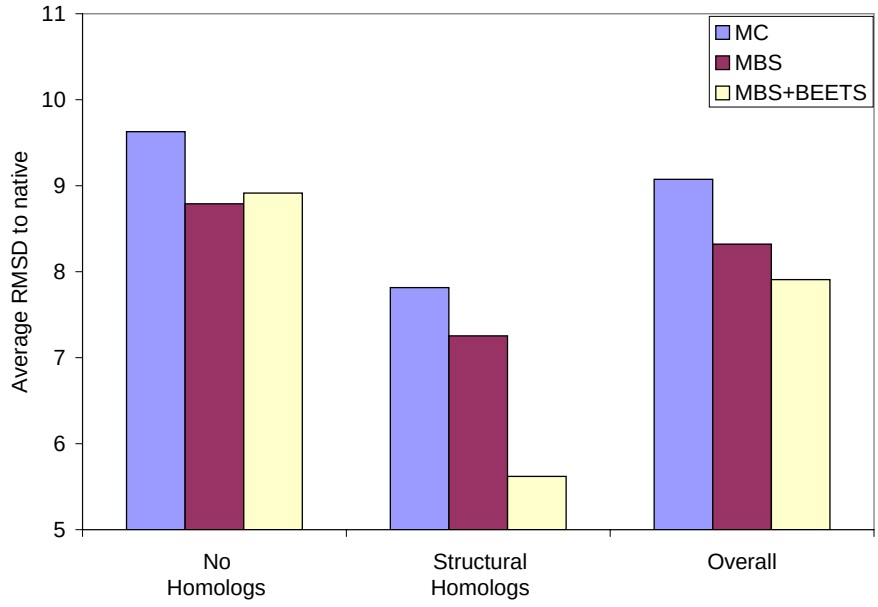


(b) Homology excluded data set, lower is better

Figure 4.19. Energy improvement of BEETS over MBS and MC. The decrease in energy in all categories attributed to BEETS indicates that BEETS effectively uses information from the substructure of known proteins independent of the presence of homologs.

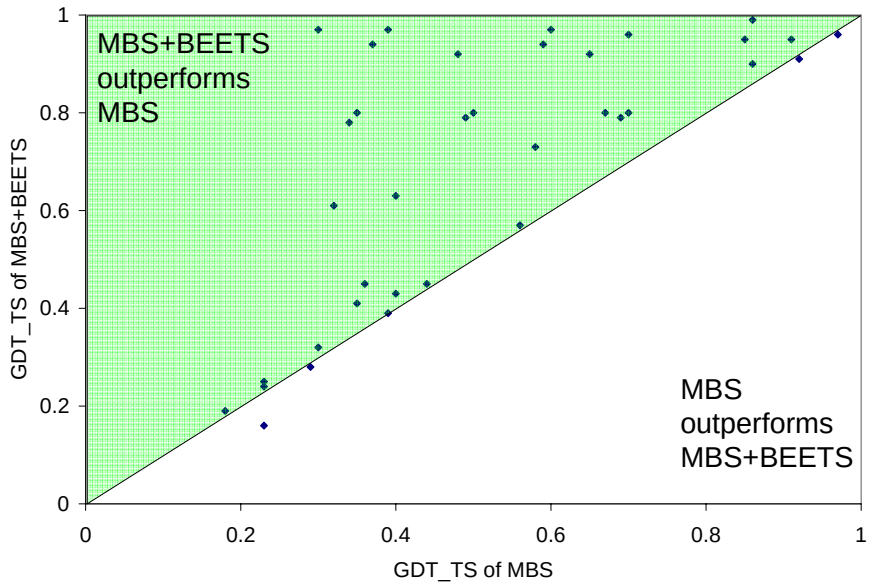


(a) Homology included data set, lower is better

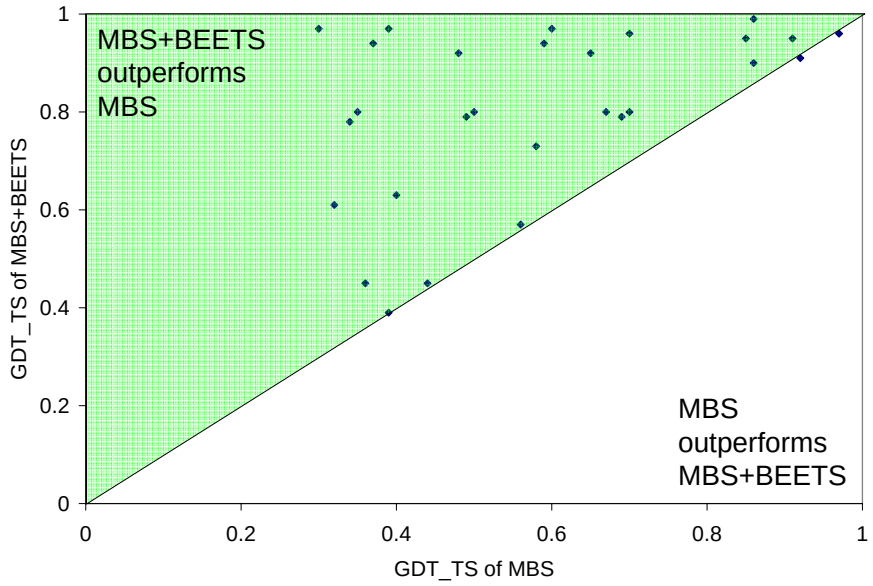


(b) Homology excluded data set, lower is better

Figure 4.20. RMSD Improvement by BEETS over MBS and MC. The decrease in RMSD by MBS+BEETS is most significant only when homologs are present. This indicates that homologs are crucial to identifying the native state.

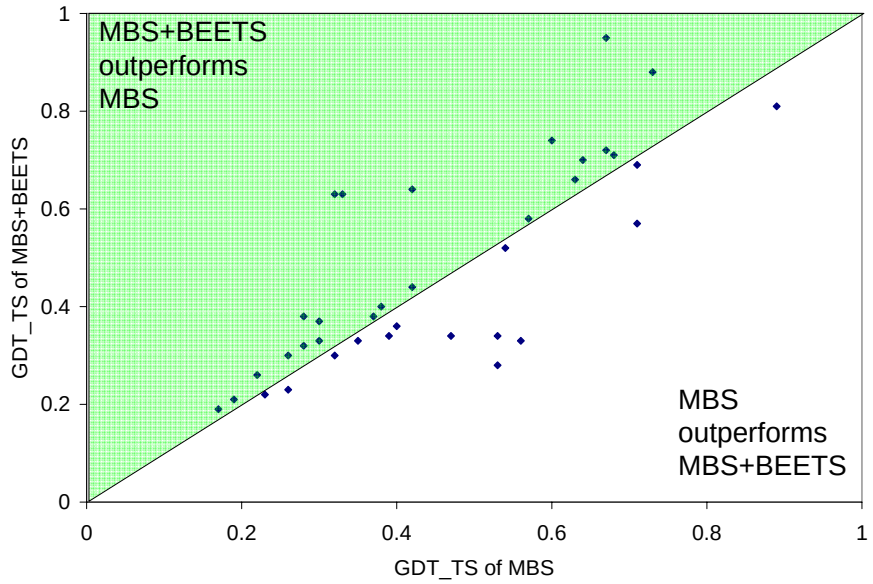


(a) Overall

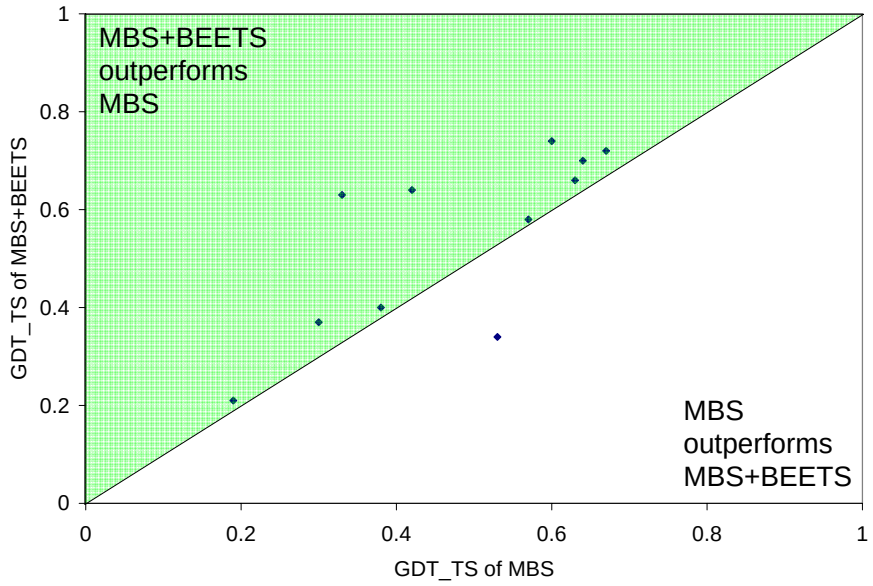


(b) Targets with sequence homologs

Figure 4.21. Change in performance measured by GDT_TS for when search uses sequence homologs (data set 1). Every dot represents an individual protein. The goal is for all proteins to have a GDT_TS of 1. It is noteworthy that MBS+BEETS outperformed MBS on 35 of 36 proteins.



(a) Overall



(b) Proteins with structural homologs

Figure 4.22. Change in performance measured by GDT when search does not use structural homologs (data set 2). Every dot represents an individual protein. The goal is for all proteins to have a GDT_TS of 1. It is noteworthy that MBS+BEETS outperforms MBS on most cases when a structural homolog exists.

CHAPTER 5

CONCLUSIONS

Studying the role information plays in search is crucial as new information is exploding all around us. Computers have made it very easy to categorize and collect data, leading to a vast repository of information for most fields. Consider for example the internet: in 2000 Google had indexed 1 billion web pages, in 2004 it was 8 billion, and by 2008 that number ballooned to 1 trillion. Similarly, the protein data bank has shown exponential growth having doubled in size since 2000 and projected to triple to 150,000 structures by 2014. Now that all this information is available, it is time to start using it.

Search is a valuable area to start researching as it is commonly used to study scientific problems. In almost all cases researchers are using a general purpose search method like Monte Carlo and are not satisfied with the results. Most of these researchers think their results will improve with additional computational resources. Unfortunately, for most problems, the size of conformation space is too large to be addressed even with increased computational resources.

Instead of increasing computational resources, my thesis shows that the use of information to choose appropriate regions, i.e. regions that are highly likely to contain the solution, is key to effective search. The more accurate the information utilized, the better the results. To identify accurate information search must be specialized for each domain. The importance of specialization is supported by others who have shown that a general-purpose optimization strategy is impossible [73]. Fortunately,

as I pointed out earlier, the availability of information is increasing at a phenomenal pace.

When information directs search towards the wrong regions resources are wasted, and search may fail. There is a vast amount of information available and not all of it will be useful, so search needs ways to differentiate between different qualities of information. This thesis introduces adaptive balancing of exploitation with exploration as an essential component in search, providing a framework that can use an information source where information quality varies. The first step to balancing exploitation with exploration is assessment of information quality. Since information is different for each domain, assessment of information quality needs to be domain specific. When high quality information is available, the information should be strongly exploited and used to guide search toward the global minimum. With lower quality information, there is a higher likelihood of that search will be guided into local minimum, and as a result exploration must be increased to recover from the mistake.

In this thesis I apply adaptive balancing of exploitation with exploration on protein structure prediction. Robust, high-throughput, high-resolution analysis of protein structure is one of the most important challenges in molecular biology. Addressing this challenge would help researchers determine protein function, cure diseases, and design novel proteins. To address this challenge, the principles of adaptive balancing are applied to two information sources for proteins: the all-atom energy function and protein data bank information.

To effectively use information from the all-atom energy function this thesis first presents model-based search. Model-based search is a new conformation space search method for finding minima in protein energy landscapes. Model-based search combines highly effective conformation space search with the ability to perform search using accurate all-atom energy information. The improvements afforded by my approach are based on two main contributions. First, my method is more effective

than previous methods at identifying and selecting the appropriate regions to focus resources. Second, enabled by the first contribution, my method is able to obtain high-quality all-atom information without incurring a significant performance penalty.

I have also developed a much more effective method of leveraging information from the protein data bank. Effective leveraging of template proteins is made possible by two contributions. First, my approach can accurately assess the quality of a template protein. This improvement is due to my hypothesis that the presence of a matching substructure in two proteins makes it more likely that residues neighboring the match also have similar structures. Enabled by effective quality assessment of template proteins my second contribution is a method to adaptively balance exploitation and exploration, allowing the template protein to be used only to the extent that current information suggests that it is accurate.

Model based search and balanced exploitation and exploration for template proteins have profound implications for protein structure prediction. To my knowledge my thesis shows the first example of structurally related, but not sequence identifiable, homologs being used to guide search. All previous approaches are limited to using sequence based homologs. With research showing that structure is 3-10 times more conserved than sequence [22], the use of structural homologs and other structurally derived information sources should significantly improve protein structure prediction.

I try to determine the value of information sources throughout this dissertation, however, I never discuss information theory [63]. It should be possible to quantify the information available from each source in a more computationally refined way. Difficulties in quantifying information include the the high variability in homolog quality, and the sparseness of protein configuration space. Work done to quantify the value of information should improve the results in this thesis. However, I believe concentrating efforts on identifying additional information sources will more quickly improve search.

While this thesis has improved search and protein structure prediction, numerous problems remain. These problems include how to model protein complexes, ligands, domain swaps, and large proteins. To solve each of these problems will require new strategies that each push the field a tiny bit further. Even with all these future advancements, inaccurate predictions will persist because each method will fail on some rare proteins. Because of these issues, I predict computational protein structure prediction will never predict structures adequate for biologists and eventually will be superseded by a method that combines computational structure prediction with high-throughput experimental methods.

If the future is a combined computational and experimental method, why does nearly everyone focus on purely computational protein structure prediction? I believe the culprit is CASP. In CASP only the amino acid sequence is distributed though easily acquired experimental data could be distributed. Some examples of useful experimental data would be presence of ligands, unassigned NMR data, or unphased crystallographic data.

Even with experimental data the primary bottleneck for larger proteins is still conformational sampling. For example, the state of the art in NMR modeling combines search and the Rosetta energy function to assign NMR restraints. With this methodology, accurate structures can be calculated for proteins as large as 20 kDa with limited experimental data [57].

In many cases improvements to computational structure prediction also improve structure determination with experimental data. To make these improvements researchers should concentrate their efforts on conformational search and energy function improvements, while not wasting resources to develop methods on things that are easy experimentally.

To improve conformational space search, I believe researchers should concentrate on multiple sequence alignments, and understanding the protein folding process. A

multiple sequence alignment contains information such as what residues are most unlikely to mutate, thereby indicating which residues are most important. Search strategies should treat these important residues differently. Understanding the protein folding process could bear fruit if proteins fold in a characteristic way that can be abstracted into a new search strategy that examines far fewer states.

To improve the energy function, researchers should figure out what is wrong and fix it! Until that happens, an interesting area of research would be to determine when to trust the energy function and when to trust homology information. I hypothesize that homology derived restraints will be more accurate in conserved loops and active sites, while the energy function will be more accurate in the core of the protein. The score function would be improved by combining the most accurate features of the energy function with homology derived restraints.

Once methods exist to accurately determine protein structure, these methods can be put to use helping cure disease and improving the efficiency of important industrial and pharmaceutical reactions. Only recently have structure calculation methods reached the tipping point where protein design is feasible [24, 60]. Although this represents a huge success, designed enzymes are always less catalytically active and smaller than enzymes produced in nature. Design of larger, more catalytically efficient enzymes will require continued algorithmic improvement. As design of enzymes improves, industry will rely more on highly effective enzymatic catalysis and less on potentially toxic chemical catalysis.

The primary bottleneck to both structure calculation and enzyme design is conformation space sampling. New algorithms developed to tackle this problem are interesting in their own right as they combine ideas from global optimization, artificial intelligence, high performance computing and biological architecture. Progress in conformation space sampling would have immediate implications to the design of novel therapeutics and industrially useful enzymes. The ability to make this progress

depends most on what information is used to decide where to explore. In protein modeling and protein design there exists a vast amount of information from experimentally determined protein structures, multiple sequence alignments, NMR, CryoEM and X-ray crystallography. Current approaches only use a small fraction of the relevant information. Since information is critical to search performance, the ability to identify and use additional information will advance many problems in computational biology. My plan is to identify and use new information sources to improve search for both protein structure prediction and protein design.

BIBLIOGRAPHY

[1] A.H. Mantawy, Y.L. Abdel-Magid, S.Z. Selim. Integrating genetic algorithms, tabu search, and simulated annealing for the unit commitment problem. *IEEE Transactions on Power Systems 14*, 3 (1999), 829–836.

[2] Aloy, Patrick, Stark, Alexander, and adn Robert B. Russell, Caroline Hadley. Predictions without templates: New folds, secondary structure, and contacts in CASP5. *Proteins: Structure, Function, and Bioinformatics 53*, S6 (2003), 436–456. `http://www3.interscience.wiley.com/cgi-bin/fulltext/106559008/PDFSTART`.

[3] Battiti, Roberto. Reactive search: Toward self-tuning heuristics. In *Modern Heuristic Search Methods* (Chichester, 1996), V. J. Rayward-Smith, I. H. Osman, C. R. Reeves, and G. D. Smith, Eds., John Wiley & Sons Ltd., pp. 61–83.

[4] Blum, Benjamin Norman. Resampling methods for protein structure prediction. Dissertation UCB/EECS-2008-184, Stanford, 2008.

[5] Bonet, Jeremy S. De, Isbell, Jr., Charles L., and Viola, Paul. MIMIC:fining optima by estimating probability densities. In *Advances in Neural Information Processing Systems*, Michael C. Mozer, Michael I. Jordan, and Thomas Petsche, Eds., vol. 9. MIT Press, 1997, p. 424.

[6] Bonneau, Richard, Strauss, Charlie E. M., Rohl, Carol A., Chivian, Dylan, Bradley, Phillip, Malmström, Lars, Robertson, Tim, and Baker, David. De novo prediction of three-dimensional structures for major protein families. *Journal of Molecular Biology 322*, 1 (2002), 65–78.

[7] Boyan, Justin Andrew. Learning evaluation functions for global optimization. Tech. Rep. CMU-CS-98-152, School of Computer Science, Carnegie Mellon University, 1998.

[8] Bradley, Philip, Chivian, Dylan, Meiler, Jens, Misura, Kira M. S., Rohl, Carol A., Schief, Willam R., Wedemeyer, William J., Schueler-Furmann, Ora, Murphy, Paul, Schonbrun, Jack, Strauss, Charles E. M., and Baker, David. Rosetta predictions in CASP5: Successes, failures and prospects for complete automation. *Proteins: Structure, Function, and Bioinformatics 53*, S6 (2003), 457–468.

[9] Bradley, Philip, Misura, Kira M. S., and Baker, David. Toward high-resolution de novo structure prediction for small proteins. *Science 309*, 5742 (2005), 1868–1971.

[10] Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., and Karplus, M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry 4* (1983), 187–217.

[11] Bystroff, Christopher, and Baker, David. Prediction of local stucture in proteins using a library of sequence-structure motifs. *Journal of Molecular Biology 281*, 3 (1998), 565–577.

[12] Chivian, Dylan, Kim, David E., Malmström, Lars, Schonbrun, Jack, Rohl, Carol, and Baker, David. Prediction of CASP-6 structures using automated robetta protocols. *Proteins: Structure, Function, and Bioinformatics 61*, S7 (2005), 157–166.

[13] Cohn, David A., Ghahramani, Zoubin, and Jordan, Michael I. Active learning with statistical methods. *Journal of Artificial Intelligence Research 4* (1996), 129–145.

[14] Cortés, J., Siméon, T., Remaud-Siméon, M., and Tran, V. Geometric algorithms for the conformational analysis of long protein loops. *Journal of Computational Chemistry 25*, 7 (May 2004), 956–967.

[15] Fernandez-Fuentes, Narcis, and Fiser, Andras. What does make a fold new. CASP 8 presentation, December 2008. `http://predictioncenter.org/casp8/doc/presentations/Fiser_newfold.pdf`.

[16] Frantz, D. D., Freeman, D. L., and Doll, J. D. Reducing quasi-ergodic behavior in Monte Carlo simulations by j-walking: Applications to atomic clusters. *Journal of Chemical Physics 93*, 4 (1990), 2769–2784.

[17] Glover, Fred, and Laguna, Fred. *Tabu Search*. Kluwer Academic Publisher, 1997.

[18] Hardin, Corey, Pogorelov, Taras V., and Luthey-Schulten, Zaida. Ab initio protein structure prediction. *Current Opinion in Structural Biology 12*, 2 (2002), 176–181.

[19] Hart, Reece K., Pappu, Rohit V., and Ponder, Jay. Exploring the similarities between potential smoothing and simulated annealing. *Journal of Computational Chemistry 21*, 7 (January 2000), 531–552.

[20] Holland, John H. *Adaptation in Natural and Artificial Systems*. The University of Michigan Press, Ann Arbor, USA, 1975.

[21] Hung, Ling-Hong, Ngan, Shing-Chung, and Samudrala, Ram. De novo protein structure prediction. In *Computational Methods for Protein Structure Prediction and Modeling 2*, Ying Xu, Dong Xu, and Jie Liang, Eds. Springer Verlag, 2007, pp. 43–64.

[22] Illergard, Kristoffer, Ardell, David H., and Elofsson, Arne. Structure is three to ten times more conserved than sequence - a study of structural response in protein cores. *Proteins: Structure, Function, and Bioinformatics 77*, 3 (2009), 499–508.

[23] Jain, A.K., Murty, M.N., and Flynn, P.J. Data clustering: A review. *ACM Computing Surveys 31*, 3 (1999), 264–323.

[24] Jiang, Lin, Althoff, Eric, Clemente, Fernando, Doyle, Lindsey, Röthlisberger, Daniela, Zanghellini, Alexandre, Gallaher, Jasmine, Bekter, Jamie, Tanaka, Fujie, asnd Donald Hilvert, Carlos Barbas, Houk, Kendal, Stoddard, Barry, and Baker, David. De novo computational design of retro-aldol enzymes. *Science 319*, 5868 (March 2008), 1387–1391.

[25] Jones, D. T., and Thornton, J. M. Potential energy functions for threading. *Current Opinion in Structural Biology 6* (1996), 210–216.

[26] Jones, David T., and McGuffin, Liam J. Assembling novel protein folds from super-secondary structural fragments. *Proteins: Structure, Function, and Genetics 53*, S6 (2003), 480–485.

[27] Kabsch, Wolfgang, and Sander, Christian. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers 22*, 12 (December 1983), 2577–637. dssp.

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

[29] Kryshtafovych, Andriy, Venclovas, Česlovas, Fidelis, Krzysztof, and Moult, John. Progress over the first decade of casp experiments. *Proteins: Structure, Function, and Bioinformatics 61*, S7 (2005), 225–236.

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

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

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

[33] Levitt, Michael. Nature of the protein universe. *Proceedings of the National Academy of Sciences 106*, 27 (2009), 11079–11084.

[34] Li, Zhenqin, and Scheraga, Harold A. Monte carlo-minimization approach to the multiple-minima problem in protein folding. *Proceedings of the National Academy of Sciences 84*, 19 (October 1987), 6611–6615.

[35] Lotan, Itay, van den Bedem, Henry, Deacon, Ashley M., and Latombe, Jean-Claude. Computing protein structures from electron density maps: The missing loop problem. In *Proceedings of the Workshop on the Algorithmic Foundations of Robotics (WAFR)* (2004).

[36] Lyubartsev, A. P., Martsinovski, A. A., Shevkunov, S. V., and Vorontsov-Velyaminov, P. N. New approach to monte carlo calculation of the free energy: Method of expanded ensembles. *Journal of Chemical Physics 96*, 3 (1992), 1776–1783.

[37] MacKay, David. Information-based objective functions for active data selection. *Neural Computation 4*, 4 (1992), 590–604.

[38] McGuffin, Liam J., Bryson, Kevin, and Jones, David T. The PSIPRED protein structure prediction server. *Bioinformatics 16*, 4 (April 2000), 404–405.

[39] Metropolis, N., and Ulam, S. The monte carlo method. *Journal of the American Statistical Association 44*, 247 (1949), 335–341.

[40] Montelione, Gaetano T., and Anderson, Stephen. Structural genomics: keystone for a Human Proteome Project. *Nature Structual Biology 6*, 1 (1999), 11–12.

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

[42] Moult, John, Fidelis, Krzysztof, Zemla, Adam, and Hubbard, Tim. Critical assessment of methods of protein structure prediction (CASP): round IV. *Proteins: Structure, Function, and Bioinformatics 45*, Suppl. 5 (2001), 2–7.

[43] Moult, John, Fidelis, Krzysztof, Zemla, Adam, and Hubbard, Tim. Critical assessment of methods of protein structure prediction (CASP)–round V. *Proteins: Structure, Function, and Genetics 53*, Suppl. 6 (2003), 334–339.

[44] Moult, John, Hubbard, Tim, Fidelis, Krzysztof, and Pedersen, J. T. Critical assessment of methods of protein structure prediction (CASP): round III. *Proteins: Structure, Function, and Genetics 37*, Suppl. 3 (1999), 2–6.

[45] Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology 247* (1995), 636–540.

[46] NIGMS/NIH Protein Structure Initiative (PSI). PSI pilot phase fact sheet. http://www.nigms.nih.gov/Initiatives/PSI/Background/PilotFacts.htm, October 2009.

[47] Okamoto, Yuko. Protein folding problem as studied by new simulation algorithms. *Recent Research Developments in Pure & Applied Chemistry 1* (1998).

[48] Orengo, C. A., Michie, A.D., Jones, D.T., Swindells, M.B., and Thornton, J.M. CATH: A hierarchic classification of protein domain structures. *Structure 5* (1997), 1093–1108.

[49] Paluszewski, Martin, Hamelryck, Thomas, and Winter, Pawel. Reconstructing protein structure from solvent exposure using tabu search. *Algorithms for Molecular Biology 1* (October 2006), 1–14.

[50] Pappu, Rohit V., Hart, Reece K., and Ponder, Jay W. Analysis and application of potential energy smoothing and search methods for global optimization. *Journal of Physical Chemistry B 102* (1998), 9725–9742.

[51] Pearlman, D. A., Case, D. A., Caldwell, J. W., Ross, W. R., Cheatham and III, T. E., DeBolt, S, Ferguson, D., Seibel, G., and Kollman, P. AMBER, a computer program for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to elucidate the structures and energies of molecules. *Computational Physics Communcations 91* (1995), 1–41.

[52] Peila, Jucjan, Kostrowicki, Jaoslaw, , and Scheraga, Harold A. The multiple-minima problem in the conformational analysis of molecules . deformation of the potential energy hypersurface by the diffusion equation method. *Journal of Physical Chemistry B 93* (1989).

[53] Petrey, Donald, and Honig, Barry. Protein structure prediction: Inroads to biology. *Molecular Cell 20* (December 2005), 811–819.

[54] Pillardy, Jaroslaw, Czaplewski, Cezary, Wedemeyer, William J., and Scheraga, Harold A. Conformation-family monte carlo (CFMC): An efficient computational method for identifying the low-energy states of a macromolecule. *Helvetica Chimica Acta 83* (2000), 2214–2230.

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

[56] Qian, Bin, Raman, Srivatsan, Rhiju Das and, Philip Bradley, McCoy, Airlie, Read, Randy, and Baker, David. High-resolution structure prediction and the crystallographic phase problem. *Nature 450* (November 2007), 259–264.

[57] Raman, Srivatsan, Lange, Oliver, Rossi, Paolo, Tyka, Michael, Wang, Xu, Liu, James Aramini Gaohua, Ramelot, Theresa, Eletsky, Alexander, Szyperski, Thomas, Kennedy, Michael, Prestegard, James, Montelione, Gaetano, and Baker, David. Nmr structures determination for larger proteins using backbone-only data. *Science 327*, 5968 (February 2010), 1014–1018.

[58] Rhee, Young Min, and Pande, Vijay S. Multiplexed-replica exchange molecular dynamics method for protein folding prediction. *Biophysical Journal 84* (2003), 775–786.

[59] Rohl, Carol A., Strauss, Charlie E. M., Misura, Kria M. S., and Baker, David. Protein structure prediction using Rosetta. *Methods in Enzymology 383* (2004), 66–93.

[60] Röthlisberger, Daniela, Khersonsky, Olga, Wollacott, Andrew, Jiang, Lin, DeChnancie, Jason, Betker, Jamie, Gallaher, Jasmine, Althoff, Eric, Zanghellini, Alexandere, Dym, Orly, Albeck, Shira, Houk, Kendall, Tafik, Dan, and Baker, David. Kemp elimation catalysts by computational enzyme design. *Nature 453* (May 2008), 190–195.

[61] Russell, Stuart, and Norvig, Peter. *Artificial Intelligence: A Modern Approach*, 2nd ed. Prentice Hall, 2002.

[62] Schueler-Furman, Ora, Wang, Chu, Bradley, Phil, Misura, Kira, and Baker, David. Progress in modeling of protein structures and interactions. *Science 310*, 5748 (October 2005), 638–642.

[63] Shannon, Claude E. A mathemtatical theory of communication. *Bell System Technical Journal 27* (July 1948), 379–423.

[64] Simons, Kim T., Riczinski, I., Kooperberg, C., Fox, B., Bystroff, C., and Baker, D. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-indepenedent features of proteins. *Proteins 34*, 1 (1999), 82–95.

[65] Skolnick, Jeffrey, Zhange, Yang, Arakaki, Adrian K., Kolinski, Andrezej, Boniecki, Michal, Szilágyi, András, and Kihara, Daisuke. TOUCHSTONE: A unified appoach to protein structure prediction. *Proteins: Structure, Function, and Genetics 53*, S6 (2003), 469–479.

[66] Söding, Johannes. Protein homology detection by hmm-hmm comparison. *Bioinformatics 21*, 7 (2005), 951–960.

[67] Stützle, Thomas, and Hoos, Holger H. Max-min ant system. *Future Generation Computer Systems 16*, 8 (2000), 889–914.

[68] Swendsen, Robert H., and Wang, Jian-Sheng. Replica Monte Carlo simulation of spin-glasses. *Physical Review Letters 57*, 21 (1986), 2607–2609.

[69] Ulrich H. E. Hansmann, Luc T. Wille. Global optimization by energy landscape paving. *PRL 88*, 6 (2002), 068105.

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

[71] Venclovas, Česlovas, Zemla, Adam, Fidelis, Krzysztof, and Mould, John. Assessment of progress over the casp experiments. *Proteins: Structure, Function, and Bioinformatics 53*, Suppl. 6 (2003), 585–595.

[72] Wales, David, and Doye, Jonathan. Global optimization by basin-hopping and the lowest energy structures of lennard-jones clusters containing up to 110 atoms. *Journal of Physical Chemistry A 101*, 28 (June 1997), 5111–5116.

[73] Wolpert, David H., and Macready, William G. No free lunch theorems for search. Tech. Rep. SFI-TR-95-02-010, Santa Fe Institute, Santa Fe, USA, 1995.

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

[75] Zemla, Adam. LGA: A method for finding 3D similarities in protein structure. *Nucleic Acids Research 31*, 13 (2003), 3370–3374.

[76] Zhang, Yang, Arakaki, Adrian K., and Skolnick, Jeffrey. TASSER: An automated method for the prediction of protein tertiary structure in CASP6. *Proteins: Structure, Function, and Bioinformatics Suppl. 7* (2005), 91–98.

[77] Zhang, Yang, Kihara, Daisuke, and Skolnick, Jeffrey. Local energy landscape flattening: Parallel hyperbolic Monte Carlo sampling of protein folding. *Proteins: Structure, Function, and Genetics 48*, 2 (2002), 192–201.

[78] Zhang, Yang, Kolinski, Adrezej, and Skolnick, Jeffrey. TOUCHSTONE II: A new approach to ab initio protein structure prediction. *Biophysical Journal 85* (2003), 1145–1164.

[79] Zhang, Yang, and Skolnick, Jeffrey. SPICKER: A clustering approach to identify near-native protein folds. *Journal of Computational Chemistry 25*, 6 (April 2004), 865–871.

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

[81] Zhou, Hongyi, and Skolnick, Jeffrey. Ab initio protein structure prediction using chunk-tasser. *Biophysical Journal 93*, 5 (September 2007), 1510–1518.