Improving protein structure prediction with model-based search

T J Brunette and Oliver Brock*

Bioinformatics Research Laboratory, Department of Computer Science, University of Massachusetts, Amherst, MA 01003-9264, USA

ABSTRACT

Motivation: *De novo* protein structure prediction can be formulated as search in a high-dimensional space. One of the most frequently used computational tools to solve such search problems is the Monte Carlo method. We present a novel search technique, called model-based search. This method samples the high-dimensional search space to build an approximate model of the underlying function. This model is incrementally refined in areas of interest, whereas areas that are not of interest are excluded from further exploration. Model-based search derives its efficiency from the fact that the information obtained during the exploration of the search space is used to guide further exploration. In contrast, Monte Carlo-based techniques lack memory and exploration is performed based on random walks, ignoring the information obtained in previous steps.

Results: Model-based search is applied to protein structure prediction, where search is employed to find the global minimum of the protein's energy landscape. We show that model-based search uses computational resources more efficiently to find lower-energy conformations of proteins than one of the leading protein structure prediction methods, which relies on a tailored Monte Carlo method to perform a search. The performance improvements become more pronounced as the dimensionality of the search problem increases. We argue that model-based search will enable more accurate protein structure prediction than was previously possible. Furthermore, we believe that similar performance improvements can be expected in other problems that are currently solved using Monte Carlo-based search methods.

Availability: An implementation of model-based search can be obtained by contacting the authors.

Contact: oli@cs.umass.edu

1 INTRODUCTION

In computational biology, many problems can be formulated as a search in a high-dimensional space. *Ab initio* or *de novo*

protein structure prediction, for example, represents the conformation of a protein by a set of parameters and then search an energy function defined over these parameters for a global minimum. The accuracy of such structure prediction depends on the accuracy of the energy function and on the effectiveness of the method used to search the energy landscape defined by that function. If the energy function accurately captures the energetics of proteins, its minimum is believed to correspond the native structure of the folded protein. In this paper, we assume to be provided with an accurate energy function and focus on the problem of searching this function for a global minimum.

For almost 60 years, one of the most universally applied search methods to determine extrema in high-dimensional spaces has been the Monte Carlo method (Metropolis and Ulam, 1949). It is applicable under very general conditions and yields a computationally efficient search, even in high-dimensional spaces.

We present a novel search method, called *model-based search*, which is equally applicable under general conditions as Monte Carlo methods but outperforms Monte Carlo-based techniques by a large margin, both computationally and in terms of the quality of the extremum obtained by the search. Similarly to Monte Carlo methods, it determines the extrema of functions through sampling. But in contrast to Monte Carlo methods, model-based search improves the effectiveness and efficiency of search by exploiting information obtained during the search. This information can provide insights into quantitative and qualitative aspects of the searched function, which can be used to guide further exploration. Monte Carlo-based methods, in contrast, use only the current state of the search to affect continued exploration and ignore the information obtained previously. Monte Carlo-based methods are memory-less, whereas model-based sampling incrementally builds a approximate model of the searched function. In regions likely to contain significant local extrema, this model becomes increasingly accurate, whereas regions unlikely to be of interest are excluded from further exploration.

We apply model-based search to the problem of protein structure prediction, which can be viewed as the search of the

---

*To whom correspondence should be addressed.

global minimum in an energy landscape defined over the conformational space of proteins (Anfinsen, 1973). We compare the performance of model-based search with that of a prominent structure prediction method Rosetta (Rohl *et al.*, 2004), Which uses a highly tailored Monte Carlo method to perform search. Our experiments show that model-based search is much more effective at finding low-energy conformations, then the Monte Carlo search performed by Rosetta. These improvements become more pronounced as the dimensionality of the search space increases, indicating that model-based search can outperform Monte Carlo-based search methods, even in very high-dimensional search spaces.

Although we present an experimental validation of model-based search only in the area of protein structure prediction, we believe that it represents a general and effective approach to search, well suited to a variety of search problems encountered in computational biology and other domains, such as chemistry, physics and economics. More specifically, we believe that model-based search can provide significant performance improvements in all problems to which Monte Carlo methods are currently applied.

2 RELATED WORK

Computational approaches to protein structure prediction use search to determine the minimum of an energy function, with the notable exception of comparative modeling (Leach, 1991), which performs structure prediction based on homologous structures found in protein structure databases. All other computational structure prediction approaches search the energy landscape associated with the conformational space of a protein for a global minimum. This minimum is believed to correspond to the native structure of the protein. Owing to the tremendous size of the search space (Levinthal, 1968), most methods search by following the gradient of the energy landscape. The most commonly applied computational paradigm to perform such gradient descent is the Monte Carlo method. It performs a random walk of the energy landscape, accepting only new states that are lower in energy than the current one.

To escape the numerous local minima encountered during this gradient descent, the Metropolis criterion (Metropolis and Ulam, 1949) is employed, accepting increases in energy with a probability inversely proportional to the increase in energy. The Metropolis Monte Carlo algorithm is based on the Boltzmann equation: increasing the temperature variable in the Boltzmann equation increases the probability of accepting neighboring states with higher energy.

The insight that controlling temperature can lead to better convergence has resulted in a variety of modifications of the Monte Carlo method. High temperatures are well suited to the coarse exploration of large regions of conformation space, whereas lower temperatures search smaller regions in more detail but are less likely to escape significant local minima. The technique of simulated annealing (Okamoto, 1998; Kirkpatrick *et al.*, 1983), for example, slowly lowers the temperature, resulting in a more detailed search focused on exploration of low-energy local minima. Jump walking (Frantz *et al.*, 1990) jumps between high and low temperatures, thereby conducting a detailed search in multiple low-energy regions. Other methods modify Monte Carlo by directly controlling the probability with which states at different energy levels are sampled (multi-canonical ensemble method) (Berg and Neuhaus, 1992; Lee, 1993) or by combining several of these heuristics (Xu and Berne, 1999). One of the most interesting techniques of exploiting temperature variation is parallel tempering. Parallel tempering performs multiple Monte Carlo runs in parallel. These runs are performed at varying temperature, thereby allowing conformation space to examined at multiple scales concurrently (Hansmann, 1997; Zhang and Skolnick, 2001).

There have also been improvements to search that are specific to the domain of protein structure prediction. The insight that amino acid sequences with a high degree of homology often fold into similar 3D structures (Jones and Thornton, 1996; Leach, 1991) has been used very successfully to reduce the overall search space and consequently to increase the efficiency of search. This is accomplished by dividing a protein into smaller fragments and finding numerous homologous candidate structures for these fragments. A search is then performed over these candidate structure, effectively reducing the relevant conformation space and improving the efficiency of Monte Carlo methods when applied to this modified search problem (Bradley *et al.*, 2003).

Despite the numerous improvements that have been suggested to the basic Metropolis Monte Carlo method over the past 50 years, all of these methods remain susceptible to local minima and computational resources are wasted as a consequence. This susceptibility is caused by the fact that these approaches are memory-less, that is, they do not exploit information obtained about the search space to guide further search. This results in repeated exploration of similar conformations and in exploration of regions unlikely to contain a relevant extremum.

To reduce the detrimental effect of local minima, search methods have to maintain information about the ongoing search. A number of such search methods can be found in the artificial intelligence (AI) literature. For example, beam search maintains multiple search paths simultaneously and biases exploration toward the most promising paths. Beam search uses a constant-width frontier during exploration, adding states in a depth-first manner based upon a heuristic (Russell and Norvig, 2003). A major drawback of beam search is that, over times, multiple searches tend to converge to a single region of the search space and consequently only small regions of the search space are explored (Russell and Norvig, 2003).

Genetic algorithms (Holland, 1975) also track multiple paths by maintaining a population of states and applying the principles of evolution to this population. States in a new generation of the population are created from the previous generation by combining two parent states. Genetic algorithms, like beam search, do not actively analyze the conformation space and suffer from the drawback of focusing too much computation on local minima.

Meta-heuristical approaches from classical AI and active learning offer promising ways to improve the performance of search through actively acquiring information and using it to guide exploration.

Tabu search (Glover and Laguna, 1997), for example, in its simplest form designates undesirable regions of conformation space 'tabu', thereby biasing computation toward more relevant regions. To declare regions tabu, however, they have first to be explored. Also, the exclusion of specific regions does not provide a way of prioritizing the exploration of the regions not excluded from the search.

Methods from active learning (MacKay, 1992; Cohn *et al.*, 1996) are capable of providing such a prioritization. They allow the determination of those regions of the search space that should be sampled to make maximum progress toward understanding the underlying function. Recently, active learning approaches have proven very effective in the area of high-dimensional robotic motion planning (Burns and Brock, 2005), a problem that exhibits a significant resemblance to protein structure prediction.

3 MODEL-BASED SEARCH

Model-based search is a sampling-based search method for finding extrema in high-dimensional search spaces. It is motivated by the fact that in many applications the information obtained during a search can provide valuable insight into the relevance of regions with respect to the search task. Making only moderate assumptions about the continuity of the function we are searching, information obtained during the search can rule out the presence of the desired extremum in a particular region of the search space. As a consequence, additional exploration should be focused on other regions.

The general objective of model-based search is to exploit information obtained during the search to guide further search space exploration in the most effective manner. To accomplish this, a model is used to represent relevant information compactly. During search, the model is used to direct further exploration. The proposed method can be viewed as an active learning technique (Cohn *et al.*, 1996) that relies on insights from a meta-heuristic search (Glover and Laguna, 1997) to generate a highly efficient search.

Model-based search proceeds by interspersing exploration of the search space with incremental updates to a model that represents the information obtained so far. The model is used to guide exploration during each iteration of the search.
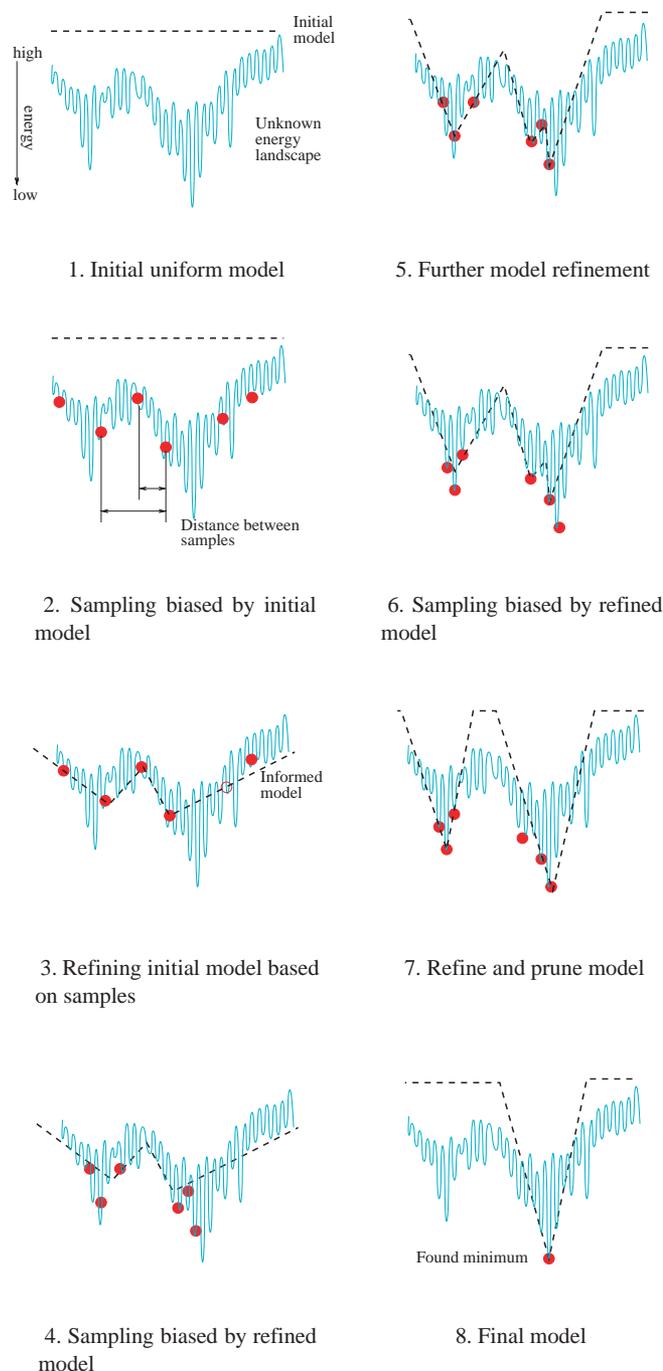

Fig. 1. Model-based search: iterative refinement of an approximate model to find a global minimum.

Figure 1 is a graphical illustration of the search for a global minimum in an arbitrary function using model-based search. In what follows we will provide additional details about each of the steps shown in that figure.

(1) Initially, the model (shown as a dashed line) contains no information about the function we are searching.

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

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

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

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

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

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

(8) The global minimum has been identified.

From this description of model-based search, it is apparent that the quality of the model critically determines the quality of the resulting search. Figure 2 illustrates that the quality of
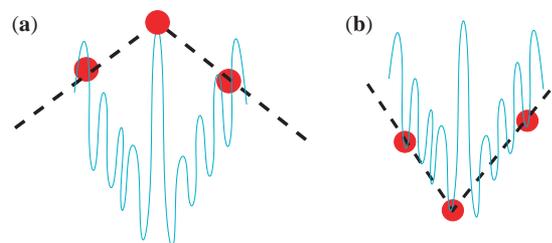
Fig. 2. Placing samples randomly increases the chance that a high-energy sample from a low-energy region will misinform the model. To avoid this, model-based search improves randomly placed samples using gradient descent. (**a**) Bad placement of samples. (**b**) Good placement of samples.

the model in turn depends on the placement of samples. In Figure 2a, samples are placed in such a way that the constructed model does not represent the underlying function well. This is addressed by including the application of gradient descent methods into the process of placing samples. After a sample has been placed randomly, gradient descent is used to find the closest local minimum. The local minimum is used to construct the model, rather than the original sample. This procedure significantly improves the accuracy of the model, as illustrated in Figure 2b.

4 PROTEIN STRUCTURE PREDICTION WITH MODEL-BASED SEARCH

In this section, we describe the application of model-based search to protein structure prediction. To perform protein structure prediction, we use model-based search to search for low-energy conformations in a protein's energy function. Since this research effort is focussed on efficient search techniques, we rely on a freely available energy function for proteins, namely, the energy function of the protein structure prediction software Rosetta (Rohl *et al.*, 2004).

The name Rosetta (Rohl *et al.*, 2004) refers to a large suite of software tools related to protein structure and protein folding (www.bakerlab.org). Here, we will refer only to the Rosetta protein structure prediction software. More specifically, we will be concerned only with *de novo* protein structure prediction, Over the past several years, Rosetta has repeatedly been demonstrated to outperform other computational methods of protein structure prediction in this category (Moult *et al.*, 2003).

To facilitate the description of implementation details for model-based search, we first describe the energy function and search method of Rosetta.

4.1 Rosetta's energy function

One of the fundamental building blocks of Rosetta is the energy function. Designed to approximate the true energy function of proteins, it computes an estimate of the energy for a particular protein conformation based on knowledge about

interactions between portions of the backbone, side chains and solvent. The design of this function was guided by observations from experimentally determined native structures (Rohl *et al.*, 2004). It has been improved continuously over the past several years and can be considered quite accurate, as shown by extensive experimental results (Moult *et al.*, 2003).

4.2 Rosetta's search method

Since we will compare model-based search with the Monte Carlo-based search strategy implemented by Rosetta, we briefly describe the latter here (Rohl *et al.*, 2004).

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

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

4.3 Implementation details

We now describe how the general search procedure from Section 3 can be adapted to the problem of protein structure prediction.

Rosetta's strategy of performing Monte Carlo steps based on protein fragments significantly reduces the search space and consequently renders search more efficient. To enable a fair comparison between Rosetta's Monte Carlo-based search strategy and model-based search, we apply this fragment-based search in both cases. Consequently, the way decoys (or samples in search space) are generated is identical in both approaches. Initially, the decoys are generated randomly. Additional samples are generated from existing samples by applying Rosetta's fragment replacement strategy.

At each stage of the search, samples are generated and their energy is determined using Rosetta's energy function. The adjacency relationship among samples is analyzed to compute and refine the model. This model represents regions that are still considered possibly to contain the global minimum. Each of these regions is represented by a sample of locally minimal energy and its nearest neighbors. Distance between sample is determined using a distance or similarity measure, such

as root mean square distance (RMSD) or global distance test total score (GDT_TS) (Zemla, 2003).

Samples with locally minimal energy relative to their neighbors are considered to be the bottom of a well. The distance to the nearest neighbors captures the size of this well.

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

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

5 EXPERIMENTAL RESULTS

The accuracy of protein structure prediction depends on the accuracy of the available energy function and on an effective search method. In our experimental evaluation we compare Rosetta's Monte Carlo-based search strategy with model-based search, both applied to Rosetta's energy function. The predictions from the implementation of model-based search described in Section 4 were compared with predictions obtained using Rosetta (Rohl *et al.*, 2004). We performed two sets of experiments: one based on 62 small proteins, ranging from 50 to 146 amino acids in length, and a second one, based on 9 proteins, varying from 60 to 414 amino acids in length.

For each protein structure prediction both algorithms produced 500 decoys (conformations of the protein considered to be candidates for the predicted native state). To generate these decoys, both algorithms were given an equal amount of computational resources. The results presented here show that model-based search consistently outperforms the search strategy implemented by Rosetta, producing decoys with much lower energy. An illustration of this performance improvement is given in Figure 3 for the protein with PDB code 2PTL. The improved prediction accuracy is apparent: the α-helix has the correct number of turns and the β-sheets are
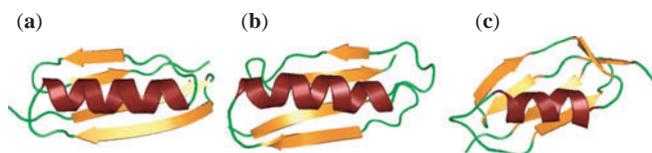
Fig. 3. The figure shows the lowest-energy decoys produced by model-based search and by Rosett's Monte Carlo-based method, alongside the experimentally determined native structure of the protein. (**a**) Model-based search, (**b**) experimentally determined native structure, and (**c**) Rosetta.
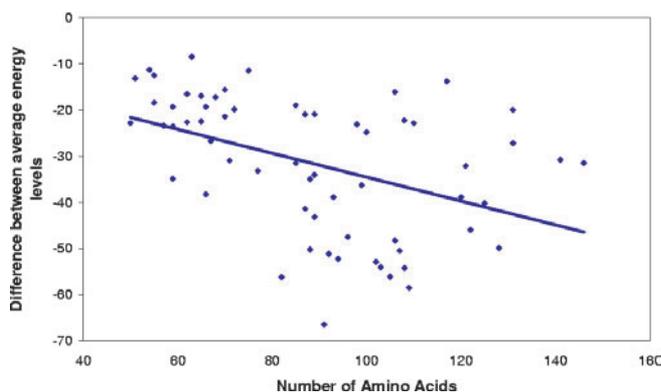

Fig. 4. In the first experiment, the structures of 62 small proteins, ranging in size from 50 to 146 amino acids, were predicted. The graph shows the difference in energy of structures obtained using model-based search and Rosetta's Monte Carlo-based search. The negative numbers indicate how much lower the energy of structures obtained with model-based search was compared with those obtained with Rosetta's search strategy. As can be seen by the trendline, model-based search finds significantly lower-energy structures with the same amount of computational resources. This performance improvement is amplified as the length of the protein increases.

predicted in the correct number and orientations. 2PTL is a 60 amino acid binding protein from the immunoglobulin L chain.

For longer proteins the performance advantage of model-based search with respect to finding lower-energy decoys becomes increasingly pronounced, as illustrated by the graphs in Figures 4 and 5. In addition, the variance of the resulting decoys is reduced significantly (Fig. 6). This is an indication that model-based search is searching the energy landscape more consistently and effectively than the Monte Carlo method used by Rosetta.

The improved accuracy of model-based search is also reflected in the fact that the distinct trajectories of the search correspond to structurally meaningful paths of the search. For example, the two distinct paths in the Hubbard plot (Hubbard, 1999) in Figure 7a for protein 2PTL (Fig. 3) correspond to structures in which the β-sheet is located on opposite sides of the α-helix. The two main clusters shown in Figure 6a represent the minima obtained along these two classes of paths.
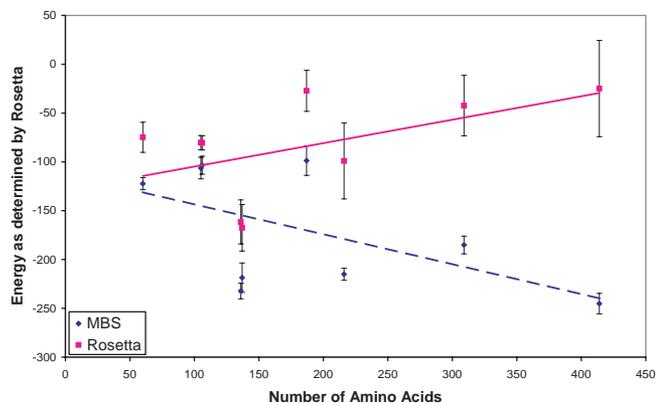

Fig. 5. In the second experiment the structure of nine proteins ranging in size from 60 to 414 amino acids were predicted. The graph shows the average energy of decoys produced by model-based search (MBS) and Rosetta for proteins of different sizes. Lower energy levels are desirable. It can be seen that model-based search outperforms the Monte Carlo method implemented by Rosetta by a large margin. The performance advantage increases as the length of the proteins increases. The trend lines indicate that this performance gap continues to widen for longer proteins.
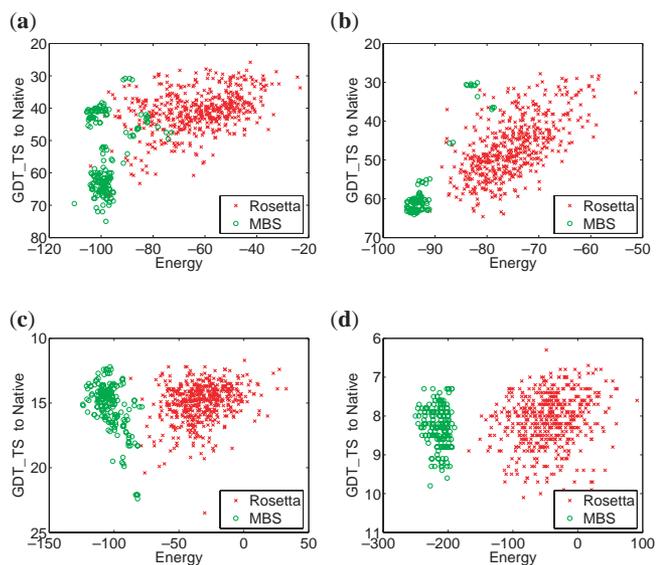

Fig. 6. These graphs compare the 500 near-native decoys produced by model-based search (MBS) with those produced by Rosetta. These graphs indicate that model-based search determines decoys of much lower energy and reduced variance. (**a**) 2PTL, 60 amino acids, (**b**) 1APC, 106 amino acids, (**c**) 1J3G, 187 amino acids, and (**d**) 100U, 414 amino acids.

The structure predictions performed by model-based search are also biologically more accurate than the ones generated by Rosetta. This is illustrated in Figure 8, which shows how much model-based search was able to improve the similarity measure GDT_TS with respect to the decoys predicted by
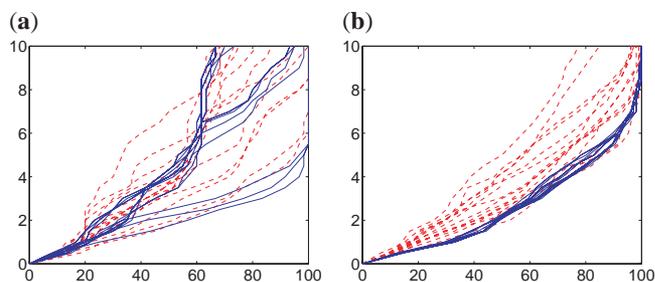
Fig. 7. Hubbard plots of the 15 lowest-energy structures produced by model-based search (solid) and Rosetta (dashed) for two proteins. The abscissa corresponds to the root mean square distance (RMSD) to native structure of the largest superimposable region expressed as a percentage of the length of the protein shown on the ordinate. This fraction, called coverage, is shown as a percentage on the ordinate. As the coverage increases, the RMSD also increases. Small slopes correspond to better predictions. (**a**) 2PTL, 60 amino acids, and (**b**) 1APC, 106 amino acids.
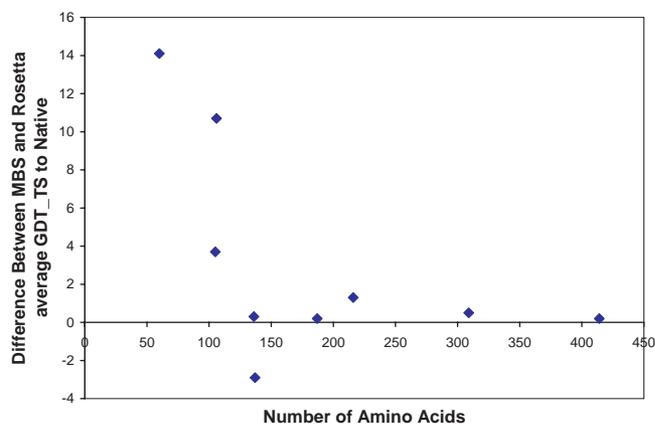
Fig. 8. A comparison of GDT_TS with the experimentally determined native structure of a number of proteins of varying length for model-based search (MBS) and Rosetta. Model-based search finds structures more similar to the native state than Rosetta. The performance improvement of model-based search becomes less pronounced as the size of the protein increases.

Rosetta. The GDT_TS measure indicates the percentage of residues of the predicted conformations that can be superimposed on the native structure within a given distance over four optimal sequence-dependent superpositions (1, 2, 4 and 8 Å) (Kinch *et al.*, 2003). GDT_TS is calculated using the software package LGA (Zemla, 2003).

Figure 8 shows that for small proteins a remarkable improvement in prediction accuracy of up to 14% can be obtained when model-based search is used instead of Rosetta. This improvement becomes much less pronounced for larger proteins, contradicting the fact that model-based search is able to determine conformations of much lower energy for larger proteins (Figs 4 and 5). This contradiction indicates that
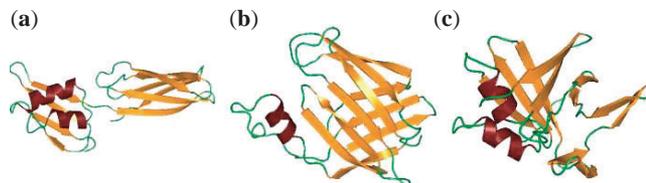

Fig. 9. Examples of protein structures for protein 1BLR (137 amino acids, a cellular retinoic acid binding protein) determined by model-based search and Rosetta, alongside the native structure. (**a**) Model-based search, (**b**) native structure and (**c**) Rosetta.

Rosetta's energy function becomes increasingly inaccurate for larger proteins.

Devising accurate energy functions is a tremendous challenge in itself. However, a more effective search procedure, such as model-based search, may be able to expose the shortcomings of existing energy functions and thus be an enabling technology to improve their accuracy. This is illustrated using protein 1BLR (Fig. 9). For this particular protein, Rosetta finds structures deemed more similar to native (albeit much higher in energy) than model-based search. (In Fig. 8, the single point below the x-axis corresponds to 1BLR.) Examination of the decoys indicates, however, that this similarity is based on disordered β-sheets in the proximity of the α-helix (only one representative example is shown in Fig. 8). In contrast, model-based search consistently finds structures in which β-sheets are clustered together and are in a favorable spatial arrangement with the α-helix. Given this more consistent search procedure, an appropriate adjustment of the parameters of the energy function relating to α-helices and β-sheets becomes possible and should lead to more accurate structure predictions.

The graph in Figure 10 compares the energy values assigned to conformations by Rosetta's energy function for the decoys obtained by Rosetta, model-based search and the native conformation of the protein. For short proteins, model-based search is able to determine conformations with energy lower than the native structure, probably indicating inaccuracies in Rosetta's energy function. For larger proteins, however, the discrepancy between the energy of decoys obtained using model-based search and the native structure cannot be explained solely by the energy function. It is apparent that the performance of search methods has to be improved further to address structure prediction for larger proteins. This motivates future improvements to model-based search.

Possible improvements to model-based search may be concerned with the underlying model used to represent the information obtained during search. Its critical importance to the quality and efficiency of model-based search is demonstrated by experiments with different underlying models. We compared the model described in Section 4 with a model that maintains only minimal energy samples and a model that maintains only the radius of the relevant regions, but not their
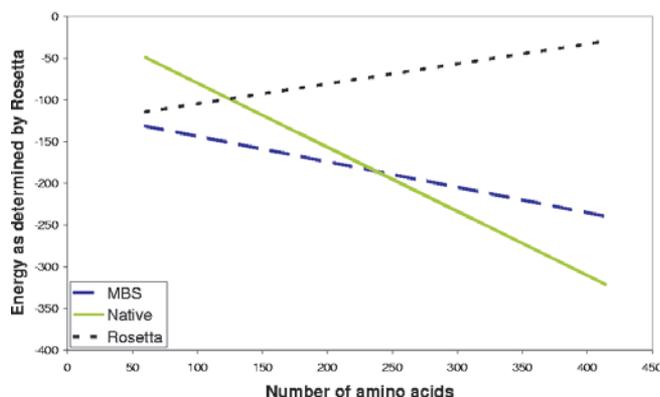
Fig. 10. A comparison of the average energy of decoys produced by model-based search (MBS) and Rosetta, and the energy of the experimentally determined native structure as determined by the energy function used for these experiments. For short proteins, model-based search finds lower-energy decoys than the energy of the native structure. This points to inaccuracies in the energy function, since the native state should represent the global minimum of the energy function.
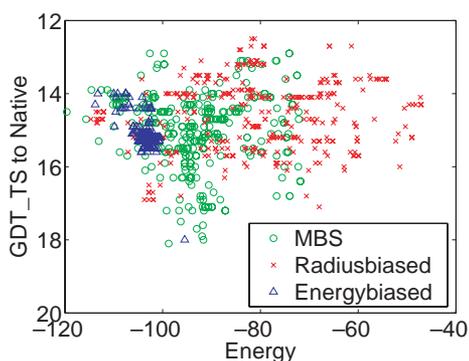
Fig. 11. The influence of the model on model-based search. A comparison of search with a model that incorporates both the energy level of a region and an approximation to the size of the region (MBS), one based entirely on the size of the region (radius-biased) and one based solely on energy level (energy-biased).

energy level. A model based solely on energy level causes model-based search to degenerate to a form of adaptive beam search, whereas a model based on the radius effectively performs uniform sampling in relevant regions by adapting the number of samples to the size of the region. The comparison of the decoys generated by model-based search with the respective models is shown in Figure 11. Not surprisingly, the most expressive model results in the lowest-energy decoys. This result provides motivation to explore alternative models for model-based search in the future with the aim of performing accurate structure prediction for larger proteins.

6 CONCLUSIONS AND FUTURE WORK

The accuracy of *de novo* protein structure prediction depends on accurate energy functions and effective search methods to find the global minimum of the energy function. To improve the accuracy of protein structure prediction, we presented a novel, general search method called model-based search. Experimental evidence shows that, using the same energy function and the same amount of computation, model-based search consistently finds lower-energy conformations than the customized Monte Carlo-based search procedure implemented by Rosetta, currently one of the leading methods of protein structure prediction. The performance improvement obtained with model-based search increases with the length of the protein and thus with the dimensionality of the associated search space. Generally, the lower energy level of conformations determined by model-based search also translates into improved structure prediction accuracy.

Model-based search differs from other search methods commonly applied to the problem of protein structure prediction in that it incrementally builds a model of the underlying search space. The model represents the information obtained so far during the search. This information is used to direct further exploration of the search space toward regions most likely to contain a relevant minimum of the energy landscape. By directing computational resources only to relevant regions of the search space, the computational burden of exhaustive search is avoided, while achieving good coverage of biologically relevant regions. This search procedure is experimentally shown to find lower-energy conformations than the Monte Carlo-based search method employed by Rosetta.

The results also indicate that the performance of model-based search degrades in higher-dimensional search spaces. Although this degradation occurs much more slowly than with other search methods, it is apparent that for very long proteins a near-optimal extremum of the energy landscape can be located only with significant computational effort. We attribute this to the fact that the amount of information obtained during the search process is not sufficient to focus the search on relevant regions. We will therefore explore the inclusion of domain information, as represented in protein structure databases, for example, to initialize the model. This additional information should result in an additional reduction of the search space and consequently significantly improve the performance of model-based search.

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

In this paper, we demonstrated that, by applying model-based search to protein structure prediction, significant improvements in finding low-energy conformations and in

prediction accuracy can be obtained. However, the application of model-based search to this specific application domain did not require any domain-specific assumptions. Therefore, we believe that model-based search can be applied to a wide variety of high-dimensional search problems. We anticipate that model-based search will be able to provide significant performance improvements for all search problems that are currently solved using Monte Carlo-based techniques.

ACKNOWLEDGEMENTS

REFERENCES

Anfinsen,C.B. (1973) Principles that govern the folding of protein chains. *Science*, **181**, 223–230.

Berg,B.A. and Neuhaus,T. (1992) Multicanoncial ensemble: a new approach to simulate first-oder phase transitions. *Phys. Rev. Lett.*, **68**, 9–12.

Bradley,P., Chivian,D., Meiler,J., Misura,K.M.S., Rohl,C.A., Schief,W.R., Wedemeyer,W.J., Schueler-Furmann,O., Murphy,P., Schonbrun,J. *et al*. (2003) Rosetta predictions in CASP5: successes, failures and prospects for complete automation. *Proteins: Stuct. Funct. Genet.*, **53**, 457–468.

Burns,B. and Brock,O. (2005) Sampling-based motion planning using predictive models. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, April 18–22, 2005, Barcelona, Spain.

Cohn,D.A., Ghahramani,Z. and Jordan,M.I. (1996) Active learning with statistical methods. *J. Art. Intell. Res.*, **4**, 129–145.

Frantz,D.D., Freeman,D.L. and Doll,J.D. (1990) Reducing quasi-ergodic behavior in Monte Carlo simulations by j-walking: applications to atomic clusters. *J. Chem. Phy.*, **93**, 2769–2784.

Glover,F. and Laguna,F. (1997) *Tabu Search*. Kluwer Academic Publisher.

Hansmann,U. (1997) Parellel tempering algorithm for conformational studies of biological molecules. *Chem. Phys. Lett.*, **281**, 140–150.

Holland,J. (1975) Adaptation in natural and artificial systems. *Artificial Intelligence*, **36**, 177–221.

Hubbard,T. (1999) RMS/coverage graphs: a qualitative method for comparing three-dimensional protein structure. *Proteins Suppl.*, **3**, 15–21.

Jones,D.T. and Thornton,J.M. (1996) Potential energy functions for threading. *Curr. Opin. Struct. Biol.*, **6**, 210–216.

Kinch,L., Wrabl,J., Krishna,S., Majumdar,I., Sadreyev,R. and Qi,Y. (2003) CASP5 assessment of fold recognition target predictions. *Proteins*, **53**, 395–409.

Leach,A.R. (1991) *Molecular Modelling—Principle and Applications*. 2nd edn., Prentice Hall.

Lee,J. (1993) New Monte Carlo algorithm: entropic sampling. *Phys. Rev. Lett.*, **71**, 211–214.

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

MacKay,D. (1992) Information-based objective functions for active data selection. *Neural Comput.*, **4**, 590–604.

Metropolis,N. and Ulam,S. (1949) The Monte Carlo method. *J. Am. Stat. Assoc.*, **44**, 335–341.

Moult,J., Fidelis,K., Zemla,A. and Hubbard,T. (2003) Critical assessement of methods of protein structure prediction (CASP)-round V. *Proteins: Stucture, Function and Genetics*, **53**, 334–339.

Okamoto,Y. (1998) Protein folding problem as studied by new simulation algorithms. *Recent Research Developments in Pure & Applied Chemistry*, **1**, 1–23.

Rohl,C.A., Strauss,C.E.M., Misura,K.M.S. and Baker,D. (2004) Protein structure prediction using rosetta. *Meth. Enzymol.*, **383**, 66–93.

Russell,S. and Norvig,P. (2003) *Artificial Intelligence A Modern Approach*, 2nd edn., Pearson Education Inc.

Kirkpatrick,S., Gelatt,C.D. and Vecchi,M.P. (1983) Optimization by simulated annealing. *Science*, **220**, 671–681.

Xu,H. and Berne,B.J. (1999) Mulitcanonical jump walking: a method for efficiently sampling rough energy landscapes. *J. Chem. Phys.*, **110**, 10299–10306.

Zemla,A. (2003) LGA: A method for finding 3D similarities in protein structure. *Nucleic Acid Res.*, **31**, 3370–3374.

Zhang,Y. and Skolnick,J. (2001) Parallel-hat tempering: a Monte Carlo search scheme for the identification of low-energy structures. *J. Chem. Phys.*, **115**, 5027–5032.