Biophysical Journal Volume 82 January 2002 36–49

A Physical Approach to Protein Structure Prediction

Silvia Crivelli,* Elizabeth Eskow,† Brett Bader,† Vincent Lamberti,† Richard Byrd,† Robert Schnabel,† and Teresa Head-Gordon‡

*Physical Biosciences and NERSC Divisions, Lawrence Berkeley National Laboratory, Berkeley, California 94720; †Department of Computer Science, University of Colorado, Boulder, Boulder, Colorado 80309; and ‡Department of Bioengineering, University of California, Berkeley, Berkeley, California 94720

ABSTRACT We describe our global optimization method called Stochastic Perturbation with Soft Constraints (SPSC), which uses information from known proteins to predict secondary structure, but not in the tertiary structure predictions or in generating the terms of the physics-based energy function. Our approach is also characterized by the use of an all atom energy function that includes a novel hydrophobic solvation function derived from experiments that shows promising ability for energy discrimination against misfolded structures. We present the results obtained using our SPSC method and energy function for blind prediction in the 4th Critical Assessment of Techniques for Protein Structure Prediction competition, and show that our approach is more effective on targets for which less information from known proteins is available. In fact our SPSC method produced the best prediction for one of the most difficult targets of the competition, a new fold protein of 240 amino acids.

INTRODUCTION

The protein folding problem in its most pragmatic guise is to predict the full three-dimensional structure of the protein molecule given only the amino acid sequence as input. When proteins are approached from a purely physical point of view, a number of issues arise in successfully meeting the demands of this difficult problem. The first is the quantitative uncertainty of the free energy function describing both the proteins' intramolecular interactions and the intermolecular interactions with aqueous solvent for arbitrary conformation. Second, there exists an exponentially large number of local minima on this solvent-averaged free energy hypersurface, a huge majority of which are false traps that make it difficult to determine the global free energy minimum that often corresponds to the correct native structure of the protein. Given the difficulty of these two obstacles, a popular alternative viewpoint is to diminish the emphasis on proteins as strictly physical systems but instead to exploit empirical structural correlations statistically evaluated or determined from machine learning methods on databases of existing protein structures.

Protein structure prediction methods can be classified into the categories of comparative modeling, fold recognition, and so-called ab initio or "new-fold" methods (Moult et al., 1999). These methods are ordered in their decreasing reliance on comparisons to known protein structures from a protein database, although structure prediction methods often combine or share underlying methodologies in these distinct categories. For example, the ab initio category covers a broad range of methodologies, from approaches that introduce tertiary knowledge through protein structure databases, to approaches that use secondary structure knowledge through prediction servers, to methods that use the information of known protein structures only in determining the weights of various terms in their simplified potential energy function (Orengo et al., 1999; Simons et al., 1999; Ortiz et al., 1999; Samudrala et al., 1999; Lee et al., 1999; Eyrich et al., 1999; Shortle, 2000). The most successful methods at present are those that can most effectively use information from the sequence and structure of known proteins to form some type of structural template for predicting tertiary structure of unknown targets. However, for targets where this information is unavailable, these methods may be somewhat less successful than those that rely more on generically applicable physical principles.

In this paper, we describe our energy-based new fold method called Stochastic Perturbation with Soft Constraints (SPSC), which uses information from known proteins to predict secondary structure but not in the tertiary structure predictions or in generating the terms of the physics-based energy function. The SPSC approach combines elements of the antlion method (Head-Gordon and Stillinger, 1993) and the stochastic/perturbation method (Byrd et al., 1994; Crivelli et al., 1999, 2000; Azmi et al., 2000). Like the antlion method, it reduces the search space by defining biasing functions based on predictions for secondary structure that in no way can serve alone as a search strategy to predict tertiary structure. However, the search is fundamentally global, and the stochastic perturbation approach is used to solve a series of global optimization problems in smaller subspaces of back-bone dihedral angles predicted to be coil. Our approach is also characterized by the use of an all atom energy function that includes a novel hydrophobic solvation function derived from experiments conducted in the Head-Gordon group, which shows promising ability for discrimination against misfolded structures. An important part of

Received for publication 23 April 2001 and in final form 24 August 2001.

Address reprint requests to Teresa Head-Gordon, Department of Bioengineering, University of California, Berkeley, Berkeley, California 94720; Tel.: 510-486-7365; Fax: 510-486-6488; E-mail: tlhead-gordon@lbl.gov.

© 2002 by the Biophysical Society

0006-3495/02/01/36/14 $2.00

the implementation of this algorithm involves grappling with effective parallelization strategies to tackle arbitrarily large proteins. Furthermore, the prediction of β-containing proteins is presented for the first time here. Finally, we discuss the results obtained in the 4th Critical Assessment of Techniques for Protein Structure Prediction (CASP4) competition and show that our method is more effective on targets for which less information from known proteins is available. In fact our SPSC method produced the best prediction for one of the most difficult targets of the competition, a new fold protein of 240 amino acids.

MATERIALS AND METHODS

Energy function

Empirical protein force fields have formed the basis of most studies of protein structure, function, and dynamics to date. We use the AMBER (Cornell et al., 1995) empirical protein force field that represents bonds and angles as harmonic distortions, dihedrals by a truncated Fourier series, and nonbonded interactions via Lennard-Jones 6-12 terms and Coulomb's Law for electrostatic interactions between point charges.

$$E_{\mathrm{AMBER}} = \sum_{\mathrm{i}}^{\#\mathrm{bonds}} k_{\mathrm{b}}(b_{\mathrm{i}} - b_0)^2 + \sum_{\mathrm{i}}^{\#\mathrm{angles}} k_\theta(\theta_{\mathrm{i}} - \theta_0)^2 + \sum_{\mathrm{i}}^{\#\mathrm{dihedrals}} k_\chi[1 + \cos(n\chi + \delta)] + \sum_{\mathrm{i}}^{\#\mathrm{atoms}} \sum_{\mathrm{i<j}}^{\#\mathrm{atoms}} \left(\frac{q_{\mathrm{i}} q_{\mathrm{j}}}{r_{\mathrm{ij}}} + \epsilon_{\mathrm{ij}} \left[\left(\frac{\sigma_{\mathrm{ij}}}{r_{\mathrm{ij}}} \right)^{12} - 2\left(\frac{\sigma_{\mathrm{ij}}}{r_{\mathrm{ij}}} \right)^{6} \right] \right). \quad (1)$$

A crucial component of our energy model is the development of additional terms to the energy function that more accurately model solvation effects, i.e., the behavior of the protein in an aqueous environment. An important demonstration of the effect of solvent is the study of native folds and misfolds of the protein sequences of hemerythrin, a predominantly α protein, and immunoglobulin VL domain, a predominantly β-sheet structure. When the sequences of the two were threaded through the others tertiary structure, the gas phase energy values for the native folds were comparable to their misfolds. However, the positive point of this work was to demonstrate that addition of a simple solvation description allowed these protein force fields to perform effectively in discriminating between correct folds and misfolds (Novotny et al., 1984). In particular, new terms are needed that model hydrophobic and hydrophilic behavior in solution.

To address this issue, an empirical solvation free energy term $E_{\mathrm{solvation}}$ has been added to the energy function used by the optimization. $E_{\mathrm{solvation}}$ models the hydrophobic effects as a two-body interaction between all aliphatic carbon centers. This description is motivated by recent experimental, theoretical, and simulation work to determine the role of hydration forces in the structure determination of model protein systems (Head-Gordon et al., 1997; Hura et al., 1999; Pertsemlidis et al., 1996; Sorenson et al., 1999). This work and that of others on the free energy of association of hydrophobic groups in water (Pratt and Chandler, 1977) show this interaction potential has two minima separated by a barrier: one for the hydrophobic molecules in contact and one for the hydrophobic groups separated by a distance of one water molecule or layer. Our new solvation term embodies these characteristics and is parameterized using a functional form of a sum of Gaussians

$$E_{\mathrm{solvation}} = \sum_{\mathrm{i}}^{\mathrm{N_c}} \sum_{\mathrm{j}}^{\mathrm{N_c}} \sum_{\mathrm{k}}^{\mathrm{M}} h_{\mathrm{k}} exp\left(- \left[\frac{r_{\mathrm{ij}} - c_{\mathrm{k}}}{w_{\mathrm{k}}} \right]^2 \right) \quad (2)$$

in which the sums i and j are over the aliphatic carbon centers, and each of the M Gaussians is parameterized by position c_{k}, depth h_{k}, and width w_{k} so as to describe the two minima and the barrier. The benefits of this description is that 1) it introduces a stabilizing force for forming hydrophobic cores, 2) it is a well-defined model of the hydrophobic effect for hydrophobic groups in water, and 3) it can be described as a continuous potential that is more computationally tractable than other solvent accessible surface area models. Its novelty in the context of structure prediction and protein energetics is the extra stabilization at a longer length scale for the hydrophobic interaction that is not described by surface area solvation models.

We have tested our solvation energy function on two α-helical proteins, the A-chain of uteroglobin (2utg A) and the four-helix bundle DNA binding protein (1pou). Using this form of potential, we found good agreement with experiment in that the potential energy of the experimentally determined structure was lower than the potential energy of any of the structures found by the global optimization algorithm. However, we used parameters values in Eq. 2 that exaggerate the stability of the contact and solvent-separated minima in comparison with the parameter values based on the experimental work mentioned above.

SPSC algorithm for global optimization

The SPSC method combines knowledge about secondary structure with a large-scale global optimization method. The secondary structure information is embodied in biasing terms added to the potential energy and later removed. The SPSC method consists of two phases. Phase 1 generates good initial configurations using secondary structure information, readily available through servers with accuracies averaging ~75% (Jones, 1999; Cuff et al., 1998). These initial configurations are generated so that they have much of the predicted secondary structure but no tertiary structure. This is an important feature of the algorithm, dramatically reducing the size of the conformational space, and allowing us to tackle reasonable-sized problems.

The second phase improves the initial structures by performing small-scale global minimizations in various subspaces of the parameter space that are randomly selected. The small-scale global optimizations use the stochastic method of Rinnooy-Kan and Timmer (1984) that provides some theoretical guarantee of success when applied to spaces of 4 to 10 variables. The small-scale minimizations are followed by full-scale local minimizations to convert the small-scale minimizers into full-scale ones. These minimizers, which are kept in a list, are clustered via pairwise root mean squared deviation (RMSD) evaluation and ordered by energy value within each cluster. A new list is formed with the lowest energy minimizer from each cluster, and phase 2 starts again with this new list. The process repeats until the stopping criteria are met. Algorithm 1 shows a framework of the SPSC method.

Phase 1: identification of configurations

Phase 1 builds initial configurations that contain predicted secondary structure obtained from the servers. Given the initial sequence, the server predicts whether each amino acid of the target protein is α, β, or coil and gives the confidence value for each prediction based on tendencies from other known proteins. The process begins with a buildup procedure similar to that proposed by Gibson and Scheraga (1987). This procedure samples on the set of dihedral angles for amino acids predicted to be "coil" some fixed number of times, and selects the angle values that produce the best

TABLE 1 Algorithm 1, the SPSC method

Phase 1: Coarse Identification of Configurations:
Generate initial configurations containing domain specific tendency information.
(a) Sampling in Full Domain
Generate the parameters of some sample configurations.
(b) "Biased" Full-Dimensional Local Minimizations:
Use server obtained information to create biasing terms for predicted secondary structure. Perform a local minimization from a subset of the sample points, using "biasing" penalty functions to superimpose predicted secondary structure on the standard energy surface.
(c) "Unbiased" Full-Dimensional Local Minimizations:
Perform a local minimization from each "biased minimizer" using the "unbiased" energy function. Save these local minimizers for step 2a.
Phase 2: Improvement of Local Minimizers:
For some number of iterations:
(a) Select a Configuration and Small Subset of Parameters to Improve:
From the list of full-dimensional local minimizers, select a local minimizer to improve. Then select a subset of the parameters of this configuration to optimize in step 2b.
(b) Small Sub-problem Global Optimization:
Apply a global optimization algorithm to the energy of the selected configuration with only the selected parameters as variables.
(c) Full-Dimensional Local Minimization:
Apply a local minimization procedure, with all the parameters as variables, to the lowest energy configurations that resulted from step 2b, and merge the new local minimizers into the list of local minimizers.
(d) Cluster Local Minima and Test for Convergence:
Cluster the list of local minimizers via pairwise RMSD and if the stopping criteria has not been met, repeat all steps of Phase 2 from a new list of local minimizers containing the lowest energy minimizer from each cluster.

partial energy for the part of the chain built so far before proceeding to the next amino acid. For amino acids that are predicted to be α or β, the dihedral angles are set to the ideal values for those types of structures. A subset of the best structures generated by this buildup procedure is then selected as starting points for full-dimensional local minimizations using the antlion strategy (Head-Gordon et al., 1991; Head-Gordon and Stillinger, 1993). The "antlion" strategy applies predicted secondary structure information in energy minimizations. It uses biasing or penalty functions designed to enforce the information from the secondary structure prediction server in step 1b of algorithm 1. The biasing functions used are described in the next two sections.

Biasing functions for α-helical proteins

The α-helix is stabilized by hydrogen bonds connecting the carbonyl oxygen of residue i to the amide hydrogen of residue $i + 4$. Two functions have proved very effective in encouraging formation of α-helices (Crivelli et al., 1999, 2000; Azmi et al., 2000). The first function,

$$E_{\phi\psi} = \sum_{\text{dihedrals}} k_\phi[1 - cos(\phi - \phi_0)] + k_\psi[1 - cos(\psi - \psi_0)] \quad (3)$$

biases the backbone torsional angles of the amino acids of a residue that is predicted to be an α-helix to be close to ideal values for an α-helix. Here ϕ_o and ψ_o are the dihedral angles of a perfect α-helix, and k_ϕ and k_ψ are force constants related to the strength of the prediction from the prediction server. The second function,

$$E_{HB} = q_i q_{i+4}/r_{i,i+4} \quad (4)$$

encourages α-helical hydrogen bonds to form between the oxygen of residue i and the hydrogen of residue $i + 4$, if residues i and $i + 4$ are predicted to be helical. In this function, the "charges," q, are the weights of the prediction output by the server, and provide a strong incentive for an intramolecular hydrogen bond to form when residues i and $i + 4$ are strongly predicted to be helical. Together these functions "bias" the protein toward forming α-helical shapes in regions predicted to be helix. These biasing functions are not part of the energy discrimination function but instead are simply soft constraints defined in the antlion approach and are therefore part of the search strategy.

Biasing functions for proteins containing β-sheets

Forming β-sheets in proteins is an intrinsically hard problem. Because α-helices contain hydrogen bonds along the backbone from neighboring amino acids, these interactions are relatively short and local, spanning only four residues. On the other hand, β-sheets usually have nonlocal hydrogen bonds, where the hydrogen bonds span many more residues. This nonlocal nature of β-sheets requires a modified approach to form them in a protein structure prediction algorithm. Secondary structure predictions provide information regarding which residues or segments of the amino-acid chain are predicted to be β-strands but not how these strands either pair up or align within a pair to form correct β-sheets.

When using secondary structure predictions of β-strands, we first use Eq. 3 to bias the backbone torsional angles to be close to ideal values for a β-strand. Here ϕ_o and ψ_o are the dihedral angles of a perfect β-sheet, and k_ϕ and k_ψ are force constants related to the strength of the prediction from the prediction server. Again, these biasing functions are not part of the energy discrimination function but instead are simply soft-constraints defined in the antlion approach.

The challenge is then to correctly pair and align the β-strands into the correct β-sheet(s) in the tertiary structure. One of the difficulties that must be overcome is the combinatorial explosion of possible β-strand matches. The choices are limited to parallel versus antiparallel orientations, and hydrogen bonds on the even or odd residues. When five strands or fewer are correctly identified, then the antiparallel orientation is always preferred. Often, the predictions do not yield strands of equal length, which further multiplies the combinations by the offset length plus one. The presence of additional strands complicates matters further by introducing even more possible matchings, and the problem grows exponentially harder with each additional β-strand. Thus, a biasing scheme that tests each possibility in turn would require many time-consuming runs.

We limit this combinatorial complexity by removing some of the potential matchings via a preprocessing step. The technique used by SPSC

is to evaluate each possible pair of strands using a scoring function based on occurrences of β-sheets in a protein database (Zhu and Braun, 1999). The scoring function returns a value representing the bonding probability of a pair of residues for forming β-sheet-type hydrogen bonds. The scores of each possible pair of β-strands, with varying alignments, are summed, and the best-scoring physically feasible set of strand pairs is selected for use in the β-biasing function given below. If more than one set of good-scoring, feasible strand pairs is identified, each is used in a separate instantiation of phase 1. Results from the various runs of phase 1, each using a different set of strand pairings in the biasing function, are compared and the best structures are selected by energy values and number of contacts (hydrogen bonds between β-strands) formed.

Because the set of hydrogen-oxygen pairs along the two strands includes both the H-bonded and non-H-bonded pairs, a biasing function that handles both types concurrently and without any explicit identification is necessary. In our work, this is accomplished by the following piecewise continuous biasing function, which couples a linear function with a sigmoid function,

$$E_{\mathrm{HB}}^{\beta} = \begin{cases} \dfrac{\epsilon\tau}{4}\left(\dfrac{r_{\mathrm{ij}} - \sigma}{\omega} + 2\right) & \text{if} \quad r_{\mathrm{ij}} > \sigma \\ \dfrac{\epsilon\tau}{1 + \exp\left(\dfrac{\sigma - r_{\mathrm{ij}}}{\omega}\right)} & \text{if} \quad r_{\mathrm{ij}} \leq \sigma \end{cases} \tag{5}$$

in which ϵ is the dielectric constant, r_{ij} is the Euclidean distance between atom i and atom j, τ is a scale factor for appropriate balance with other forces in the model, σ is the sigmoid offset from the origin, and ω scales the sigmoid width. The linear term ($r_{\mathrm{ij}} > \sigma$) attracts atom pairs from afar to be at least the distance of a typical non–H-bonded pair. Then the attractive force within the sigmoid term ($r_{\mathrm{ij}} \leq \sigma$) decays as the two atoms move nearer to each other. There is still enough attraction, however, for the biasing function in conjunction with the other terms in the energy function to encourage a hydrogen bond to form between the most likely hydrogen-bonded pair yet not too much force to disrupt a non–H-bonded pair. The three parameters in the biasing function, τ, σ, and ω, affect the formation of hydrogen bonds in β-sheets. Our current set of experimental parameters $\tau = 2.09$, $\sigma = 16$, $\omega = 4$ appears to work well, but we still believe there is room to improve this biasing function by tuning these parameters. Again, the biasing term $E_{\mathrm{HB}}{}^{\beta}$ in Eq. 5 is not part of the energy function but part of the search strategy.

Phase 2: improvement of local minimizers

A number of the best minimizers generated in phase 1 are used as starting configurations in phase 2. This phase, which accounts for most of the computational effort of the method, improves those local minimizers through a sophisticated global minimization algorithm. The basic idea of this phase is to select a number of promising configurations from the list of local minimizers and then select small subsets of their variables for improvement followed by full variable local minimizations on the best resulting configurations. This process expands a subtree of local minimizers from each minimizer selected. The strategy for selecting configurations from the list in step 2a of algorithm 1 is to use a combination of breadth (expand different subtrees) and depth (expand the subtree with the lowest energy structure). Thus, for some number of iterations of phase 2, the least developed subtree is selected, and the lowest energy configuration in this tree that has not already been expanded is chosen for improvement. After a number of iterations of this "balancing" stage, the remaining iterations of phase 2 correspond to the "nonbalancing" stage. In this stage, the lowest energy configuration is selected regardless of which tree it comes from. We have found that this heuristic in the search of the configuration space contributes to the success of this method mainly because the energy function is not monotonically decreasing.

Once a configuration is chosen, a small subset of its variables is selected for modification by global optimization. The subset of variables consists of a small number of dihedral or torsional angles of the protein picked randomly from the set of amino acids predicted to be "coil" by the secondary structure prediction. Once the subset has been determined, a stochastic global optimization procedure is executed to find the best new positions for the chosen dihedral angles while holding the remaining dihedral angles fixed. This stochastic global optimization approach provides a theoretical guarantee of finding the global minimum, and tests have shown it to be efficient compared with other global optimization approaches for problems with small numbers of parameters. The global optimization method samples over the parameter space, and it selects a sample point to be a start point for a local minimization only if that sample point has the lowest energy of all neighboring points within a certain distance metric. If a sample point lies within the distance metric of another point that is lower in energy, it is assumed to lie within an existing basin of attraction and is rejected from further computational consideration. Because the probabilistic theoretical guarantee is easier to satisfy computationally for small dimensional problems, we select a subset size of ~6 to 10 variables (from the space of torsion angles of the protein). This global optimization algorithm is general in the sense that a parameter space of arbitrary dimension can be explored, however, in practice, the amount of work required to reach the theoretical guarantee is prohibitive for large subspaces.

The small-scale global optimization expands a tree of configurations that present significantly different shapes than the root. Those configurations represent local minimizers in the small subspace of chosen parameters (torsion angles). A number of those conformations with the lowest energy values are selected for local minimizations in the full variable space. These full-scale local minimizations are less likely to produce major structural changes but can cause important, more local refinements throughout the protein. The new full-dimensional local minimizers are then merged with those found previously, and the entire phase is repeated a fixed number of times.

Periodically during phase 2 the minima are ordered and clustered to decide whether to terminate phase 2 or continue. The clusters are defined so that members of each cluster are within ~5 Å RMSD of the lowest energy configuration in that cluster. The number of clusters indicates the number of diverse structures that exist at this stage.

Stopping criteria

We have determined some guiding criteria for deciding when the global optimization algorithm has converged. At the end of each global optimization run, a list is kept of all minima obtained, and this list is clustered via pair wise RMSD evaluations, and ordered by energy value within each cluster. In early stages of the global optimization runs, we see that the number of distinct clusters expands in number and that significant energy lowering is observed as compactness increases and tertiary structure is formed within the lowest energy structures of these clusters. In midstages of our global optimization approach, it is not uncommon for our energy to "stall," i.e., the global optimization run may have worked on a subset of coil residues that contributed to greater diversity in cluster number but no improvement in the lowest energy clusters of results found up until this point. After this point, several outcomes are possible with the next run of the global optimization algorithm. One possibility is that working on a different subset of coil residues produces further energy lowering of the lowest cluster. The algorithm is judged to have not converged, and more runs are planned. Second, it may happen that one of the structures from the higher energy clusters yields a new energy minimum structure that is now the lowest energy cluster found thus far. Again, the algorithm is judged to have not converged and more runs are planned. Third, there is a further blossoming of the number of distinct clusters. Again, the algorithm is judged to have not converged, and more runs are planned. Finally, if one (or more if resources permit) subsequent run of the global optimization

algorithm finds no change in cluster number and no further lowering of energy in our lowest energy cluster, the algorithm is assumed to have converged at this point.

Parallel implementation of SPSC

The amount of computational time needed to solve realistic problems of interest with the SPSC method makes the use of parallel computers a necessity. In fact, our current runs on the Cray T3E at NERSC take many processors for many hours to converge. Although algorithmic improvements may decrease this time, the need to solve considerably larger problems and the exponential nature of the number of minimizers almost guarantee that these will always be very large applications.

The SPSC method is highly but not straightforwardly parallelizable. The main problems are how to partition the load and keep it balanced considering that the work is dynamically generated and its computational time unknown and how to efficiently gather the partial results generated so far without jeopardizing the scalability of the code. We have applied a hierarchical approach to deal with these issues (S. Crivelli, T. Head-Gordon, submitted manuscript).

The approach divides the system into three different categories of processors and two levels of work, each dealing with different types of tasks and granularities. In the first category is a central scheduler that divides the work coarsely, assigns it to the intermediate processors, collects the results, and keeps the global information updated. The central scheduler maintains the current list of minimizers and distributes it to the processors in the next category according to some heuristic. In the next category are the supervisors. Each supervisor receives a specific task from the central scheduler, i.e., a minimizer and a subspace to perform a small-scale global optimization. They split the work at a finer level and distribute it among their workers. The supervisors control the work of their workers and manage the local information in their group but have no knowledge about the computation in other groups. Finally, in the third category are the workers that perform smaller subtasks such as sampling and local minimizations, and report the results to the supervisors.

Because the amount of computational work assigned to each supervisor and its workers is large but unknown, it is hard to make an efficient assignment of the work. Thus, some of the supervisors and their workers may be working for quite some time, while other supervisors and their workers may be sitting idle waiting for them to finish. We have implemented an efficient dynamic load balancing strategy that reassigns idle workers to busy supervisors thus changing the virtual communication tree among the processors as the computational tree changes. This is efficient because rather than moving tasks and incurring significant overhead, the idle supervisors only communicate their workers' ids to the busy ones. The busy supervisors, in turns, only need to update their table of workers.

The hierarchical approach is scalable to large number of processors and has been shown to run effectively on up to 256 processors of the Cray T3E at NERSC. In fact, as the number of processors increases, new categories can be added to the hierarchy to avoid communication bottlenecks.

RESULTS

Our collaborative team recently participated in CASP4 for the first time, where we competed in the "new fold" or ab initio category. The CASP4 experiment ran from May 11 through September 15, 2000. During this time the amino acid sequences for 43 target proteins, whose structures were in the process of being determined experimentally, were released to participants for "blind" prediction. Predictions were due on various dates during that time frame, and the amount of time allocated for predicting a specific target was variable. Up to five models of structure prediction for each target were accepted, however, the submission specified as "model 1" (M1) was predominantly used for evaluation of structure predictions. We submitted models for eight different target proteins, sometimes submitting more than one model per target. We always chose as M1 the prediction with lowest potential energy. In all cases, we submitted predictions for the entire sequence of the target.

Targets T0091, T0097, and T0105 were in the fold recognition/new fold category, with T0097 having a bigger percentage of structure homology to existing structures than the other two targets. T0098 was classified as a new fold but later considered as fold recognition/new fold. Conversely, T0110 was classified in the fold recognition category but later switched to the new fold one. Target T0124 was a new fold. It should be emphasized that we treated all targets as if they were new folds, using only secondary structure information in all cases.

First we present results on secondary structure prediction accuracy for all eight targets, showing the overall effectiveness of phase 1 of SPSC described in Materials and Methods. Then we discuss tertiary structure results for the eight targets compared with the other CASP4 submissions as a function of difficulty of the targets based on the criteria used in CASP4. Finally, we present more detailed performance of our results for our eight target submissions.

Secondary structure prediction accuracy

The SPSC method takes advantage of existing secondary structure prediction methods. These methods, in general, are much more advanced than methods predicting overall tertiary structure. Given their widespread use, it would be impractical not to attempt to use them in this problem domain, and only by using them can we realistically tackle reasonable-sized problems in protein structure prediction at this time. Forming predicted α-helical segments using the biasing functions of phase 1 of SPSC is not very difficult due to the local nature of the hydrogen bonds in α-helices (see description of the method). However, this task is much more difficult in the case of β-sheet formation because β-strands may be located at distant parts of the protein, and there may be only one or two hydrogen bonds formed to hold β-strands together in a β-sheet. We had not worked on targets with β-sheets before CASP4, and we developed our β-biasing techniques as we were actually predicting for CASP4.

Fig. 1 shows three lines of secondary structure for each target consisting of 1) the secondary structure predictions we used or some combination of prediction servers (top), 2) the secondary structure in our submitted model 1 (middle), and 3) the secondary structure of the target protein (bottom). The darker lines in the figure represent helical segments and the light segments are β-strands. For targets T0091, T0099, T0110, and T0124, some parts of the secondary structure of

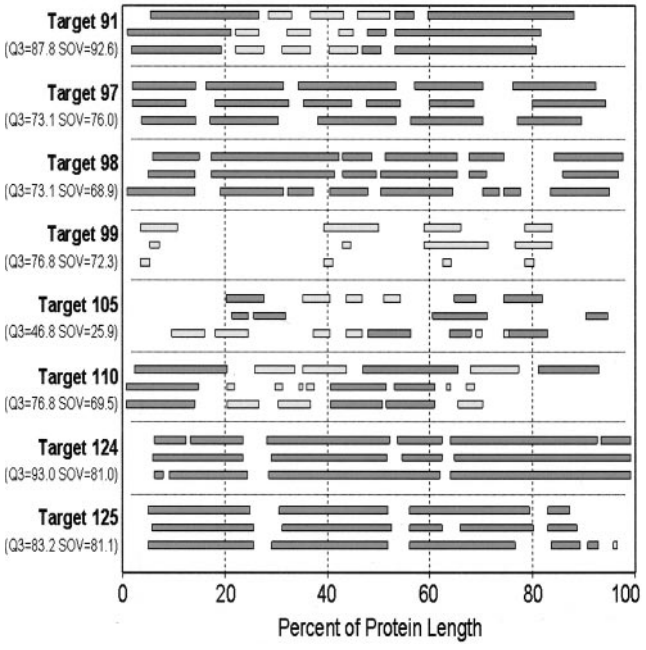


FIGURE 1 Secondary structure accuracy for our eight predictions in CASP4. The three lines for each target represent (*top*) the secondary structure predictions used to generate initial configurations in phase 1 of SPSC, (*middle*) the secondary structure of our model 1 submission, and (*bottom*) the secondary structure of the experimental structure. The dark lines denote α-helical and the lighter lines are β-strands. To the left, "Q3" and "SOV" are measures of the percentage of secondary structure accuracy of our model 1 submissions.

our models are closer to the target's actual secondary structure than what was predicted by the server we used. This shows that our method not only can incorporate predicted secondary structure in phase 1 but sometimes can improve upon it significantly in phase 2 of SPSC. On the other hand, for targets T0097, T0105, and T0125, the predicted secondary structure was closer to the target's actual secondary structure than the secondary structure obtained in our models, although this deterioration is not nearly as significant as the improvements in the previous set of cases. Target T0098 had the predicted secondary structure implemented accurately, but the secondary structure prediction was different from the target secondary structure.

The overall secondary structure accuracy for the M1 we submitted for each target is evaluated by two numbers that are given to the left of the figure. The "Q3" percentage is measured by the percentage of helical, β, and coil residues predicted correctly over the number of all residues of the protein (Zemla et al., 1997). The segment overlap measure (SOV) is a more structurally meaningful measure of secondary structure prediction accuracy that also ranges from 0 to 100 and is defined in Zemla et al. (1999). For target T0105, the SPSC method did not form the predicted β-sheet, and although the individual β-strands in the model are not very far off from being within hydrogen bond distance, the Q3 and SOV scores are both very poor for that target. Except for target T0105, the Q3 measures of our models range from 73.1 to 93.0, and the SOV measures range from 68.9 to 92.6, showing that the incorporation of predicted secondary structure into the submitted models, via the initial configurations generated by phase 1 of SPSC is largely successful. It is also interesting that in some cases where the predicted secondary structure was in error, our algorithm was able to correct this error to some extent in making the final tertiary structure prediction.

Tertiary structure prediction performance

The organizers of CASP4 have ranked all of the targets with respect to the percentage of sequence or structure homology found in existing structural databases. The "easiest" targets have a high percentage of sequence similarity to known proteins, "harder" targets have some structural similarity to known proteins, and the hardest targets have very little similarity to known proteins and are called "new folds." The CASP4 organizers classified all 43 targets into eight difficulty bins. In Fig. 2, we plot the comparative difficulty of each protein we attempted in CASP4 versus the comparative accuracy of our prediction to that of other groups. The comparative accuracy measure is based on the overall accuracy measure used in CASP, GDT_{TS}. The *GDT* is the global distance test that measures the largest set of contiguous residues whose RMSD from the target is under a certain distance cutoff. The measure GDT_{TS} is the *GDT* total score, measured as

$$GDT_{\mathrm{TS}} = (GDT_{\mathrm{P1}} + GDT_{\mathrm{P2}} + GDT_{\mathrm{P4}} + GDT_{\mathrm{P8}})/4.0 \tag{6}$$

in which each GDT_{Pn} is an estimate of the percent of residues under distance cutoff $\leq$ nÅ. The comparative accuracy ranking is the percent of groups with poorer GDT_{TS} scores on that protein. Fig. 2 *a* shows the comparative ranking among all M1 predictions, and Fig. 2 *b* shows the comparative ranking among all submitted models for a given protein target.

Fig. 2 shows that as the difficulty of the targets increases, the GDT_{TS} percentile of our models ranked against all other M1 submissions generally increases as well. In other words, the SPSC method is relatively more effective on targets where less information from known proteins is available. Because SPSC uses knowledge only in forming secondary structure, but not in the prediction of overall tertiary structure, it provides a complementary strength to most methods used for CASP4 predictions that are more successful when knowledge from known proteins is available. In what follows, we discuss in more detail the results for all the targets that we submitted.

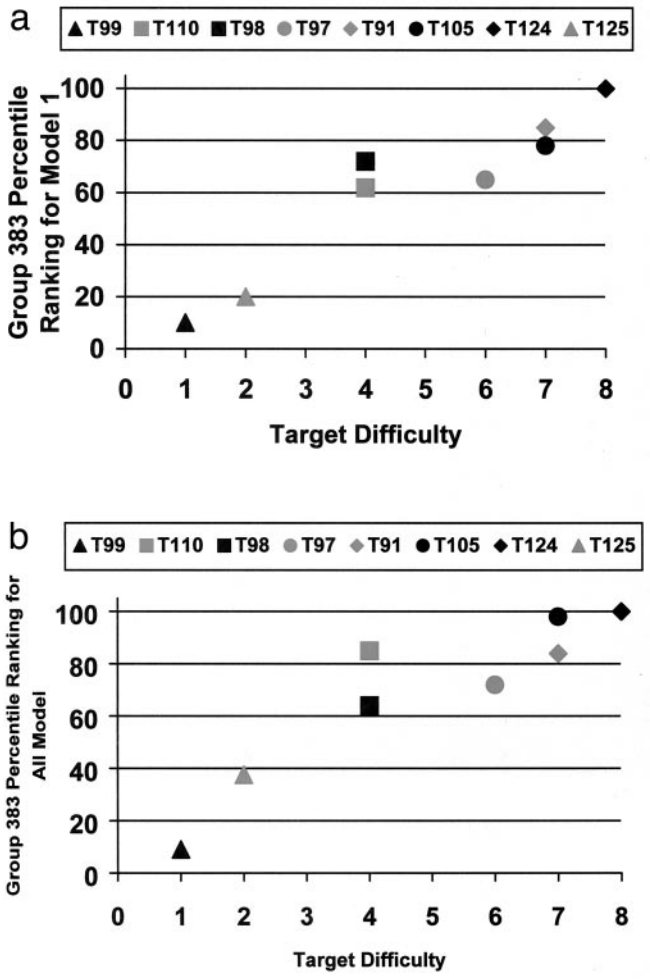


FIGURE 2 Difficulty of CASP4 targets as rated by the CASP4 organizers versus the percentile ranking of our group's (383) model 1 submissions using SPSC. The percentile ranking of our models generally increases with target difficulty. (*a*) Percentile ranking with respect to target difficulty for model 1 predictions. (*b*) Percentile raking with respect to target difficulty for all models submitted.

Target T0091: hypothetical protein HI0442, *H. influenzae*

We submitted only one model for target T0091. Matching between the secondary structure of the models and the target is quite good as illustrated in Fig. 1. Our overall GDT_{TS} score is above average as seen in Fig. 3, especially for distance cutoffs beneath 5 Å, and overall percentile ranking was ~80% for this target (Fig. 2). Better results on this target should be obtainable with additional research in the following areas: 1) our biasing strategies for β-sheet proteins and 2) experiments with the optimization of various dimer pairings. Confirmation of point 2 is pending because the experimental Cartesian coordinates have not been made available to the CASP4 predictors as of this date.

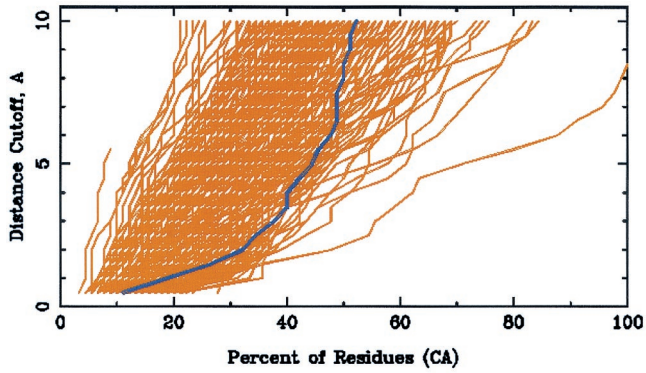


FIGURE 3 Blind prediction results on hypothetical protein HI0442, *H. influenzae* (T0091). Our GDT_{TS} score as a function of distance cutoff for M1 (*dark blue*). We did not submit any other models for this target.

Target T0097: C-terminal domain of ERp29, rat

We submitted two models for target T0097 (Liepinsh, Mkrtchian, Barishev, Sharipo, Ingelman- Sundberg, Otting, submitted manuscript). Matching between the secondary structure of the models and the target is quite good as illustrated in Fig. 1. Unfortunately, Fig. 4 *a* shows that the overall packing of these secondary structure elements is incorrect with the first and the second helices adopting a mirror image of the respective helices in the target, the third and forth helices arranged correctly, and the fifth helix packing incorrectly against helices three and four. Our GDT_{TS} score is average when compared against all the models submitted for this target (Fig. 4 *b*). However, the overall RMSD of 10.2 Å for our M1, although not great, is comparable with those of the best scoring groups for this target. The RMSD for the C_α atoms between residue 123 and residue 164 (40% of the total number of residues) is 4.84 Å.

A possible problem with the prediction for target T0097 could be related to the solvation term of the energy function. Fig. 5 shows the hydrophobic (white) and hydrophilic (blue) patterning from two different (left and right), but equivalent views between the experimental structure (bottom) and our M1 prediction (top) of the target protein. We see that more hydrophobic surface is exposed in the experimental structure relative to our M1 prediction (left), but from a different view (right) our M1 shows a little more hydrophobic exposure. This seems to suggest that 1) the hydrophobicity term lacks specificity, and/or 2) the balance of stabilization of hydrophobic solvation against underlying AMBER force field is not optimal. In regard to point 1, we believe that our solvation function has some specificity in the sense that the greater number of aliphatic carbons that a side chain possesses, the greater amount of hydrophobic attraction there is between it and other hydrophobic amino acids. A clear indication of point 2 is the bending of the last helix

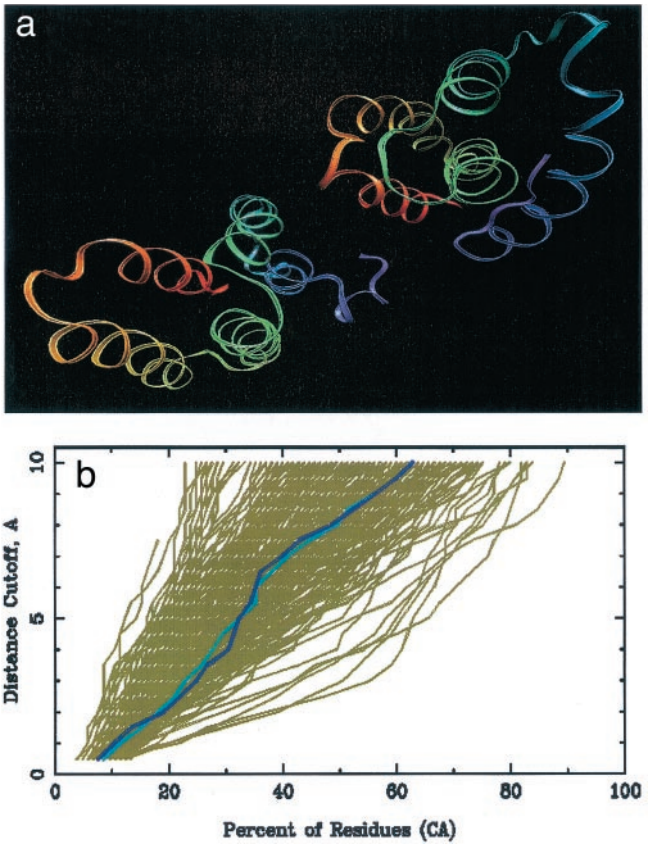


FIGURE 4 Blind prediction results on C-terminal domain of ERp29, rat, PDB code 1G7D (T0097). (*a*) Comparison of experimental nuclear magnetic resonance structure (*lower left*) and our M1 prediction (*upper right*). (*b*) Our GDT_{TS} score as a function of distance cutoff for M1 (*dark blue*) and M2 (*light blue*) with respect to all other submissions.

for our M1 (Fig. 4 *a*). This suggests that the energy function finds greater stabilization for a structure that slightly unravels or bends the helix, instead of packing against helix 3 and 4 as in the correct structure.

Our results for T0097 clearly point out the need for more diversity of starting points in early stages of the global optimization. To decrease the computational time, a single structure was created for the α-helical targets. This starting structure was formed by performing local minimizations using the biasing terms, but the buildup phase was omitted. A thorough analysis of all of the structures generated throughout all phases of the global optimization algorithm for target T0097, regardless of their energy value, reveals that the first two helices adopt essentially the same juxtaposition of order, i.e., we never sampled a configuration with the correct packing order of helix 1 and 2 because we always descended from parent populations that ordered them incorrectly. Although our method is designed to overcome such barriers, a single starting configuration may take longer runs that we could not afford during the CASP4 competition.

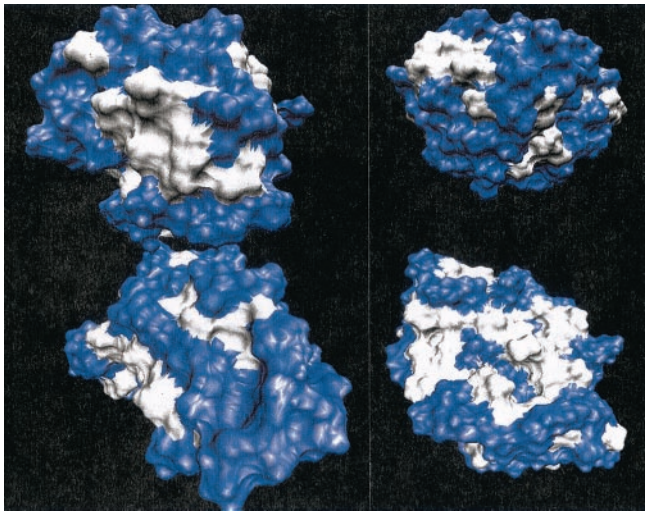


FIGURE 5 The patterning of two different views of C-terminal domain of ERp29, rat, PDB code 1G7D (T0097). The hydrophobic (*white*) and hydrophilic (*blue*) patterning from two different (*left* and *right*), views for the experimental structure (*bottom*) and our M1 prediction (*top*), of the target protein.

Target T0098: C-terminal domain of Spo0A, *B. stearothermophilus*

We submitted three models for T0098 (Lewis et al., 2000). Incorrect secondary structure predictions (prediction of a single α-helix rather than two short α-helices in two cases) had a negative impact on our results, which rely heavily on secondary structure information. The Q3 factor for M1 is 73.10, whereas the SOV is 68.90. Regarding tertiary structure prediction, the overall RMSD for the C_α atoms for M1 is 13.7 Å, and the longest segment with RMSD less than 5 Å is 39 (32% of the entire sequence.) Our relative performance is shown in Fig. 6.

After CASP4, we carried out a test in which we used perfect secondary structure prediction obtained from the experimental structure. The results, although not converged yet, reveal much better structures than our submitted models. This confirms our statement that the SPSC method relies heavily on secondary sructure predictions. Another problem that we observe is that, as with target T0097, tertiary structure predictions for this target were created with a single starting configuration. This results in structures that do not show a great deal of diversity.

Target T0099: a synthetic construct

Our submission for target T0099 was relatively poor in comparison to other predictions (Fig. 7 *a*) because this target had high sequence homology, meaning that it closely matched known proteins. For methods that use tertiary structure knowledge based on sequence or the Protein Databank (PDB), this was a rather easy target and was ranked

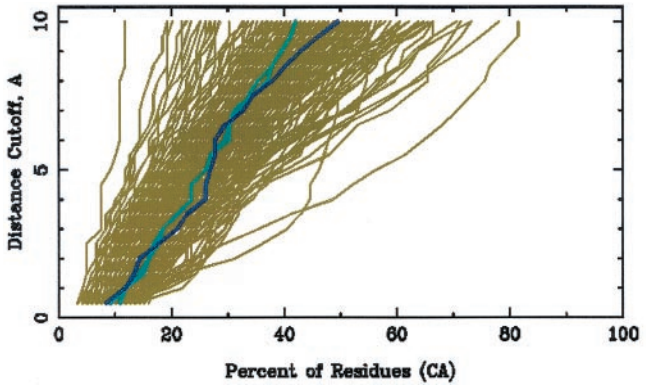


FIGURE 6 C-terminal domain of Spo0A, *B. stearothermophilus*, PDB code 1FC3 (T0098). Our GDT_{TS} score as a function of distance cutoff for M1 (*dark blue*) and M2 (*light blue*) with respect to all other submissions.

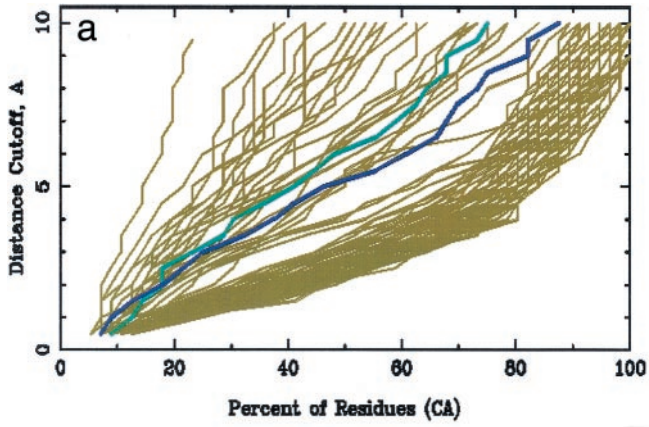
accordingly as shown in Fig. 2. However, this is a β-sheet protein, and while we did not get the overall topology right, without using any knowledge of sequence homology we predicted a good portion (30/54 amino acids) with an RMSD of under 5.0 Å. The overall RMSD for our target T0099 model 1 submission is 7.79 Å (see Fig. 7 *b*).

Target T0105: protein Sp100b, human

We submitted three models for this target. Secondary structure predictions were weak from both Jpred and Psi-Pred servers and substantially different between them. Therefore, we created our own predictions by combining the strongest predictions from the servers with our own consensus analysis from other neural network predictions. This resulted in secondary structure for M1 with a Q3 of 46.80 and an SOV of 25.90. However, secondary structure was better for M2 with a Q3 factor of 52.10 and a SOV of 39.70. Despite this problem, our group scored very well on this target (Fig. 8) with one of the best RMSDs over the entire structure (~11 Å for the three models) and one of the best GDT_{TS} scores (Fig. 2). Not surprisingly, M2 did better than M1 because it had better secondary structure manifested in phase 1. An interesting aspect of our predictions for this target is that we generated a number of distinct initial configurations by using the buildup algorithm (see Materials and Methods). Although the number of starting structures is small (only 10), it made a substantial difference creating a more diverse population of structures than in those cases in which we do not use the buildup phase.

Target T0110: ribosome-binding factor A, *H. influenzae*

This is the only target for which we submitted five models. This was due to two reasons: 1) we did not have time to run until convergence was achieved and 2) we did not have time



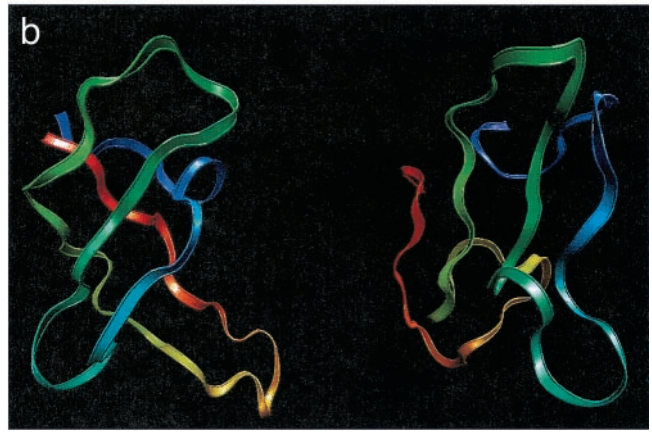


FIGURE 7 Blind prediction results (T0099). (*a*) Our GDT_{TS} score as a function of distance cutoff for M1 (*dark blue*) and M2 (*light blue*) with respect to all other submissions. (*b*) Comparison of experimental nuclear magnetic resonance structure (*left*) and our M1 prediction (*right*).

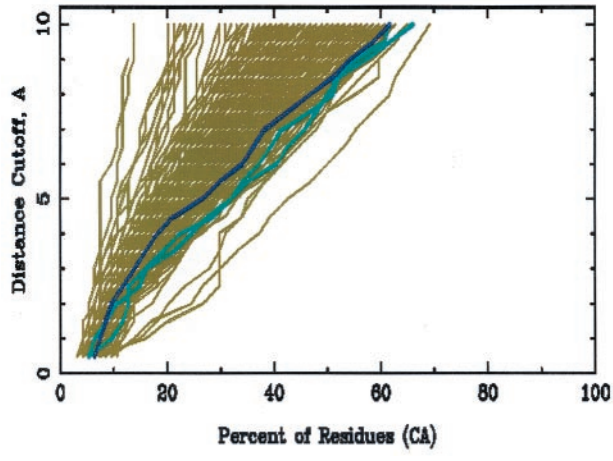


FIGURE 8 Blind prediction results on protein Sp100b, human (T0105). Our GDT_{TS} score as a function of distance cutoff for M1 (*dark blue*) and M2 (*light blue*) with respect to all other submissions.

to perform a full scale minimization in Cartesian coordinates, which usually changes the ordering of the structures according to their energy value. This was an important issue because we made the decision to submit the lowest energy value in internal coordinates for M1, which is not necessarily the same as the lowest energy structure in Cartesian coordinates. In fact, it was not. After submitting our results to CASP4, we performed the minimizations and found that our original M1 should have been M5, whereas the original M3 should have been M1.

We did very well in secondary structure for this target, improving substantially the poor predictions of Jpred and Psi-Pred. The Q3 factor for what ended up being M1 was 85.30 and the SOV was 86.30. For the structure submitted as M1 the Q3 was 76.80 and SOV was 69.50. We also did well in tertiary structure prediction, although not as well as with secondary structure. The overall RMSD for our submitted model 3 (actual M1) was 8.9 Å. The GDT_{TS} value for this model was also good (see Fig. 9). We hope that future runs of T0110 run until convergence is achieved will further improve these results.

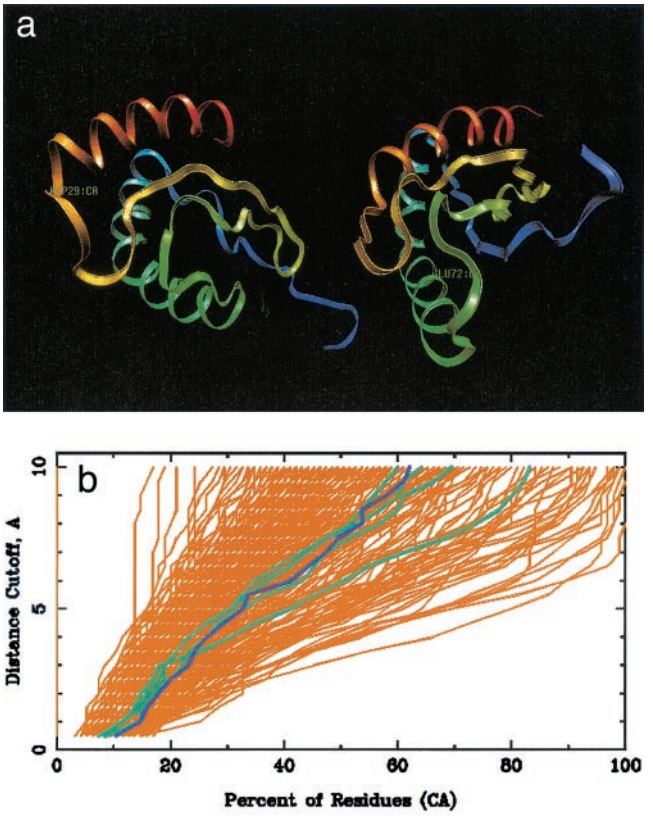


FIGURE 9 Blind prediction results on ribosome-binding factor A, *H. influenzae* (T0110). (*a*) Comparison of experimental nuclear magnetic resonance structure (*right*) and our M3 prediction (*left*). (*b*) Our GDT_{TS} score as a function of distance cutoff for M1 (*dark blue*) and M2 (*light blue*) with respect to all other submissions. Our M3 submission was our best.

Target T0124: phospholipase C beta C-terminus, turkey

Target 124 was considered to be one of the most difficult targets, or new fold, by the CASP4 organizers. It was also a difficult target from an optimization point of view with 242 amino acids, 4102 atoms, and over 12,000 Cartesian coordinates. Our M1 submission (Fig. 10 *a*) was among the best predictions for this target (Fig. 2) with the best GDT_{TS} score of any group (Fig. 10 *b*) and an overall RMSD of 8.46 Å. The secondary structure was very well formed for this target. Fig. 1 shows that some parts of the secondary structure of our models are closer to the target's secondary structure than what was predicted with a Q3 factor of 93.00 and an SOV of 81.00 for M1. Our longest continuous segment under cutoff RMSD of 5.0 Å was 99, which represents 40% of the entire structure and the best submitted at CASP4. The results for this target again show the value of a physics-based optimization method that does not rely on known protein structures for predicting proteins with new folds.

Our submission for target T0124 had not converged by typical standards of our global optimization algorithm. Typically when we have "converged" we see a contraction of the number of distinct structure clusters and a nonchanging energy of the lowest energy minimizer within each of these clusters. After the deadline for submission to CASP4, an additional run of phase 2 of SPSC was performed for target 124. This new run lowered the energy of the previous lowest energy minimizer from −4528 to −5095 kcal/mol, resulting in a new RMSD of 7.7 Å. Furthermore, in Fig. 11 we show the hydrophobic (white) and hydrophilic (blue) patterning for the experimental structure (center), our M1 prediction (right), and the next generation structure (left) were quite good. It is also apparent from the figure that the lowering of energy in the post-CASP4 run was due to further core packing.

Target T0125: Sp18 protein, *H. fulgens*

Target T0125 (N. Kresge, V. D. Vacquier, C. D. Stout, submitted manuscript) was considered to be a somewhat easy target by the CASP4 organizers. The secondary structure was very well predicted and subsequently well formed by our biasing phase. The deadline for submission for this target was simultaneous with the end of the competition, and we simply ran out of time to run until convergence was achieved, but it is interesting to see what type of structures are being found during early stages of the algorithm. We plan on continuing the runs necessary for convergence to ascertain our method's success on this target.

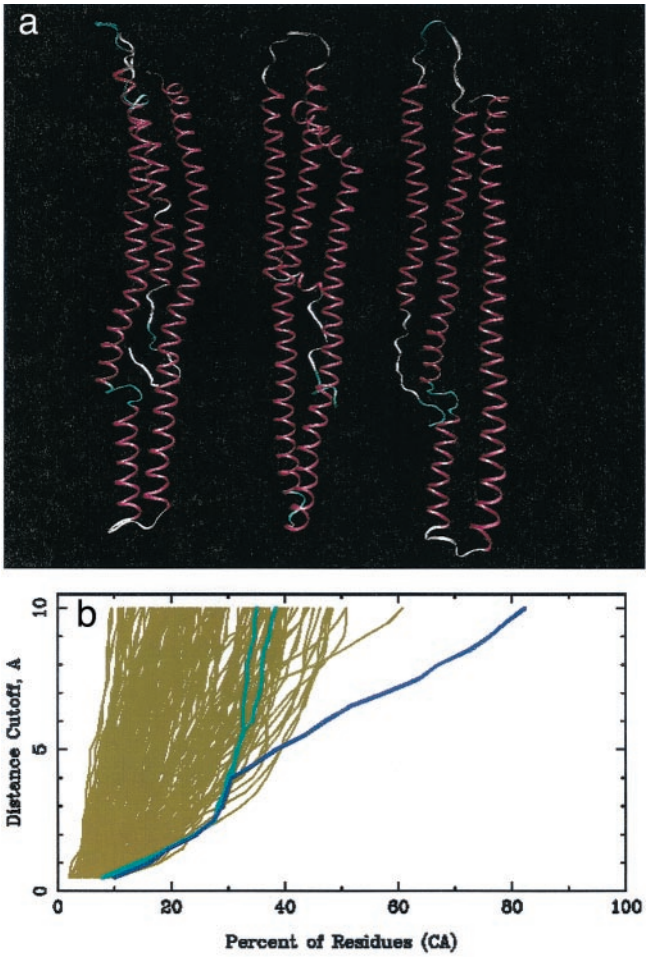


FIGURE 10 Results on phospholipase C β C-terminus, turkey (containing 242 amino acids). (*a*) Ribbon structure comparison between experiment (*center*), submitted M1 prediction (*right*), our lowest energy submission, and next generation run of the global optimization algorithm (*left*). This new run lowered the energy of our previous best minimizer, resulting in a new structure with an RMSD of 7.7 Å. (*b*) Our GDT_{TS} score as a function of distance cutoff for M1 (*dark blue*) and M2, M3 (*light blue*) with respect to all other submissions.

Computational costs associated with the CASP4 predictions

We give an estimate of the run times and amounts of computation for two CASP4 predicted targets, T0099 and T0124, representing respectively the smallest and largest targets we predicted. For target T0099, phase 1 (step 1a of algorithm 1) generated 10 starting configurations. A local minimization of ~12,000 steps using β-biasing (step 1b of algorithm 1) was applied to each. After phase 1, the 10 initial minima were clustered, resulting in three diverse clusters of which the lowest energy minima were passed on to phase 2. A total of 21 iterations of phase 2 were computed, nine of which used the balancing paradigm for choosing configurations to minimize, and 12 used the nonbalancing paradigm. The structure we submitted to CASP4 had the

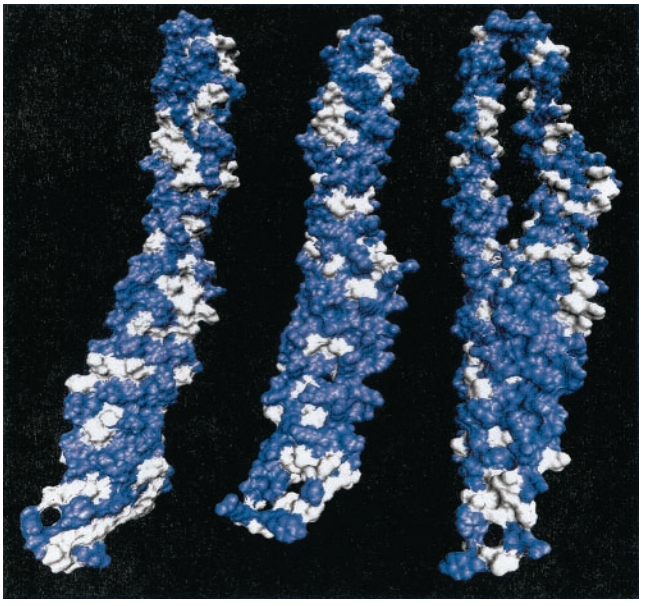


FIGURE 11 Results on phospholipase C β C-terminus, turkey (containing 242 amino acids). The hydrophobic (*white*) and hydrophilic (*blue*) patterning for the experimental structure (*center*), our M1 prediction (*right*), and the next generation structure (*left*) that indicates that the lowering of energy in the subsequent run was due to further core packing.

lowest overall energy, and the total computation time was roughly 36 h of elapsed time using 10 Avalanche (300 mHz) workstations at the University of Colorado. For target T0124, phase 1 used only one extended conformer as the starting configuration, and local minimization with α-biasing was done in portions over different segments because the configuration was too long to bias all the helix at once. It took ~25,000 iterations for all of the α-helices to form. Phase 2 ran for a total of 21 iterations (14 balancing and

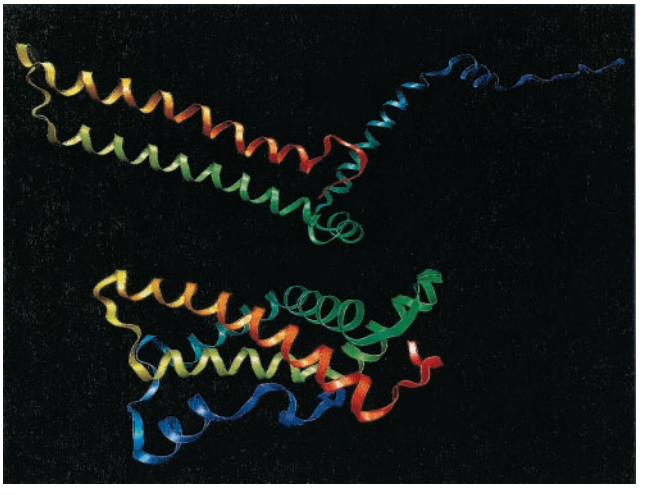


FIGURE 12 Results on Sp18 protein, *H. fulgens* PDB code 1GAK (T0125). (*a*) Comparison of experimental structure (*bottom*) and our M1 prediction (*top*).

TABLE 2 Discriminatory power of energy function

Target number	Model 1 prediction energy (kcal/mol)	Experimental structure energy (kcal/mol)
T0091	N/A	Coordinates not released
T0097	−2144	−2100
T0098	−2422	−2451
T0099	−856	−812
T0105	−1835	−1826
T0110	N/A	N/A*
T0124	−5115	−5142
T0125	N/A	N/A*

Evaluated energies (Eqs. 1 and 2) for our model 1 prediction versus experimental structure energies.

*Experimental structure had fewer number of amino acids than sequence given during CASP4. Because we made a prediction based on sequence given, our submitted structure has more amino acids than experimental target coordinates released after CASP4. This makes straightforward energetic comparisons between our models and experiment ambiguous and difficult.

seven nonbalancing), running for 36 h on 120 processors on the Cray T3E at NERSC.

Energy function discrimination

Further data that we have developed after the CASP4 conference was the energy ordering of our submitted predictions relative to the experimental structures. We were pleased to find that the experimental structures were as low or lower in energy than the models we submitted to CASP4. Table 2 provides the energies of our submitted models and the experimental structural for all targets for which we submitted predictions. Table 2 verifies that the energy function that includes our unique solvation model is behaving appropriately. Although we certainly propose to further improve the discriminatory power of our energy function in the proposed research, the overall energy discrimination between folds and misfolds was good.

DISCUSSION AND CONCLUSIONS

The area of protein structure prediction encompasses methods at two extremes: 1) reliance on a database of tertiary structures (bioinformatics) and 2) so called "ab initio" in which a (free) energy surface is formulated, and a search technique is developed to find the global minimum that is hypothesized to correspond to the native state structure (physical based). Practical implementations occur on a continuum between these two points of view, but our method is unquestionably placed much closer to 2 than to 1. When approaching protein structure prediction from a more physical view, then the approach should be judged based on the quality of the free energy surface, and the efficiency and effectiveness of the search technique. It is also a standard of the field that participation in the blind-prediction CASP contest is necessary to be credible, similar to the standard that code implementation must accompany theoretical assessments of scalabilty or speed of a new approach in the simulation or ab initio electronic structure areas.

An additional criterion for ab initio prediction is that a target's fold class or size should not be a limitation in the proposed method. In fact many ab initio prediction groups in CASP3 restricted their prediction to easy to medium targets that were primarily α-helical, and only in the most recent CASP4 (ending December 2000) did the ab initio field progress to do blind prediction more uniformly across all general classes of all-α, all-β, and α/β proteins. We too have extended the SPSC technique from only α-helical proteins to all general classes of protein structure that now includes β-sheet and mixed α/β proteins, which we present here for the first time.

An additional difference of our approach to others is the aqueous solvation description that is used to define the energy function. The solvation function is typically an implicit one in a search strategy such as this, and hence the complexity of the empirical solvation term we use is comparable with parameterized solvent accessible surface area terms. The functional form involves free energy of stabilization of hydrophobic groups in water at two length scales: at hydrophobic contact (like a solvent accessible surface area hydrophobic solvation term) and when they are separated by a water layer (unlike any empirical function used in prediction). This is motivated by our experimental work in hydrophobic solvation where we see scattering evidence of this new length scale, and is a well-defined physical solvation term (Sorenson et al., 1999). What we have learned about protein energetics is that there is an important additional length scale of the protein-solvent interaction that defines a mean force surface. We have further tested it to show that we can make reasonable predictions and very good predictions on the most difficult proteins of the CASP4 competition.

The performance of the SPSC algorithm at the CASP4 competition can be summarized as follows. 1) The method is more effective on targets for which less information from known proteins is available. In fact, as the difficulty of the targets increased, our group's percentile ranking increased as well. Our SPSC method produced the best prediction for one of the most difficult targets of the competition (T0124 a new fold protein of 240 amino acids). 2) The method's atom-based energy function and novel solvation function derived from experiments apparently did a very reasonable job of discriminating against misfolds from correct folds. We always submitted our lowest energy result as our first model, with second, third, etc. models always ranking the second, third, etc. highest in energy when clustered. 3) The method produced reasonable results with a very small number of initial configurations (1–10 compared with 1000-millions in other methods). This suggests that the global optimization algorithm used is more effective than the search mechanisms used by the other groups. 4) The method

takes a lot of computational time to converge, which makes it considerably more expensive than other knowledge-based methods that use more information from the databases. Our groups' overall performance was hurt by only making predictions on eight targets (of 17 new fold targets), of which only six were considered for the new fold category.

The performance of our SPSC method in the CASP4 competition is very encouraging, but there are several ways in which we believe our structure prediction approach can be improved significantly.

The first area of improvement is through further development and testing of our solvation function in the context of our global optimization approach to build in more specificity and therefore discrimination against energetically low-lying misfolds. During the CASP4 competition, we did not have time to test an appropriate balance or weighting of terms in Eqs. 1 and 2 to fully optimize a free energy function that can discriminate between correctly folded structures and those that are incorrectly folded. Overall the balance of the contributing forces seem reasonable because our lowest energy models scored in the 80 to 100 percentile of the CASP4 competition for targets T0091, T0105, T0110, and T0124, and our submitted energy values are lower, or are only slightly higher in energy, than the experimental structures (Table 2). However, we will modify our total free energy function to optimize a balance among the contributions from AMBER, the hydrophobic solvation function, and the addition of a more sophisticated treatment of electrostatics such as generalized born approaches (Onufriev et al., 2002). Decoy databases developed by Levitt and co-workers and other members of the protein and experimental communities (http://dd.stanford.edu/) provide means for testing of the energy function (Park et al., 1997; Park and Levitt, 1996).

Second, we will improve the SPSC approach for β-sheet and loop formation. We have achieved moderate success in matching pairs of β-strands via a preprocessing step described in Materials and Methods. The next step is to develop a more comprehensive approach for matching pairs of β-strands and sampling loop regions. We propose to experiment with variants of the biasing function to determine the most effective functional form and strategy for allowing strand-pairings to remain at least in close proximity, if not completely bonded. At the same time, it is essential to allow enough flexibility to obtain structural changes in the portions of the protein not predicted to have secondary structure, such as loops and turns.

A crucial property of any successful large-scale global optimization method is its ability to explore a diverse set of configurations. Our CASP4 results show that our global optimization method produced reasonable results with a very small number of initial configurations (typically 1–10), demonstrating that the global optimization is able to qualitatively improve on starting structures, even from a vastly under-represented population of search space. Our approach contains several techniques for accomplishing this, including the semirandom generation of initial configurations in phase 1, the explicit following of several paths during the balancing portion of phase 2, and the global perturbations of selected parameters in the small-scale global optimizations in phase 2. To improve our approach, we need to both explore new ways to generate structurally "different" configurations and ways to avoid redundant work on structurally similar configurations. Furthermore, we plan to significantly improve the computational efficiency to both increase our diversity of structures and to at least double the number of targets that we attempt in future CASP competitions.

Head-Gordon and Crivelli gratefully acknowledge support from the DOE/MICS program, U.S. Department of Energy contract number DE-AC-03-76SF00000, and the National Energy Research Supercomputer Center (NERSC) for significant T3E computer resources. Byrd, Eskow, and Schnabel gratefully acknowledge support from Air Force Office of Scientific Research grant F49620-00-1-0162, Army Research Office grant DAAG55-98-1-0176, and National Science Foundation grant CDA-9502956. S. Crivelli would like to thank Dr. Jon Sorenson for his invaluable help during the CASP4 competition. We thank Francesca Verdier and NERSC for needed cycles during the CASP4 competition.

REFERENCES

Azmi, A., R. H. Byrd, E. Eskow, R. Schnabel, S. Crivelli, T. M. Philip, and T. Head-Gordon. 2000. Predicting protein tertiary structure using a global optimization algorithm with smoothing. *In* Optimization in Computational Chemistry and Molecular Biology: Local and Global Approaches. C. A. Floudas, P. M. Pardalos, editors. Kluwer Academic Publishers, Dordrecht, The Netherlands. 1–18.

Byrd, R. H., T. Derby, E. Eskow, K. P. B. Oldenkamp, and R. B. Schnabel. 1994. A new stochastic/perturbation method for large-scale global optimization and its application to water cluster problems. *In* Large-Scale Optimization: State of the Art. W. Hager, D. Hearn, P. Pardalos, editors. Kluwer Academic Publishers, Dordrecht, The Netherlands. 69–81.

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

Crivelli, S., T. Head-Gordon, R. H. Byrd, E. Eskow, and R. Schnabel. 1999. A hierarchical approach for parallelization of a global optimization method for protein structure prediction. *In* Lecture Notes in Computer Science, Euro-Par '99, P. Amestoy, P. Berger, M. Dayde, I. Duff, V. Fraysse, L. Giraud, D. Ruiz, editors. 578–585.

Crivelli, S., T. Philip, R. Byrd, E. Eskow, R. Schnabel, R. Yu, and T. Head-Gordon 2000. A global optimization strategy for predicting protein tertiary structure: a-helical proteins: proceedings for new trends in computational methods for large molecular systems. *Comput. Chem.* 24:489–497.

Cuff, J. A., M. E. Clamp, A. S. Siddiqui, M. Finlay, and G. J. Barton. 1998. Jpred: a consensus secondary structure prediction server. *Bioinformatics.* 14:892–893.

Eyrich, V. A., D. M. Standley, A. K. Felts, and R. A. Friesner. 1999. Protein tertiary structure prediction using a branch and bound algorithm. *Proteins Struct. Funct. Genet.* 35:41–57.

Gibson, K. D., and H. A. Scheraga. 1987. Revised algorithms for the build-up procedure for predicting protein conformations by energy minimization. *J. Comp. Chem.* 8:826–834.

Head-Gordon, T., J. Arrecis, and F. H. Stillinger. 1991. A strategy for finding classes of minima on a hypersurface: implications for approaches to the protein folding problem. *Proc. Natl. Acad. Sci. U.S.A.* 88:11076–11080.

Head-Gordon, T., J. M. Sorenson, A. Pertsemlidis, and R. M. Glaeser. 1997. Differences in hydration structure near hydrophobic and hydrophilic amino acid side chains. *Biophys. J.* 73:2106–2115.

Head-Gordon, T., and F. H. Stillinger. 1993. Predicting polypeptide and protein structures from amino acid sequence: antlion method applied to melittin. *Biopolymers.* 33:293–303.

Hura, G., J. M. Sorenson, R. M. Glaeser, and T. Head-Gordon. 1999. Solution x-ray scattering as a probe of hydration-dependent structuring of aqueous solutions. *Perspect. Drug Discov. Des.* 17:97–118.

Jones, D. T. 1999. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* 292:195–202.

Lee, J., A. Liwo, D. Ripoll, J. Pillardy, and H. Scheraga. 1999. Calculation of protein conformatio by global optimization of a potential energy function. *Proteins Struct. Funct. Genet.* [Suppl.] 3:204–208.

Lewis, R. J., S. Krzywda, J. A. Brannigan, J. P. Turkenburg, K. Muchová, E. J. Dodson, I Barák, and A. J. Wilkinson. 2000. The trans-activation domain of the sporulation response regulator Spo0A revealed by X-ray crystallography. *Mol. Microbiol.* 38:198–212.

Moult, J., T. Hubbard, K. Fidelis, and T. Pedersen. 1999. Critical assessment of methods of protein structure prediction CASP: round III. *Proteins Struct. Funct. Genet.* [Suppl.] 3:2–6.

Novotny, J., R. E. Bruccoleri, and M. Karplus. 1984. An analysis of incorrectly folded protein models: implications for structure prediction. *J. Mol. Biol.* 177:787–818.

Onufriev, A., D. Bashford, and D. A. Case. 2002. Modification of the generalized born model suitable for macromolecules. *J. Phys. Chem. B.* In press.

Orengo, C. A., J. E. Bray, T. Hubbard, L. LoConte, and I. Sillitoe. 1999. Analysis and assessment of ab initio three-dimensional prediction, secondary structure, and contacts prediction. *Proteins Struct. Funct. Genet.* [Suppl.] 3:149–170.

Ortiz, A., A. Kolinski, P. Rotkiewicz, B. Ilkowski, and J. Skolnick. 1999. Ab initio folding of proteins using restraints derived from evolutionary information. *Proteins Struct. Funct. Genet.* [Suppl.] 3:177–185.

Park, B., E. S. Huang, and M. Levitt. 1997. Factors affecting the ability of energy functions to discriminate correct from incorrect folds. *J. Mol. Biol.* 266:831–846.

Park, B., and M. Levitt. 1996. Energy functions that discriminate X-ray and near native folds from well-constructed decoys. *J. Mol. Biol.* 258: 367–392.

Pertsemlidis, A., A. M. Saxena, A. K. Soper, T. Head-Gordon, and R. M. Glaeser. 1996. Direct, structural evidence for modified solvent structure within the hydration shell of a hydrophobic amino acid. *Proc. Natl. Acad. Sci. U.S.A.* 93:10769–10774.

Pratt, L. R., and D. Chandler. 1977. Theory of the hydrophobic effect. *J. Chem. Phys.* 67:3683–3704.

Rinnooy Kan, A. H. G., and G. T. Timmer. 1984. Stochastic methods for global optimization. *Am. J. Math. Management Sci.* 4:7–40.

Samudrala, R., Y. Xia, E. Huang, and M. Levitt. 1999. Ab initio protein structure prediction using a combined hierarchical approach. *Proteins Struct. Funct. Genet.* [Suppl.] 3:194–198.

Shortle, D. 2000. Prediction of protein structure. *Curr. Biol.* 102:1–10.

Simons, K. T., R. Bonneau, I. Ruczinski, and D. Baker. 1999. Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins Struct. Funct. Genet.* [Suppl.] 3:171–176.

Sorenson, J. M., G. Hura, A., K., Soper, A. Pertsemlidis, and T. Head-Gordon. 1999. Determining the role of hydration forces in protein folding. *J. Phys. Chem. B.* 103:5413–5426.

Zemla, A., C. Venclovas, K. Fidelis, and B. Rost. 1999. A modified definition of SOV, a segment-based measure for protein secondary structure prediction assessment. *Proteins Struct. Funct. Genet.* 34: 220–223.

Zemla, A., C. Venclovas, A. Reinhardt, K. Fidelis, and T. Hubbard. 1997. Numerical criteria for the evaluation of ab initio predictions of protein structure. *Proteins Struct. Funct. Genet.* [Suppl.] 1:140–150.

Zhu, H., and W. Braun. 1999. Sequence specificity, statistical potentials, and three-dimensional structure prediction with self-correcting distance geometry calculations of beta-sheet formation in proteins. *Protein Sci.* 8:326–342.