Exploratory Studies of Ab Initio Protein Structure Prediction: Multiple Copy Simulated Annealing, AMBER Energy Functions, and a Generalized Born/Solvent Accessibility Solvation Model

Yongxing Liu and D. L. Beveridge*
Chemistry Department and Molecular Biophysics Program, Wesleyan University, Middletown, Connecticut

ABSTRACT **A theoretical and computational approach to ab initio structure prediction for polypeptides in water is described and applied to selected amino acid sequences for testing and preliminary validation. The method builds systematically on the extensive efforts applied to parameterization of molecular dynamics (MD) force fields, employs an empirically well-validated continuum dielectric model for solvation, and an eminently parallelizable approach to conformational search. The effective free energy of polypeptide chains is estimated from *AMBER* united atom potential functions, with internal degrees of freedom for both backbone and amino acid side chains explicitly treated. The hydration free energy of each structure is determined using the Generalized Born/Solvent Accessibility (GBSA) method, modified and reparameterized to include atom types consistent with the *AMBER* force field. The conformational search procedure employs a multiple copy, Monte Carlo simulated annealing (MCSA) protocol in full torsion angle space, applied iteratively on sets of structures of progressively lower free energy until a prediction of a structure with lowest effective free energy is obtained. Calibration tests for the effective energy function and search algorithm are performed on the alanine dipeptide, selected protein crystal structures, and united atom decoys on barnase, crambin, and six examples from the Rosetta set. Specific demonstration cases of the method are provided for the 8-mer sequence of Ala residues, a 12-residue peptide with longer side chains QLLKKLLQQLKQ, a de novo designed 16 residue peptide of sequence (AAQAA)$_3$Y, a 15-residue sequence with a β sheet motif, GEWTWDATKTFTVTE, and a 36 residue small protein, Villin headpiece. The Ala 8-mer readily formed an α-helix. An α-helix structure was predicted for the 16-mer, consistent with observed results from IR and CD spectroscopy and with the pattern in ψ/φ angles of known protein structures. The predicted structure for the 12-mer, composed of a mix of helix and less regular elements of secondary structure, lies 2.65 Å RMS from the observed crystal structure. Structure prediction for the 8-mer β-motif resulted in form 4.50 Å RMS from the crystal geometry. For Villin, the predicted native form is very close to the crystal structure, RMS values of 3.5 Å (including sidechains), and 1.01 Å (main chain only). The methodology permits a detailed analysis of the molecular forces which dominate various segments of the predicted folding trajectory. Analysis of the results in terms of internal torsional, electrostatic and van der Waals and the electrostatic and non-electrostatic contributions to hydration, including the hydrophobic effect, is presented. Proteins 2002;46:128–146.** © 2001 Wiley-Liss, Inc.

INTRODUCTION

Anfinsen's experiment[1] on the reversible denaturation of ribonuclease demonstrated that the tertiary structure of proteins in solution may be determined by the amino acid sequence, and extensive subsequent experiments have demonstrated thermodynamic reversibility for many small, single domain proteins[2] and a number of more complex cases (for a recent survey and discussion see Dill[3]). These results gave rise to the idea known today as the "thermodynamic hypothesis," i.e., that the native form of a protein corresponds to global minimum on the conformational free energy surface. This, in turn, implies that a purely theoretical/computational approach to protein structure prediction from amino acid sequence should be feasible, at least in principle. In practice, the difficulty of determining accurate energy functions, the essentially rugged character of the free energy landscape, and the dimensionality of the configuration space of a polypeptide in solution, makes ab initio prediction of protein structure from sequence a very challenging problem. The field of ab initio structure prediction has spawned a large literature (for a recent review see Osguthorpe[4]), and is featured prominently in the CASP series of blind prediction contests designed to objectively document progress in protein structure prediction by various methods.[5]

A problem at the opposite end of the computational spectrum is the dynamics of proteins in the vicinity of a known equilibrium structural form (such as that obtained

Grant sponsor: NIH; Grant number: GM 37909.

*Correspondence to: D. L. Beveridge, Chemistry Department and Molecular Biophysics Program, Wesleyan University, Middletown, CT 06457. E-mail: dbeveridge@wesleyan.edu

Received 26 January 2001; Accepted 7 August 2001

DOI 10.1002/prot.10020

from crystallography). This problem has been treated successfully via molecular dynamics (MD) simulation based on all-atom empirical energy functions and Newtonian equations of motion.[6,7] Recent initiatives in this area have produced considerable refinements in MD energy functions.[8–10] MD based on these force fields has been used effectively in the investigation of protein unfolding pathways in solution.[11] The longest MD on polypeptides in water are the 50 ns trajectory on a β-heptapeptide by Daura et al.[12] and a 1.2 microsecond trajectory on Villin[13] (see also Lee et al.[14]). However, protein structure prediction directly from MD starting from a denatured form is not immediately feasible, due to (1) the limited time scale (approximately nanoseconds) currently accessible to this technique and (2) the general inability of MD sampling to escape from metastable local minima. The problem is only compounded by the need to include solvation, which increases the degrees of freedom that must be treated in a given calculation. The results of Duan et al.[13] on villin provide a critical demonstration case of the problem with MD folding. Lee et al.[14] have used a scoring function on Villin MD structures with some success at discriminating the native form.

The scoring function in ab initio protein structure prediction is an "effective" free energy, i.e., effective in the sense that some degrees of freedom are by necessity integrated out.[15,16] The use of MD potentials in protein structure prediction schemes was discussed at the 1998 CASP3 meeting[5] as a refinement tool, applied mainly to predictions from knowledge based methods with the hope of further improvements. Early reports on this strategy ("introducing the physics") did not show dramatic success. However, solvent effects were either not included explicitly in the models, or not estimated accurately enough. Since it seems reasonable for polypeptide and protein structure prediction schemes to build systematically on the considerable work invested in the general formulation and extensive parameterization efforts involved in developing MD force fields, we have proceeded to explore this approach, including solvent in the model. In this project, we explore the feasibility of transferring the energy functions designed for protein MD simulations over to protein structure prediction, treating the effective free energy of hydration using Generalized Born (GB) theory for electrostatics and solvent accessibility (SA) calculations for non-electrostatic components. GBSA calculations[17] have proved to be quite accurate approximations to more rigorous theoretical methods and agree well with observed values.[18–21] Combined, the MD energy functions for the internal energy and the GBSA method for solvation provide a well-defined effective free energy function for scoring the relative stability of polypeptide chain conformations in water, and a possibly viable point of departure for systematic studies of polypeptide structure (where applicable) and protein structure prediction.

To make ab initio protein structure prediction tractable, one requires a search engine that is rapid but not vulnerable to getting trapped in the rugged features of the configurational free energy surface. In this study, increased speed (compared with MD), is obtained by reducing the degrees of freedom in the prediction scheme to torsional motions. We treat the sampling problem by a multiple copy Monte Carlo Metropolis simulated annealing (MCSA) approach,[22] implemented here in a iterative protocol in which progressively lower free energy structures are periodically culled and form the basis for successive refinements. This computational strategy takes advantage of the "free energy funnels" by which proteins in solution are thought to overcome the Levinthal paradox.[23–25] Examination of snapshots of intermediary structures along the prediction trajectory can be used to determine the relative importance of factors such as hydrophobic collapse and helix nucleation. At a more detailed level, determination of the relative contributions from internal torsional, electrostatic and van der Waals energies and the electrostatic and non-electrostatic contributions to hydration, (including the hydrophobic effect) forms a basis for assessment of relative importance of diverse chemical forces in structure prediction trajectories and folding processes.

Subsequent to CASP3, Simmerling et al.[26] demonstrated preliminary success with MD refinement of knowledge based structure predictions and a molecular mechanics/Poisson Boltzmann (MMPB) scoring function. In studies more closely related to those described herein, several groups have been pursuing protein structure predictions using MD energy functions and implicit solvent models, each with a unique and independent implementation.[27,28] Our particular emphasis in the present project is on a fully ab initio approach, wherein the only input data is the amino acid sequence of the polypeptide with no knowledge based information such as helix propensities, fold recognition or homology modeling involved. Obviously, a scheme involving additional strategic and particularly knowledge based elements could be devised that would be more practical and efficient, but our aim here was to see how far we could get with strict observance of the Anfinsen "sequence determines structure" paradigm with MD potential functions. A second emphasis in this project is to determine the contributions to the effective free energy from the internal torsional, electrostatic and van der Waals energies, and the electrostatic and non-electrostatic contributions to hydration (including the hydrophobic effect), assess the relative importance of various chemical forces at various stages of prediction trajectories, and examine the extent to which this model performs according to conventional wisdom(s) about protein folding. We anticipate that larger and more complex structures, particularly those involving multiple β-strands, will be difficult for this (or any other fully ab initio method) to predict. Here we expect that a systematic study of a series of modestly sized case studies of increasing difficulty, with the capabilities and limitations of the methods fully exposed, will be a valuable benchmark for further methodological improvements.

BACKGROUND

Protein Structure Prediction

Recent progress in the area of protein structure prediction and the related protein folding problem is documented in a number of recent articles and reviews.[5] Studies of this type are motivated by a pressing need for accurate structure predictions on the large number of protein sequences is identified in genomics initiatives, with the ultimate objective of inferring function from structural homologies. There are also questions about the fundamental nature of protein folding to be solved, such as the role of funnels vs. pathways in the route from unfolded structures to a native form, and the relative importance and timing of phenomena such as nucleation of secondary structural elements, hydrophobic collapse, and the role of molten globule forms and other structures as possible intermediates in the folding process.

Research in protein structure prediction is being pursued in the field along two major lines of investigation. One approach is "comparative" in nature,[29] with information from known structures systematically and strategically incorporated into prediction algorithms. Two classes of comparative structure prediction methods are homology modeling[29] and threading,[30] which in some cases provide a basis for "knowledge based potentials" that are supplied to search engines for optimization. Some concerns have been expressed about the meaning and significance of these functions.[31] The second approach, responding to the challenge posed by the Anfinsen experiment, is the ab initio prediction of structure simply from a knowledge of the corresponding amino acid sequence and specific interaction energies implied by the sequence.[4] In ab initio folding, one must deal with the construction of reliable potentials describing the interactions within the polypeptide backbone and various amino acid side chains on one hand, and with water and possibly added salt on the other. In most approaches, a search engine must be applied to a scoring function to obtain predictions of the native form of the protein. The empirical nature of comparative methods obviates obtaining reliable new knowledge about the molecular physics of structural stability and folding processes. A fully ab initio approach, if demonstrably successful at accounting for structures, can be subjected to analyses that can reveal the relative importance of various intrinsic energies and solvent effects at various stages of the calculation and thus provide interpretation as well as prediction.

Ab Initio Potential Functions

A number of schemes for deriving empirical potential functions for proteins as a function of conformation have been set forth.[15,16,32,33] Included among these are lattice based potentials[34] and united residue potentials with one or more virtual atoms representing each amino acid side chain[32] as well as the minimalist Geocore approach.[35] While these potentials have been used successfully to predict a number of structures, in most cases considerable atomic details are neglected. Thus, there is the possibility that the physics is compromised to some extent, since late stage folding is likely to involve a subtle balance of diverse forces.[3] If parameterization of a potential is based on known protein structures, the strictly ab initio character of the approach is compromised. Nevertheless, an accurate prediction tool derived on this basis is of considerable interest and utility, and can also be incorporated into hybrid approaches. Meanwhile, considerable efforts have been applied to the development of all-atom potential functions for protein atoms and solvent molecules for use in MD simulations. Empirical potential functions developed by research collaboratories that have nucleated around the AMBER,[8] CHARMM,[9] and GROMOS[36] MD and molecular modeling programs are parameterized mainly on the basis of experimental data and quantum mechanical calculations on prototypes of macromolecular constituents, which are then presumed to transfer to macromolecules in solution. The accuracy with which such potentials perform on the equilibrium structure and dynamics of proteins[7,37] and also nucleic acids[38] is well documented. Solvent molecules are typically included via explicit representations in the MD modeling, which makes studies of this genre quite computationally intensive.

The use of all-atom potentials and an explicit model for solvation is currently out of the question for high throughput protein structure prediction. However, solvent water plays an important role in protein stability and thus a good model for solvation is expected to be essential to accurately represent the physics in an ab initio modeling approach. A class of solvent models in which water is treated as a polarizable dielectric continuum and the electrostatics of solvation is computed via the Poisson Boltzmann (PB) equation has received considerable recent attention.[39] A number of variations on continuum solvent models are being pursued, including full finite difference Poisson Boltzmann (FDPB) calculations[40] and the ACES method of Lazaridis and Karplus.[16] Several recent studies have demonstrated that the more simplified Generalized Born model ,[17,21] well parameterized, can be used as a rapid estimator of FDPB results, which in turn agree well with results on solvation free energy computed using the perturbation method applied to fully explicit molecular dynamics simulations. Further modifications have brought this model into line for modeling both solvation energy and pKa shifts simultaneously,[41] correcting a subtle inconsistency in earlier versions of the method. The GBSA method provides agreement to ~5% for calculated and observed free energies of hydration for small molecules and molecular ions.[42] A combined GBSA-AMBER potential was used to predict polypeptide loop geometries[43] and in a study closely parallel to this work, GBSA-OPLS-AA potential is being used to score protein folding decoys.[28] In passing, we note several current initiatives underway in which MDs on proteins and nucleic acids are being performed in a GB solvent.[19,44]

In this project, an effective free energy function is constructed from MD potentials, modified GB approach for computing the conformation dependent electrostatic free energy of polypeptide structures and protein solvation, and the solvent accessibility (SA) method to estimate the non-electrostatic contribution. The latter includes both solvent entropy (cavitation) and solute-solvent van der Walls interactions and is capable of providing quantitative estimates of the hydrophobic effect. Moreover, the effect of added salt can be included in GB, following Jayaram et al.[41] and Tsui and Case,[19] by adding a Debye-Huckel term to the solvation energy. This extends the scope of the modeling to include environmental conditions closer to that found in vivo if desired.

Conformational Sampling

In a "brute force" conformational search method for protein structure prediction, all conformational space would be randomly sampled in order to find the global minima. It is, however, well known that protein (not to mention solvent) conformational space is too large to be fully sampled in a reasonable time scale, and finding the global minimum is never guaranteed. Proteins fold to a native form reliably in the order of seconds in spite of this problem,[45] presumably by utilizing funnels on the free energy landscape. For recent relevant commentary see the reviews by Dill and Chan[24], Simmerling et al.,[26] Karplus,[25] and Honig.[47] Diverse approaches have been described in the literature designed for treating the sampling problem in protein structure prediction calculations (for a current review, see Hansmann and Okamoto[48]). Recent developments include genetic algorithms,[49] conformational space annealing,[50] Newton Operator Torsional Dynamics,[51] statistics biased Monte Carlo conformational search,[52] self-guided molecular dynamics simulation,[53] and the convex global under-estimator (CGU) method.[54] Some methods go beyond the random search paradox via the explicit idea of a protein folding free energy landscape and funnel theory[55]; Honig[47] has recently described how this idea is in fact implicit in most if not all of the current approaches. The multicopy MCSA protocol has reasonable prospects of dealing with problems inherent in the ruggedness of the free energy landscape. Studies on prototype cases show impressive efficiency on the optimization of a simple mathematical function representative of the protein folding problem at low dimensionality.[22] Simulated annealing of polypeptide fragments is a key component of the currently most successful of knowledge based approaches to protein structure prediction in the CASP3 and CASP4 structure prediction contests[56] (see also Kawai et al.[57]).

METHODS

The structure prediction protocol used in this study, which utilizes an effective free energy scoring function based on *AMBER* and GBSA, and a multiple copy MCSA search engine is implemented in this laboratory in a program called *REFOLD*. Details of the *REFOLD* protocol are as follows.

Conformational Free Energy

Conformations of the polypeptide backbone and amino acid side chains are developed in Cartesian coordinates for energy evaluation and in internal coordinates for sampling protocols. The two sets of coordinates are, of course, rapidly interconvertable during the calculation. The internal coordinate set is composed bond lengths, bond angles, dihedral angles, and improper dihedrals, including three dummy atoms for the optimization. We make the usual separation of hard modes and soft modes, and assume the bond lengths and bond angles of the polypeptide to be fixed. Conformation sampling is thus carried out exclusively in torsion angle space. We construct an effective free energy ΔG^{eff} for a polypeptide chain in solution factors as

$$\Delta G^{\text{eff}} = \Delta G_{\text{int}} + \Delta g_{\text{sol}} \tag{1}$$

where ΔG_{int} is the free energy intrinsic to the polypeptide per se and Δg_{sol} is the solvation free energy. Note uppercase in the rhs of equations denotes quantities intrinsic to the polypeptide chain per se and lower case is used to distinguish solvation quantities.

Scoring Intrinsic Free Energy

At constant temperature, the intrinsic free energy can be expanded as

$$\Delta G_{\text{int}} = \Delta U_{\text{int}} - T\Delta S_{\text{int}} \tag{2}$$

where ΔU_{int} and ΔS_{int} are the internal energy and entropy, respectively. We neglect the effects of configurational averaging on internal energy, and estimating it from corresponding values on the Born Oppenheimer Energy surface, vis.

$$\Delta U_{\text{int}} \cong \Delta E_{\text{int}} \tag{3}$$

An internal entropy for a structure can be estimated by the quasiharmonic method.[58] As noted by Lazardis and Karplus, the vibrational entropy of a folded protein is large but there is evidence that it is similar to that of any single unfolded conformer (see also Lee et al.[14]). The large entropy of the unfolded state follows from the multiplicity of conformers of similar energy, which is important to thermodynamics but not a problem for scoring functions or prediction trajectories. Thus, we follow the practice of neglecting the entropy contribution from the internal motions of the polypeptide chain. The intrinsic energy ΔE_{int} is obtained from AMBER united atom empirical potential functions.[59,60] With bond lengths and angles fixed, the internal energy is composed of dihedral, electrostatics and VDW contributions,

$$\Delta E_{\text{int}} = \Delta E_{dihedrals} + \Delta E_{es} + \Delta E_{vdW} \tag{4}$$

where

$$E_{dihedrals} = \sum_{dihedrals} \frac{V_n}{2}\,[1 + \cos(n\phi - \gamma)], \tag{5}$$

$$E_{el} = \sum_{i<j} \frac{q_i q_j}{\varepsilon R_{ij}} \tag{6}$$

$$E_{vdw} = \left[\left(\frac{A_{ij}}{R_{ij}^{12}}\right) - \left(\frac{B_{ij}}{R_{ij}^{6}}\right)\right] \tag{7}$$

where the definition of terms and force field parameters are just those listed by Weiner et al.[59] With these approximations, the effective free energy function energy function reduces to

$$\Delta G^{\text{eff}} \cong \Delta E_{\text{int}} + \Delta g_{\text{sol}} \tag{8}$$

Scoring Solvation Free Energy

The solvation free energy is divided further into two contributions, electrostatic (el) and nonelectrostatic (nel). The electrostatic component of solvation energy is calculated from modified GB model.[41] Various uses of the GB model for the calculation of solvation energy of a molecule have been extensively documented.[18] The defining equations are

$$\Delta g_{el} = \sum_{i=1}^{n} [(\Delta g_{scr})_i + (\Delta g_{pol})_i] \tag{9}$$

where

$$(\Delta g_{scr})_i = -\frac{1}{2} \times 166\left(1 - \frac{1}{\varepsilon}\right) \sum_{j=1}^{n} \frac{q_i q_j}{f_{GB}} \tag{10}$$

$$(\Delta g_{pol})_i = -166\left(1 - \frac{1}{\varepsilon}\right) \frac{q_i^2}{\alpha_i} \tag{11}$$

with $f_{GB} = (r_{ij}^2 + \alpha_{ij}^2 e^{-D})^{0.5}$, $D = r_{ij}^2/(2\alpha_{ij}^2)$, $\alpha_{ij} = (\alpha_i \alpha_j)^{0.5}$ and ϵ is the dielectric constant. The Born radii α_i are calculated using formulae of Cramer and coworkers.[61] However, only a set of small molecules with energy minimized structures were used for previous parameterization. In this work, we use the peptide itself and the MD ensemble as the training set and obtain parameters for each of the *AMBER* atomic types. Structures for this purpose were obtained either from MD simulation or random generation of conformations in the torsional space. In case of random generation, some conformations close contacts are screened out. For each structure in the training set, the electrostatic portion of the solvation free energy was calculated using the FDPB method in program *DELPHI II* . In the FDPB calculation, partial charges and atomic radii are obtained from *AMBER* parameter file. The grid size used was 0.25 Å. With the *DELPHI* values as the ideal values, a simulated annealing optimization procedure was applied to obtain the atomic type parameters. The results are given in Table I.

The non-electrostatic contributions to the solvation free energy is calculated from solvent accessible area for the various atoms, SA_i. The solvent accessible surface area is the sum of all atomic contributions and is amenable to rapid calculation[17,62] as

$$SA_t = S_i \prod_{\substack{j=1 \\ j \neq 1}}^{N} (1.0 - P_i P_{ij} P_{ijk}/S_i) \tag{12}$$

where P_i is atomic parameter,

$$S_i = 4\pi(r_i + r_s)^2 \tag{13}$$

TABLE I. GBSA Parameters Used in This Work for United Atom Types in Amber Force Field

Atomic type	MGB	ASA
C3	0.968	2.847
C	0.913	1.918
O	0.980	1.237
N	0.848	1.210
H	0.914	1.316
CH	0.926	3.000
C2	1.021	2.999
OH	0.898	1.333
HO	0.904	0.822
N3	0.822	0.536
H3	0.919	0.528
N2	0.879	1.532
CA	0.927	2.871
CC	1.016	2.999
NA	0.828	1.377
CP	0.934	3.000
NB	0.926	2.115
CF	0.892	2.997
O2	0.995	1.915
SH	1.011	1.801
HS	0.955	1.167
LP	0.985	0.300
S	0.990	0.638
CD	0.898	1.770
C*	0.917	2.593
CG	0.962	2.688
CN	0.906	2.647
CB	0.922	2.576

C* = sp^2 aromatic carbon in 5-membered ring with one substitute.

$$P_{ij} = \pi(r_i + r_s)(r_j + r_i + 2r_s - r_{ij})\left(1.0 + \frac{(r_i - r_j)}{r_{ij}}\right) \tag{14}$$

and

$$p_{ijk} = \prod_{k} \frac{r_{jk}^2}{r_{ij}^2} \tag{15}$$

where the product function is over all atoms bonded to atom i.

For GBSA parameterization purposes, three model peptides were constructed that include all 20 amino acids in a random order and solvated in a box of TIP3P water molecules. After minimization and equilibration, 1 ns MD simulation is run using AMBER 5.0 molecular dynamics simulation package and the united atom force field.[59] From the trajectory of each molecule, 1,000 conformations were extracted as the training set for the parameterization. For each conformation in the training set, the electrostatic part of the solvation free energy was calculated using Poisson Boltzmann (PB) method by program Delphi.[63] The solvent accessible area is calculated by the slicing algorithm using Molecular Dynamics Tool Chest.[64] The GB parameters were fit to the PB results. The ASA parameters are fit to the numerical slicing algorithm using simulated annealing optimization. The results are listed in Table I.

Analysis of Results

For analysis and interpretation of results, the formulation of configurational free energy as described above is readily decomposed into various physically meaningful contributions. The intrinsic term can be divided into torsional, electrostatic, and van der Waals contributions as per Equations 1–7. The solvation energy is likewise decomposed into electrostatic and non-electrostatic contributions, such as

$$\Delta g_{nel} = \gamma \cdot SA \tag{16}$$

Here, the parameter γ is set to 7.2 cal/Å^2 based on calibration from experimental data[17] on the solubility of aliphatic alcohols. The parameter γ is composed of two parts,

$$\gamma = \gamma_{\mathrm{cav}} + \gamma_{\mathrm{vdW}} \tag{17}$$

which define the contributions to the non-electrostatic free energy from cavitation effects and solute-solvent van der Waals interactions. Values for these quantities suggested by Jayaram et al.[65] are $\gamma_{\mathrm{VDW}} = +47$ cal/Å^2[66] and $\gamma_{\mathrm{cav}} = -39.8$ cal/Å^2.

Multicopy Simulated Annealing

All structure predictions begin with an extended form of the polypeptide chain as the initial configuration. From this structure, a set of n replicas is spawned, Each is subjected to an independent simulated annealing procedure, so that the random numbers applied to select and perturb torsional displacements send each replica down a unique pathway on the free energy landscape. In this study, all torsional angles were allowed to move, with equal probabilities applied to backbone and side chain torsions. The side chains are flexible during the folding process. At the outset of a prediction trajectory, van der Waals steric clashes are immediately encountered, since we do not otherwise prevent clashes during each Monte Carlo move. It is inadvisable to spend much computer time on structures that clash out, so first we compute the VDW contribution to the intrinsic energy of each replica. When this interaction is more than three times larger than the initial structure, a short simulated annealing in Cartesian space is performed to try to relax the structure. If the energy still remains above the 3× threshold, we discard the move.

After a certain number of steps depending on the size of system, the annealed conformations are sorted with respect to effective free energy. The half with higher values are discarded. The lower energy structures are each duplicated and form the basis for another iteration of optimization. The temperature annealing schedule is started at 800 K with 100 to 500 K steps of Metropolis sampling. Subsequent temperatures typically chosen are 700, 600, 500, 400, 350, and 300 K. If an optimization at 300 K is not converged, we start again at 800 K using the current configuration. This procedure is repeated until the energy changes between two successive steps is < 0.1 kcal/mol. We note the similarities in MCSA to genetic algorithms for search procedures.[49] In MCSA, each of the copies is a conformational "mutation," which is subjected to evolutionary pressure applied by simulated annealing and the successive cycles of pruning the prediction ensemble. If a large number of copies were considered, a Monte Carlo Metropolis pruning could be applied; here we just take the best 8/16 at each stage of the annealing schedule. The use of "crossovers" as illustrated by Unger and Moult[49] on a 2D lattice model of protein folding is a possible line of subsequent refinement (vide infra).

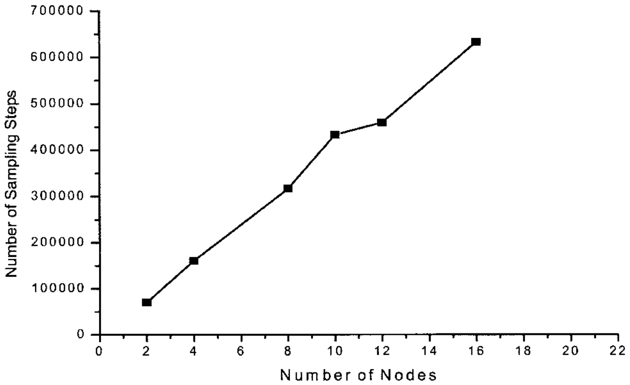


Fig. 1. Benchmarking *REFOLD* MCSA on our local Beowulf PC cluster. From the graph, the program scales linearly up to 16 processors running under MPI/LAM.

The prediction scheme as described above is implemented on a local Beowulf class PC cluster[67] comprised of 60 Pentium and AMD Athlon PCs interconnected by fast ethernet. An advantage of MCSA optimization is that the network communication is no longer a bottleneck for parallel computation as in MD. Implementation of the calculation using Message Passing Interface (MPI) protocols shows a linear scale for parallelization up to 16 nodes. Theoretically, the scaling of multicopy algorithm on multiprocessors is almost linear if the communication time compared with the time taken by optimization within each replica is neglected. In reality, the scale of the parallelization is limited by the communication mechanism for each platform. A benchmarking of the computational process from our PC Cluster is provided in Figure 1.

Calibration: Ala Dipeptide

An MD simulation of 1 ns was performed on the prototype system alanine dipeptide in solution including explicit water molecules, starting from the all *trans* extended form. The simulation involved minimization followed by heating to 300 K over 10 picoseconds. and a 30-ps period of equilibration. MD was then performed from the equilibrated structures for 1 ns, during which time the dipeptide interconverts among several major conformations. The probability density map as a function of Ramachandran ψ and ϕ angles from the MD on Ala dipeptide is shown in Figure 2. As expected, the large peak in the β sheet region of the probability map corresponds with the region of the conformational global minima. As a test, we applied the REFOLD to this molecule starting from an extended structure, and compared the results with those from MD.

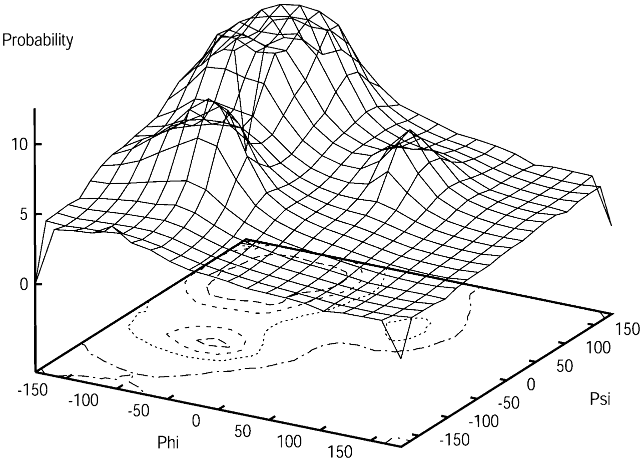


Fig. 2. Probability distribution map in ϕ/ψ space of Ala dipeptide from 1 ns MD simulation using *AMBER* 5.0 with the united atom force field. The three major clusters are α-R, β and (small peak) α-L.

The results are shown in Figure 3, with the results from single copy and multicopy minimizations given in Figures 3(a) and (b), respectively. On the density map of the single copy optimization, the global minima is sampled as well as two local minima whereas with multicopy, essentially all replicas are converged to the global minima. This comparison shows clearly that the multicopy protocol has advantages for structure predictions with respect to this class of problems.

Calibration: Performance of ΔG^{eff} on Structures From Protein Crystallography

Further calibration of *REFOLD* was carried out based on two protein crystal structures, Barnase (1BNI) and Crambin (1crn) with the aim of determining how well structures obtained from minimization of the effective energy function agree with protein crystal structures to begin with. Refold calculations were carried out beginning from the known crystal structure, submitted to a full temperature annealing cycle applied from 800 to 300 K. At each temperature cycle, 8 copies and 1,000 search steps were carried out for each copy. The resulting calculated structure with lowest energy was compared with the native structure by the Local Global Alignment (LGA) utility of Zemla[68] at http://PredictionCenter.llnl.gov, with a 10 Å cutoff. The resulting RMSD is 3.3 Å for Barnase and 2.7 Å for Crambin, i.e., reasonable close accord. The local RMSD histograms are shown in Figure 4 for Barnase and Crambin, and show some variation by residue and some patterning, such as a trend towards higher RMS at the C-terminal ends, which may or may not be significant.

Calibration: Performance of ΔG^{eff} on Decoys

Decoys are non-native structures for a given protein that are closely related to the native form but differ in salient details, and differentiation of decoys from the native structure is a primary means of preliminary testing of effective energy functions in protein folding.[69] Several compilations of protein structure decoys are available,[70] but generally the decoy sets are not sufficiently detailed to be applied to united atom or all atom potential functions. Thus, we generated sets of structures for barnase and crambin from ambient temperature and high temperature MD simulations including explicit water. Trajectories of 1 ns at 300 K and 5 ns at 600 K were carried out for each case using *AMBER* 5.0 by standard protocol using the Cornell et al. force field[8] and PME boundary conditions.[71] The MD at 300 K provides a distribution of structures in the vicinity of the native form, and the 600 K MD provides a set of decoys. Structures were culled from each trajectory at equally spaced intervals, and used as input to a *REFOLD* effective free energy calculation. The resulting energy profiles from REFOLD calculations for native and decoy structures of crambin and barnase are shown in Figure 5. The results indicate that the native and decoy

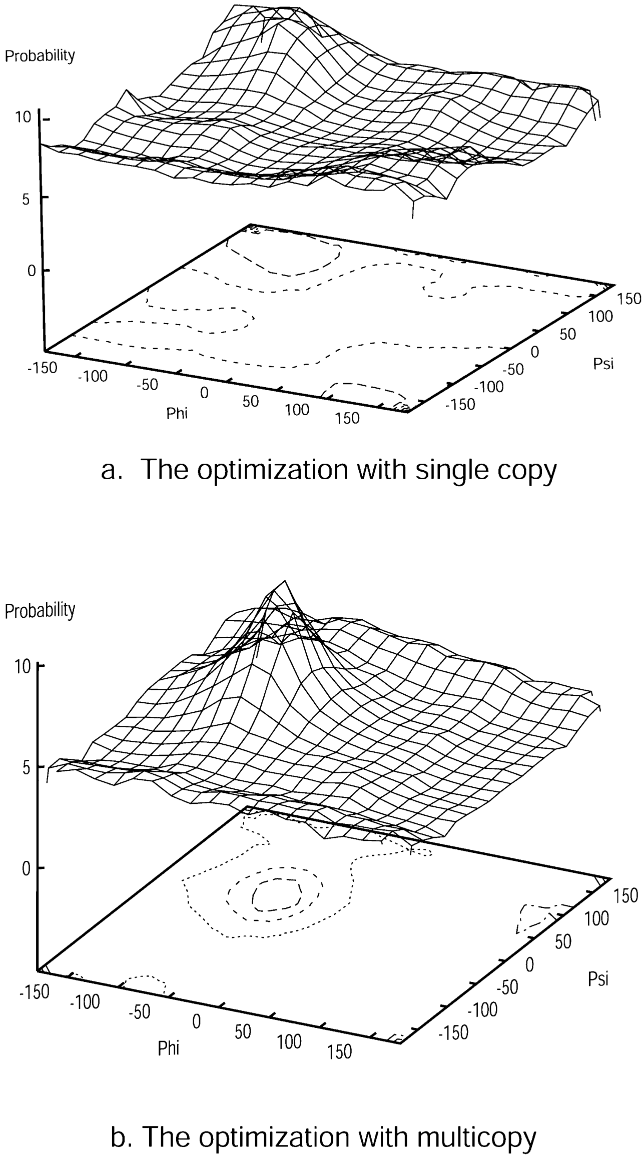
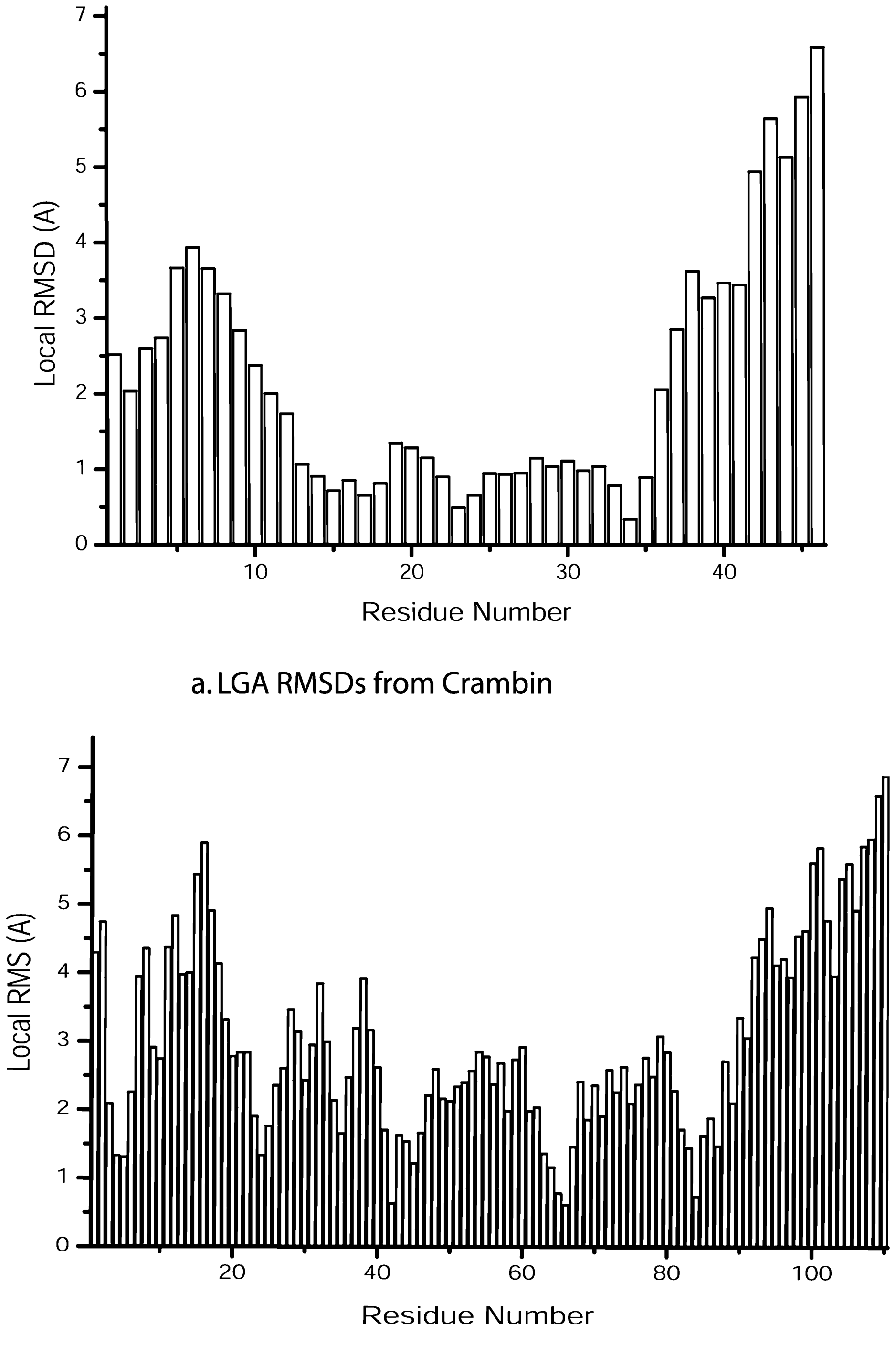


Fig. 3. Comparison of the efficiency of a single copy and multicopy algorithm results on ϕ/ψ population maps for the Ala dipeptide: (**a**) single copy algorithm; (**b**) multicopy algorithm.



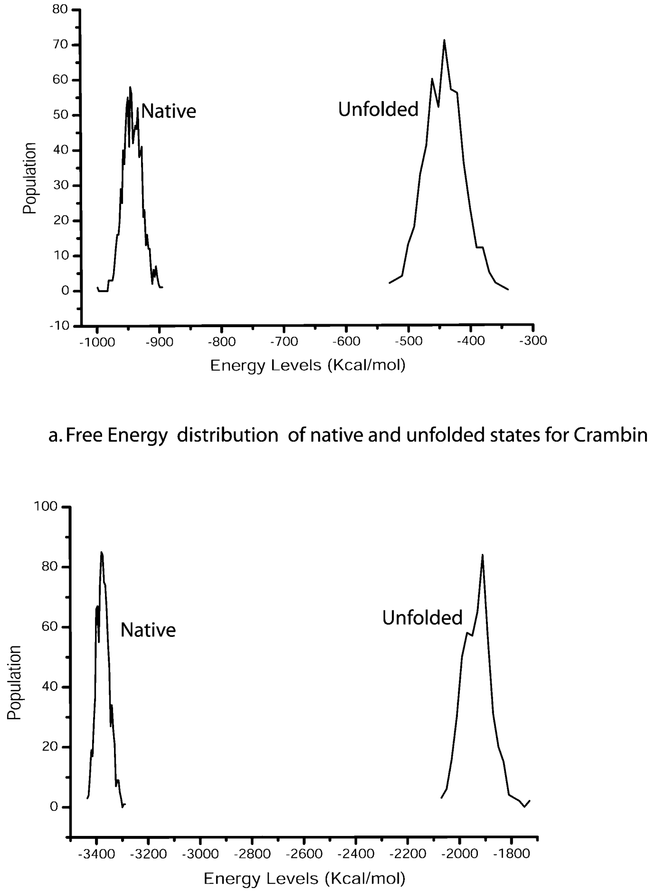
a. LGA RMSDs from Crambin

b. LGA RMSDs from Barnase

Fig. 4. Histogram of local RMSDs between calculated structures and crystal structures: (**a**) crambin and (**b**) barnase.

sets for barnase and crambin are well differentiated with respect to effective free energy.

For a further test of our scoring function on decoys, we randomly selected 6 proteins from the Rosetta web site.[56] We randomly selected one decoy conformation from 1,000 conformation for each protein provided by the Rosetta decoy set. The corresponding crystal conformation was obtained from the Protein Data Bank and the residues



a. Free Energy distribution of native and unfolded states for Crambin

b. Free Energy difference of native and unfolded states for Barnase

Fig. 5. Free energy distributions of native and decoy conformations for (**a**) crambin and (**b**) barnase.

absent in the decoy form were deleted. Using Amber 5.0, 50 steps of the steepest decent minimization were performed on the decoy form to relieve any VDW repulsions. The energy components and the total energy were calculated for each conformation based on our effective potentials. The results are summarized in Table II. For all six Rosetta proteins, the native form has lower free energy than the corresponding decoy. A more comprehensive study of decoys is, of course, appropriate, but these results combined with those on barnase and crambin described above are considered sufficient for proceeding with exploratory studies.

RESULTS

REFOLD prediction trajectories were carried out for 5 polypeptides of various lengths ranging from 8 to 36 amino acids, chosen as a series of progressively more challenging cases. The smallest case was an 8-residue polyalanine sequence, followed by a 12-mer with more complex side chains, for which a crystal structure is available. A 16-mer with 40–60% α-helix based on CD served as the next most complicated case. The preliminary tests carried out so far conclude with a 15-residue polypeptide exhibiting β-turn geometry, and the 36-residue villin headpiece. Villin is an advantageous case to study because of the parallel MD studies available on this system for comparison by Kollman and coworkers.[13,14] For Villin, we provide a more detailed analysis of results and demonstrate how the results from our calculations can be decomposed as a function of the prediction coordinate and used to understand the chemical forces active at various stages of a prediction trajectory.

AAAAAAAA (Ala 8)

Alanine has the greatest tendency among amino acids to form α-helices, and we first applied to a sequence of 8 consecutive Ala residues capped at the C-terminal and N-terminal ends with acetyl (Ac) and N-methyl-amine (NHMe) groups. The prediction process was initiated from a fully *trans* extended form of the molecule. This calculation involved 16 replicas on 16 PC nodes. The result is shown in Figure 6 in a series of four panels, which depict the initial form, the structure after 1/3, 2/3 and at completion, respectively. The total length of the folding involved 70 K steps of MCSA. At conclusion, the structure has clearly adopted a reasonably regular form of α-helix. This is actually a "toy" problem, since the molecule is expected to be insoluble in water. This results allows us to point up the difference between structure prediction based on a scoring function and a well-defined statistical mechanics treatment of protein folding, in which non-local entropy effects act to thermodynamically destabilize folded forms.[72] Similar test cases have served in other structure predictions.[51]

QLLKKLLQQLKQ (Hill et al., 12-mer)

REFOLD was next applied to a 12-residue peptide Ac-QLLKKLLQQLKQ-NHMe, a synthetic peptide comprised of a larger variety of amino acid residues with longer side chains than Ala, for which a crystal structure was reported by Hill et al.[73] The prediction process was initiated from a linear structure, also using 16 copies, one per processor. The prediction converged after ~230 K steps of simulated annealing, and prediction trajectory is depicted in Figure 7. The prediction yielded a lowest energy structure with an RMSD of 2.65 Å (heavy atoms) with respect to the Hill et al. crystal structure. The helix region of the predicted and observed structure overlap very well. The side chain of one residue from the tail section of the prediction lies in the opposite direction from that of the crystal structure, possibly a consequence of the flexibility of side chains in solution. In this case, several of later stage structures in the prediction were close in energy. These structures and their corresponding RMS values with respect to the crystal form are shown in Figure 8.

$(AAQAA)_3$ Y (Scholtz et al. 16-mer)

A 16-mer peptide is a de novo designed peptide that folds in solution[74] to a structure with a mixture of α-helix and random coil. *REFOLD* was again initiated from a fully extended conformation, with optimization was carried out on 16 copies. The calculated folding trajectory is shown in

TABLE II. Energy Comparison Between Native Protein and Rosetta Decoys (Kcal/mol)

	1aa2	1lfb	1lis	1r69	1ris	1vls
$\Delta G_{\text{Native-Decoy}}$	−2060.6	−249.8	−868.9	−474.5	−642.6	−705.0

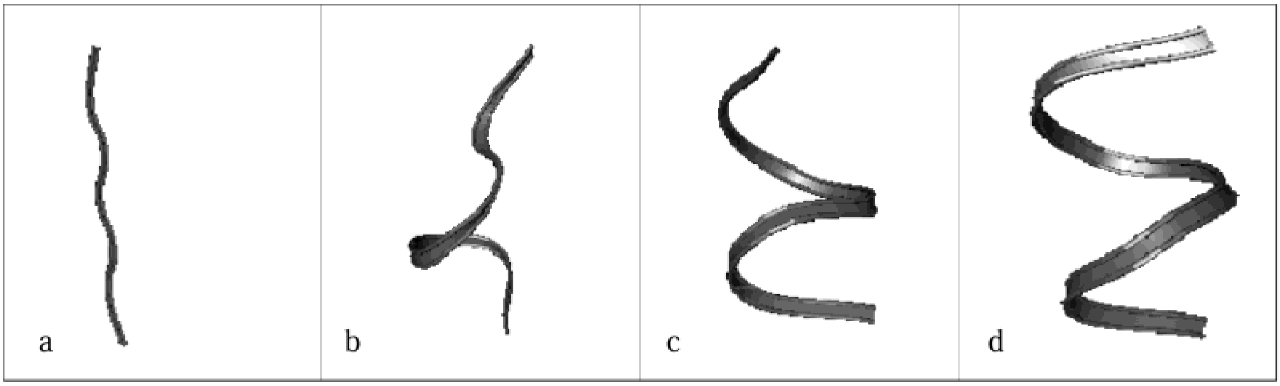


Fig. 6. *REFOLD* results on Ala-8. **a:** Starting structure. **b:** Early stage. **c:** Late stage. **d:** Predicted structure.

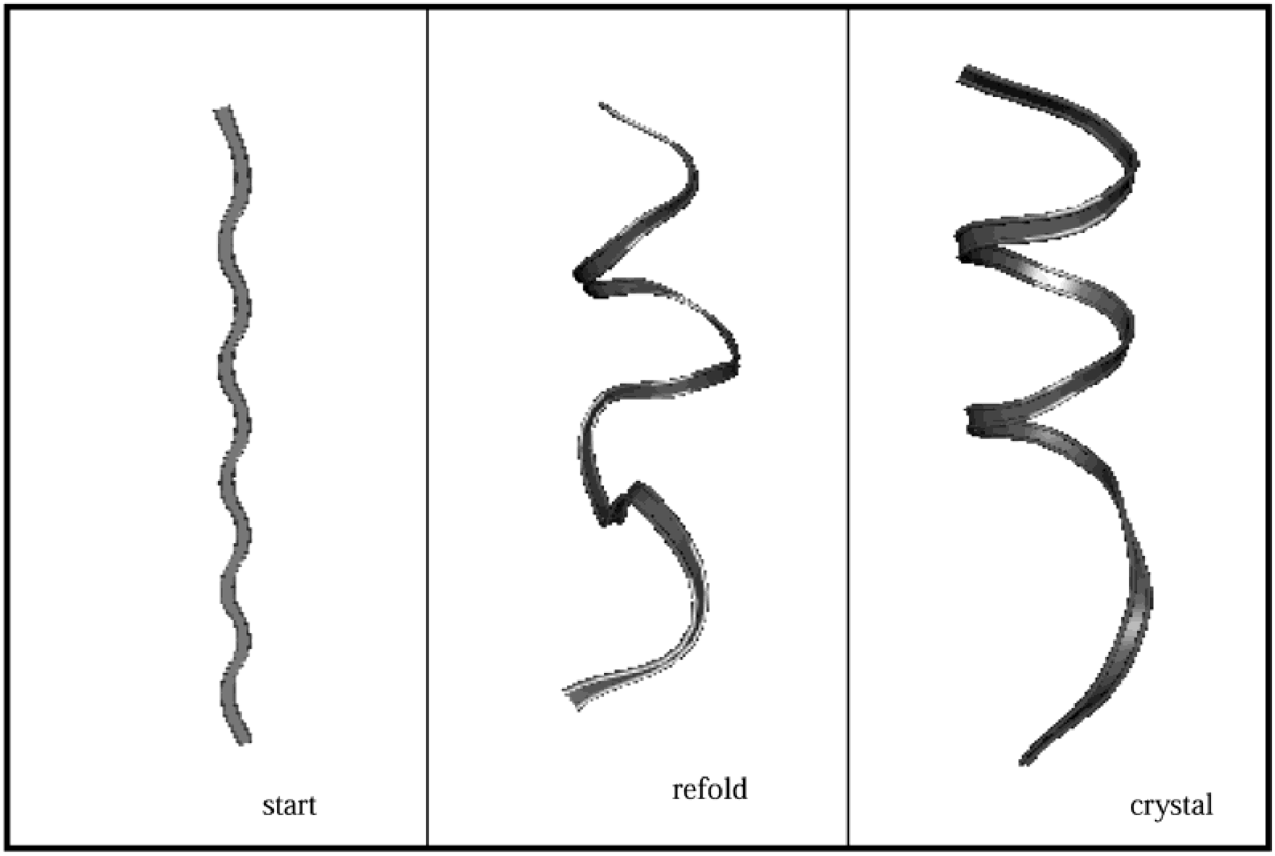


Fig. 7. *REFOLD* results on Hill et al.[73] 12 mer.

Figure 9. The predicted form shows extensive α-helix, with frayed C-terminal and N-terminal ends. There is no crystal structure available for this sequence, so in this case the predicted structure must be assessed by other means. First, PROCHECK[75] was used to make a Ramachandran plot (Fig. 10), from which one can see that all backbone angles lie in the allowed regions. Here 10 residues are located in the most favored region, four of them are located in the next most region, and all are within in the region denoted "generous." No residue is located in any disal-

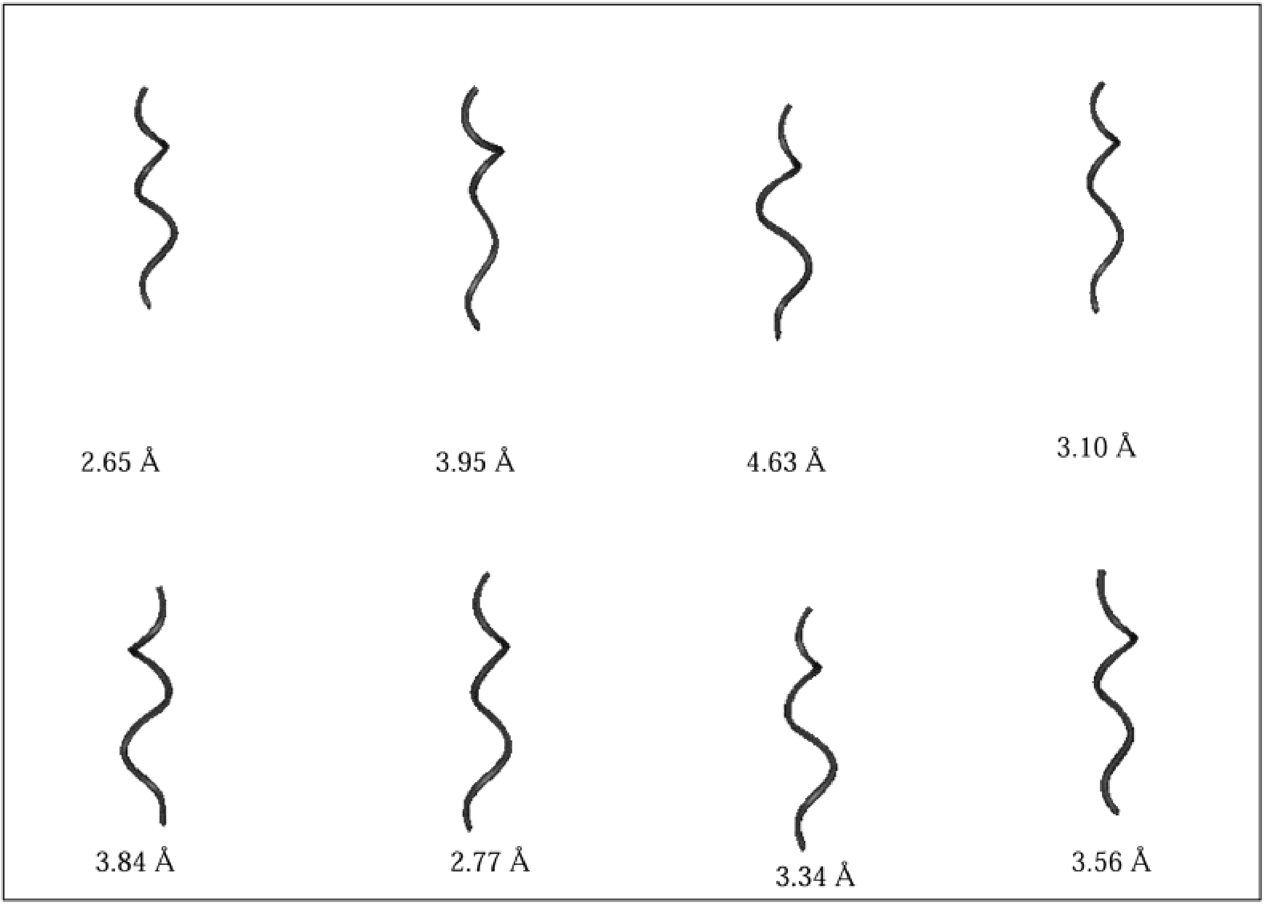


Fig. 8. Set of low score predicted structures for Hill et al.[73] 12 mer, including RMSD values to crystal structure.

lowed region. Figure 11 shows that the α-helix is formed in the central region, with reduced solvent accessibility. The predicted can be compared with helix probability predictions calculated by the program CapHelix based on Lifson-Roig calculation.[76] Similar comparison between the Lifson-Roig model and isotope labeled sequence YAAKAAAAK-AAAAKAAH has been reported earlier,[77] which shows that residues at the N-terminus and in the center are more likely to be helical than residues in the C-terminus. For the Scholtz 16-mer, the Lifson Roig calculated probability as a function of residue positioning is shown in Figure 12. From the plot, we can see that the most probable region for a helix is at or near in the middle of the sequence, consistent with the predicted helix structure from the MCSA algorithm. The CapHelix calculation also suggests higher values for nucleation probability at both end positions, which seems to match what is observed in the prediction trajectory.

GEWTWDATKTFTVTE (PDB Code 1GB1)

This sequence has a β-strand segment (residue 41–56) chopped from a NMR determined structure, the immunoglobulin binding domain of streptococcal protein.[78] In *REFOLD*, an extended structure of this sequence folds successfully to a β motif (Fig. 13). The RMSD between the predicted and crystal form is 4.5 Å.

Chicken Villin Headpiece (PDB Code 1VII)

Villin is a 36 residue protein with a three helix structure as determined by NMR.[79] *REFOLD* applied to this sequence beginning with an arbitrary V-shaped extended structure yielded the prediction trajectory shown in Figure 14. The RMSD between the predicted structure and the crystal structure is 3.4 Å (including side chains) and 1.01 Å (backbone only). An RMS coverage graph for this prediction is shown in Figure 15(a)[80] and in an local RMS histogram from the LGA utility in Figure 15(b). Some 90% of the amino acid residues in Villin are predicted with 1.5 Å, which we consider excellent results considering the practical resolution of the methodology. The local RMS histogram shows the lowest local RMSD is in the longest helix region (helix 3).

An analysis of the calculated free energy components of Villin as a function of progress along the prediction trajectory is shown Figure 16. The sum total decomposed into contributions intrinsic to the molecule and from solvation are shown in Figure 16(a). The tradeoff between decreasing internal energy and increasing (less negative)

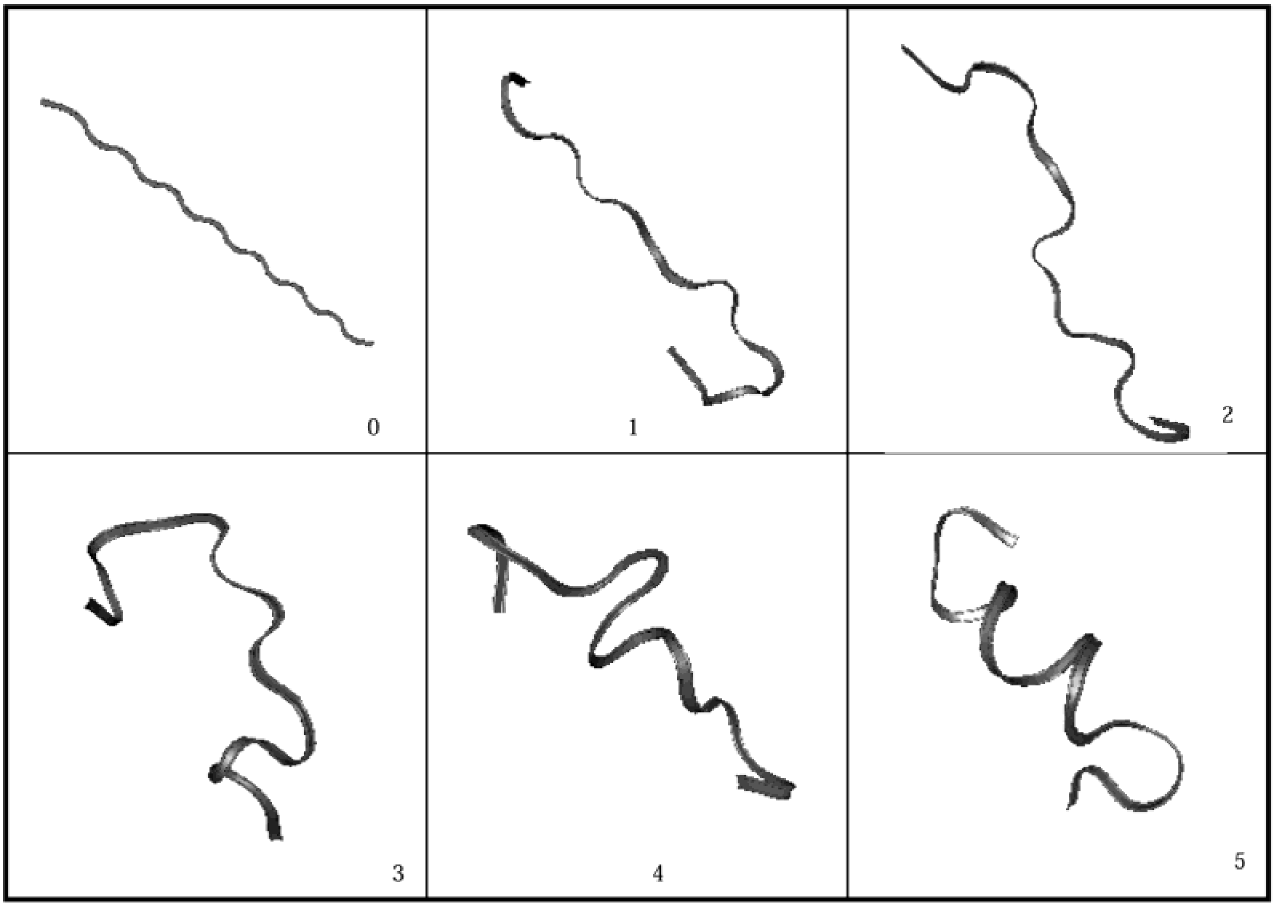


Fig. 9. *REFOLD* results on the Scholtz et al.[74] 16-mer peptide. **0:** Starting structure. **1–4:** Intermediates along the prediction trajectory. **5:** Predicted structure.

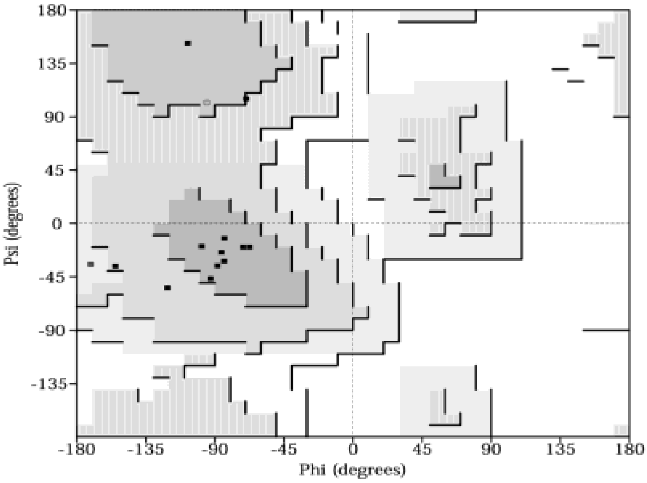


Fig. 10. Ramachandran plot of calculated values of ϕ,ψ for *REFOLD* result on Scholtz et al.[74] 16-mer.

solvation energy is clearly observed. While the intrinsic contribution clearly controls the overall prediction trajectory, if the solvent term were to be neglected a much different result would have been obtained. A further decomposition of the intrinsic and solvation contributions

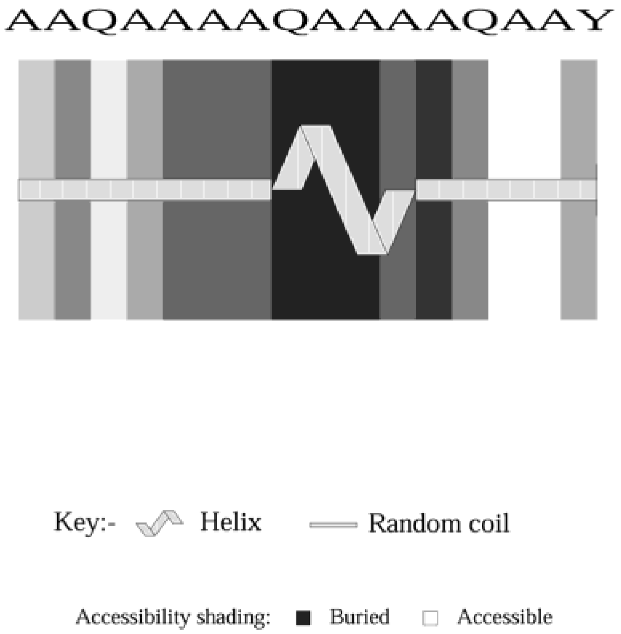


Fig. 11. Secondary structure and solvent accessibility plot of *REFOLD* structure on Scholtz et al.[74] 16-mer.

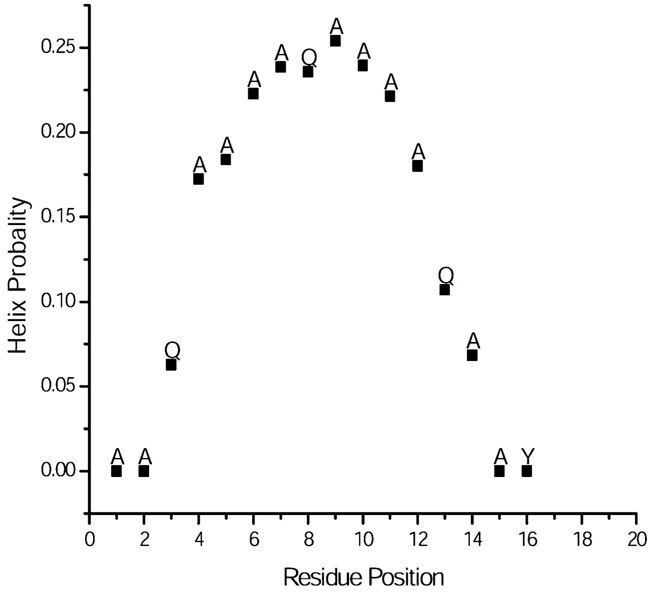


Fig. 12. Plot of calculated helix probability vs. residue position for Scholtz et al.[74] 16 mer. Calculated helix probabilities were generated from a modified Lifson-Roig calculation using the program CapHelix.[76]

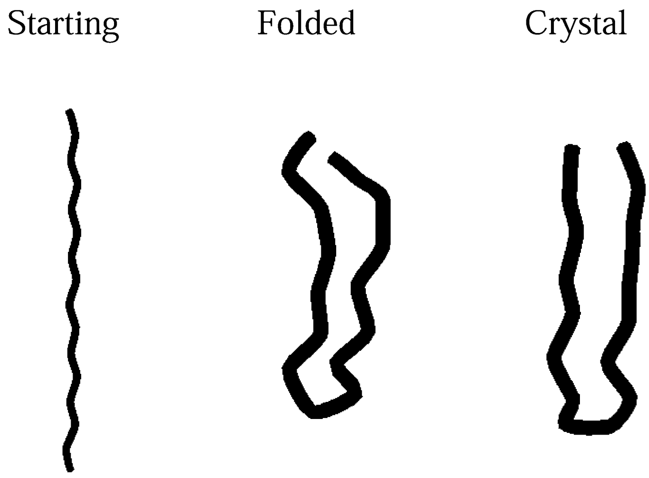


Fig. 13. *REFOLD* results on a β segment sequence GEWTWDATKT-FTVTE.

is shown in Figures 16(b) and (c), respectively. The intrinsic contribution to the effective free energy function is seen to be dominated by electrostatics in the early stages of the prediction and by the van der Waals interactions in late stage refinement. The calculated solvation energy features opposing contributions from electrostatics and non-electrostatic contributions, with electrostatics of solvent polarization dominant. For the non-electrostatic component, the cavity term is seen to increase along the folding trajectory, reflecting the free energy required to create room in the solvent for the folded protein. It is conventional to associate the burial of non-polar residues with hydrophobic effect, which is factored out of the total and also shown in Figure 16(c). This is seen to contribute only a fraction of the total.

In order to compare the results obtained on energy components as a function of prediction trajectory for all the cases studied, we fit the profile of each of the free energy components for each case to an exponential decay function of the simple form,

$$E = E_0 + Ae^{-(t/\tau)}$$

in which E_0, A, and τ_0 are fitting parameters. From the results, the magnitude of each contribution can be compared and the correlation "time" τ gives an idea on how fast from the initial folding the decay of a term occurs. The analysis is summarized in Table III. Preliminary consideration of the results shows each case to be somewhat different, which suggests to obtain the most information from this type of approach we will need to study a large set of cases grouped by classifications such as fold category. This is quite feasible, but well beyond the scope of the present study.

DISCUSSION

In the preceding, we have described a set of exploratory studies in ab initio protein structure prediction, using an *AMBER* united atom empirical energy functions, a GBSA implicit solvent model, and a multiple copy MCSA search engine combined into a prediction protocol we call *REFOLD*. The results show that *REFOLD* as implemented recovers the structures of test cases consisting of diverse structural motifs within 6 Å. In particular, a prototype α/β motif was treated successfully and, in our most challenging test case, the 36-residue villin headpiece was predicted to within 3.5 A RMS (including sidechains) and 1.10 A RMS (backbone only), surprisingly good for a pure "sequence determines structure" ab initio structure prediction. Analysis of the results in terms of energy components shows the general counterbalance of intrinsic energy and solvation effects over the course of the unfolding trajectory to be successfully reproduced. For the largest and most protein-like case studied, early stage folding is characterized by offsetting contributions from the terms intrinsic to the polypeptide chain and terms from solvation. Intermediate stage folding by compensation of solute contributions and solvent contributions and late stage folding by subtle balance of various terms.

Certain specific details are worth noting, both on the evolution of the calculated structure of villin from an extended to folded form, and as an illustration of how the methodology yields information on the how and why of structure prediction. Examination of the results for villin with respect to secondary and tertiary structure shows that the two orders of structure form simultaneously as the prediction evolves. One of the ideas about protein folding is that certain cases pass through molten globule states stabilized by hydrophobic forces. The extent to which this is observed in *REFOLD* can be readily examined by a more detailed analysis of the structures and calculated properties such as radius of gyration (all heavy atoms) along the prediction trajectory. The results of for villin heavy atoms are shown in Figure 17, but show no evidence of a plateau, only a monotonic decline. Consider-

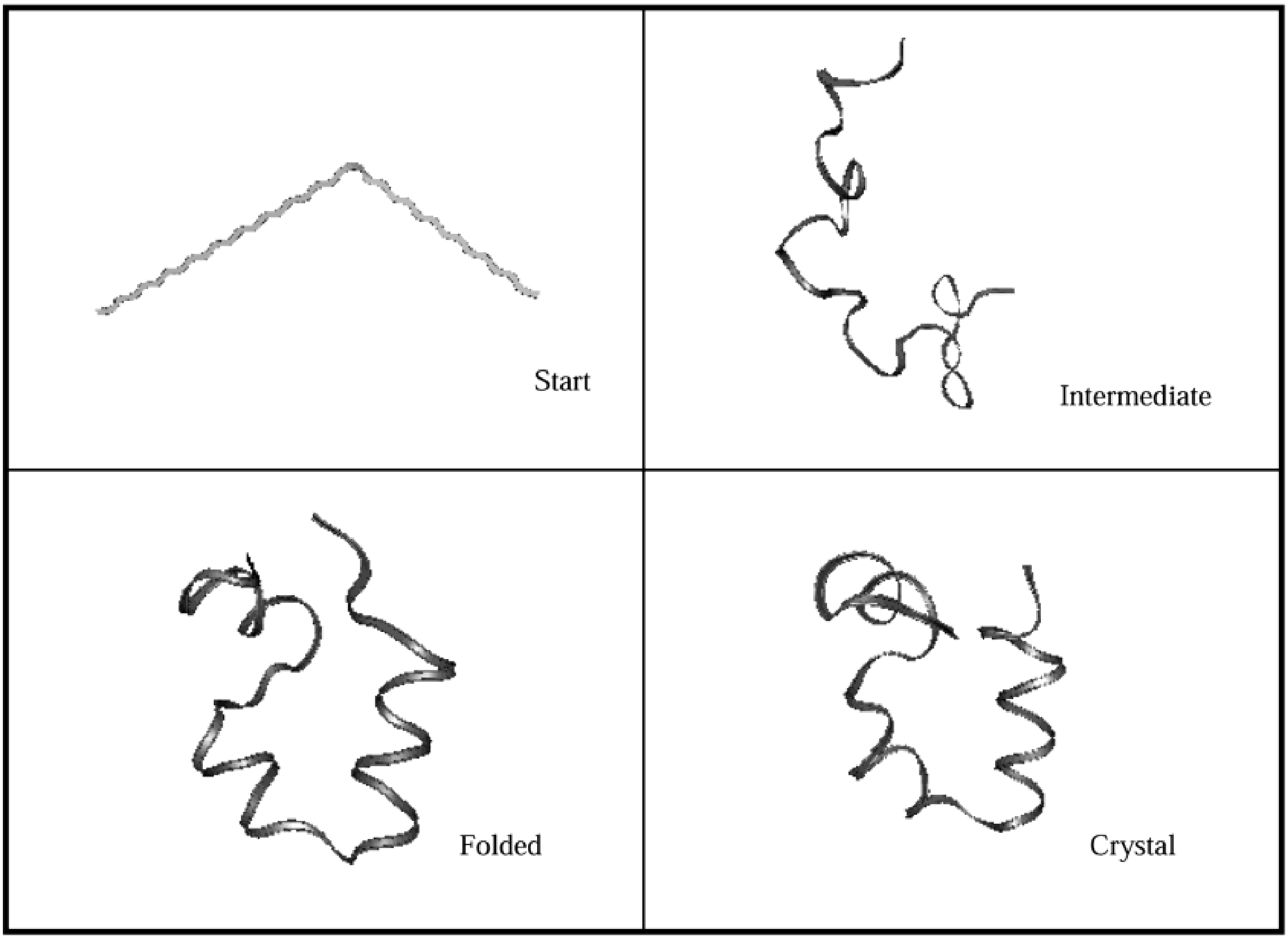


Fig. 14. *REFOLD* results on villin.

ing this results and the contribution of non polar groups to the non-electrostatic component of the solvation energy (Fig. 18) above shows van der Waals solvation terms are more important in the early stages of the prediction trajectory, and the cavity terms associated with the hydrophobic effect are more important at later stages.

The good results on villin are tempered somewhat by our experience so far in going to larger proteins, but a systematic study of the reliable limit of *REFOLD* as implemented is a work in progress. The CASP4 prediction contest closed in September of 2000, and we had only sufficient time to prepare one prediction, the 128 residue target T0110 of sequence MAREF KRSDR VAQEI QKEIA VILQR EVKDP RIGMV TVSDV EVSSD LSYAK IFVTF LFDHD EMAIE QGMKG LEKAS PYIRS LLGKA MRLRI VPEIR FIYDQ SLVEG MRMSN LVTNV VREDE KKHVE ESN. This forced us to jump from a 36 to a 128 residue prediction without the benefit of an aufbau process, and we proceeded on this with considerable trepidation. In this case, our prediction protocol was not purely ab initio, but involved knowledge based secondary structure prediction for the initial structure from a consensus of results from http://jura.ebi.ac.uk:8888 plus some selective elimination of nonglobular intermediate results. The prediction was submitted, and in the on subsequent assessment turned out to be 14.3 Å from the crystal structure, an α/β motif. The discrepancies are nontrivial, indicate for more complex cases there is much to do before truly satisfactory predictions can be achieved. A Hubbard type RMS coverage map for all the predictions made on T0110 in CASP4 is shown in Figure 19. Compared with all the other predictions on this target, our submission was nonetheless in the middle of the pack. The best predictions on this target were quite good, but also involved extensive use of knowledge based information and or statistical potentials.

The results from this exploratory study and those of this genre emerging from other groups are encouraging enough to undertake further testing and improvements, and also to explore how an ab initio *REFOLD* protocol might be incorporated into an expanded structure prediction protocol for late-stage refinement of knowledge-based results. Specific deficiencies in the potentials can be identified with the aid of sequence dependent and sequence independent RMS coverage graphs and LGA local RMS histograms. Our choice of intrinsic energy function was dictated by a decision to restrict ourselves to published force fields, but

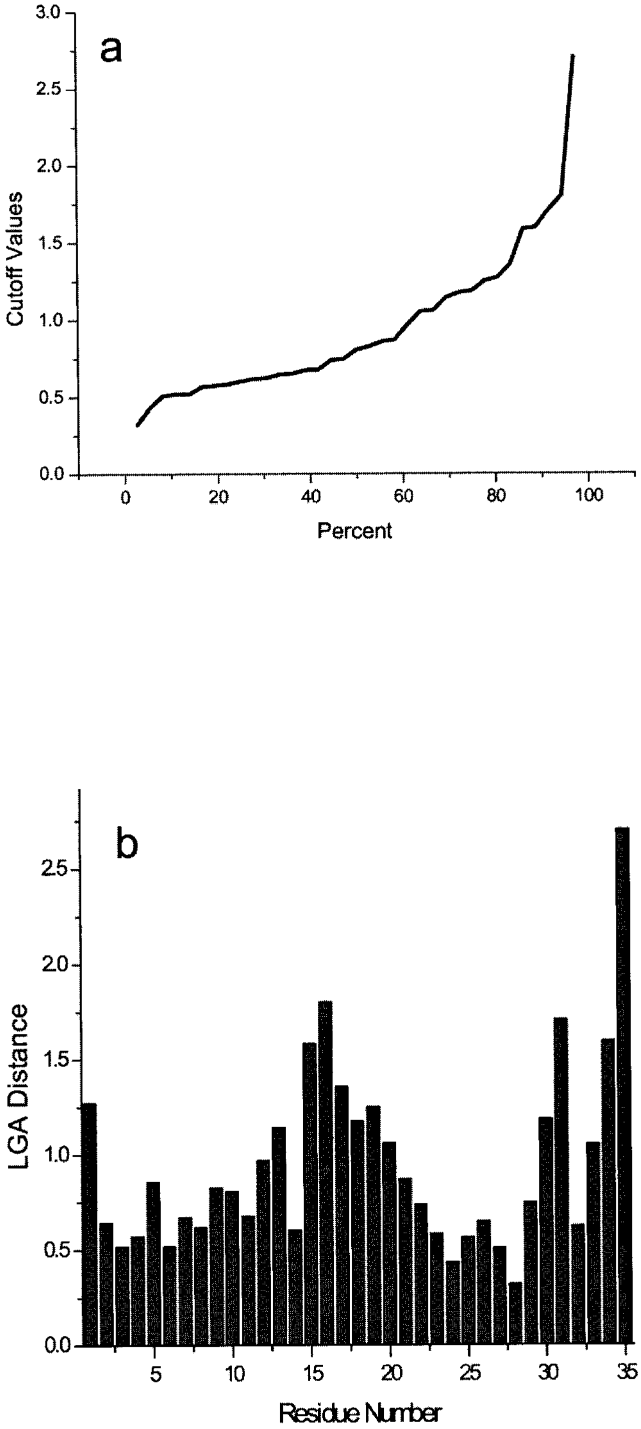


Fig. 15. Hubbard plot and LGA local RMSD histogram for villin.

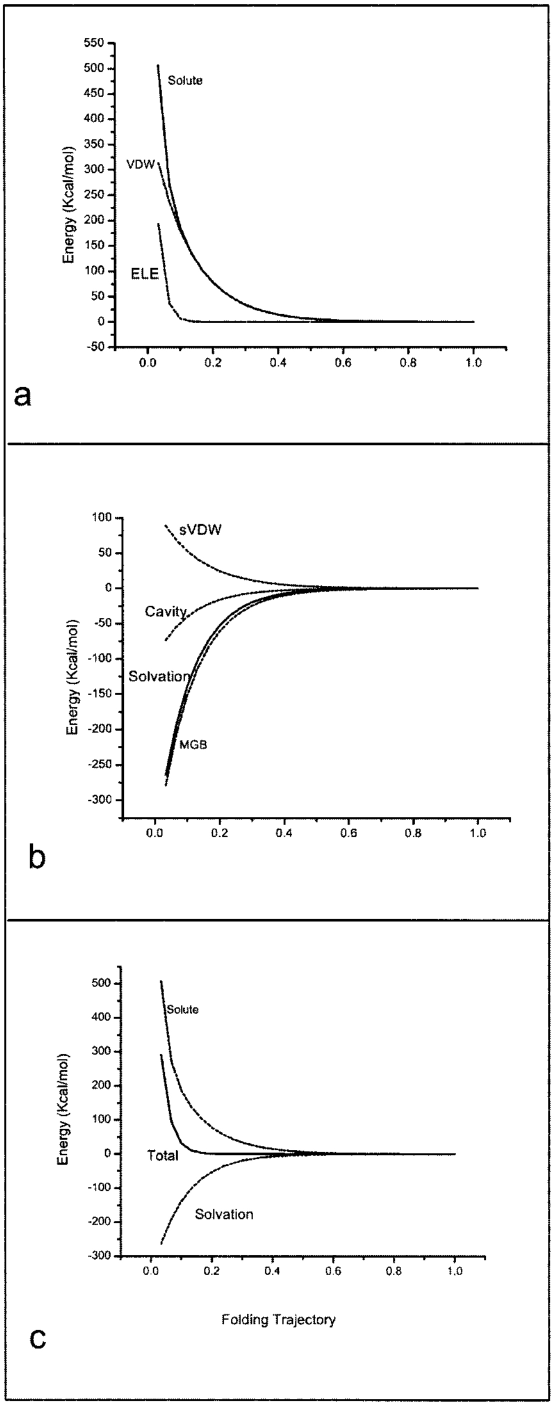


Fig. 16. Calculated free energy components of villin as a function of progress along the prediction trajectory.

an improved all atom AMBER force field could be used to construct a second-generation, united atom effective-free energy function. Other force field options are available as well,[16] and can be used as a sensitivity test and determine the extent to which predictions of this genre depend on fine differences in force fields. With regard to the solvent model, studies that provide an improved treatment of the interior of proteins in GBSA type models have been described,[81] and it will be interesting to see what difference this effect makes in predictions. In the area of sampling, the MCSA search procedures could be heuristically fine tuned with respect to dominant contributions from various terms at various stages of the prediction trajectory. Beyond a basic implementation of simulated annealing, there are a host of advanced Monte Carlo based sampling methods that hold further promise for improved

TABLE III. Results of Single Experimental Fitting of Free Energy Components Villin

Terms	DIH	ELE	VDW	MGB	sVDW	Cavity	Total	F[a]	W[b]
E_0	−762.2	−62.5	−365.6	−760.2	163.43	−138.70	−989.47	11.1	16.9
A	954.6	1023	413.9	−377.7	115.18	−100.10	883.70	8.6	11.28
T_0	38.2	0.02	0.12	0.11	0.13	0.11	0.03	0.19	0.08

[a] Hydrophobic contributions to non-polar solvation free energy from hydrophobic residues (ACFILMV).

[b] Hydrophobic contributions to non-polar solvation free energy from hydrophilic residues (DEGHKNPQRSTWY).

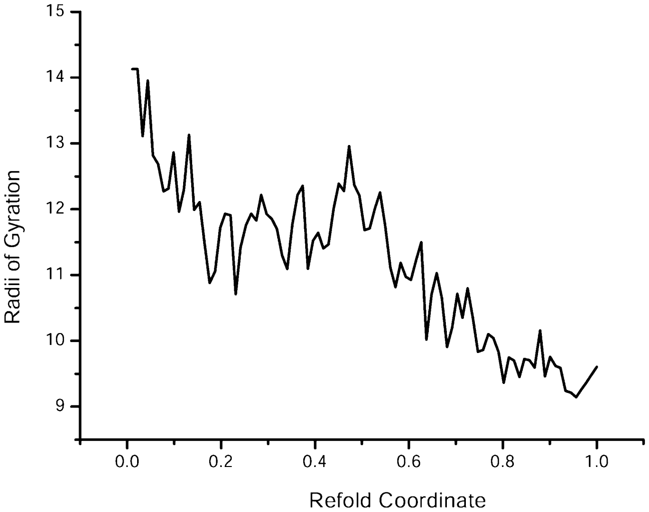


Fig. 17. Calculated radii of gyration along the *REFOLD* prediction trajectory for villin.

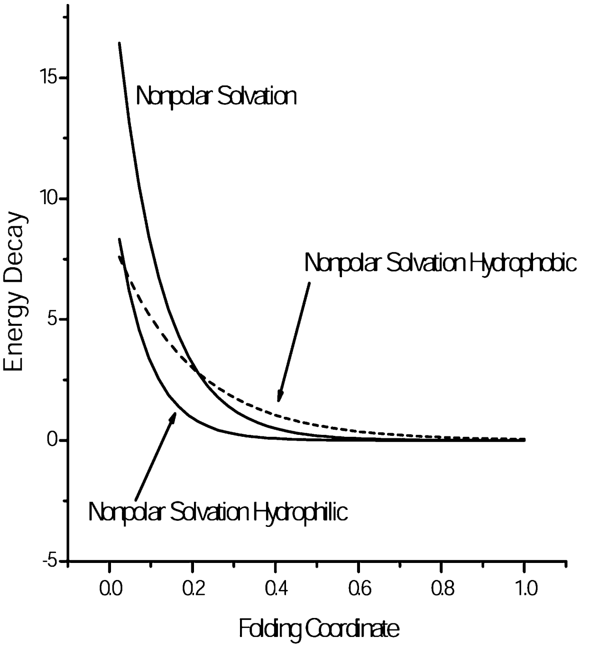


Fig. 18. *REFOLD* results on exponential decay of non-polar solvation contributions from both hydrophobic and hydrophilic groups for Villin. Thick line: total non-polar solvation. Thin line: non-polar solvation contributions from hydrophilic groups. Dashed line: non-polar solvation contributions from hydrophobic groups.

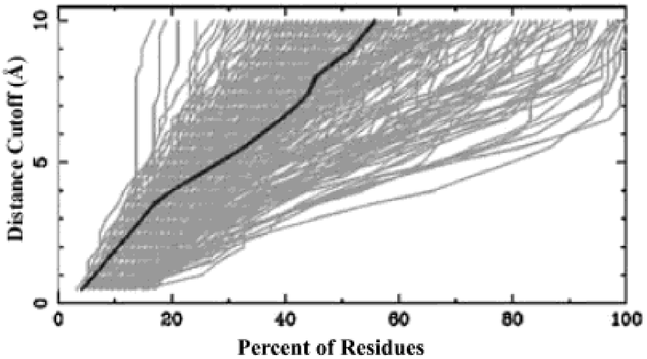


Fig. 19. Hubbard plot from CASP4 for target T0110. Bold line: *REFOLD* prediction.

performance.[48] For the most part, these have been tested only on prototypes, and so far not a lot of cross comparison is available. The results of this study, extended with results on more previous CASP targets or structures from other sources, contribute further benchmark cases from which one may quantitatively document progress in methodological developments.

In concluding, ab initio structure prediction as described herein is only part of a more general scheme purely theoretical approach to the problem. Commentary by K. A. Dill at CASP4 noted that one way to distinguish true physics based predictions from other methodologies that had previously been classed as ab initio is a legitimate dependence on temperature and Boltzmann's equation. While MCSA employs an effective temperature in setting an annealing schedule, this is an effective temperature and does not yield a method in compliance with Dill's criteria. To move in this direction, the *REFOLD* type protocol, either with or without knowledge-based input, could serve as the initial stage of a hierarchy of steps that, in the final analysis, produces a properly defined Boltzmann statistical mechanics treatment of the problem. We are investigating the following series of steps, using villin as a test case: (1) use *REFOLD* to obtain a low resolution prediction for the structure of the native state (N-search); (2) use this structure as a basis for a search for substrates, treating the protein plus explicit solvent using a high temperature MD followed by quenching at well defined intervals to identify substates (n-search). An alternative possibility is to use a hybrid MC MD protocol for substate search.[82] A final step would be to determine the statistical weights of substates (w_n-search) by clustering methods or 2D RMS.[83] The result would be a structure prediction protocol that at the final step is consistent with statistical mechanics and displays proper temperature-dependence, at least in principle. The practical viability of this scheme of course remains to be demonstrated, but individual elements are each obviously feasible. Current studies of villin and several other cases are in progress to provide a demonstration.

SUMMARY AND CONCLUSIONS

Studies of ab initio protein structure prediction, using an *AMBER* united atom empirical energy functions, a GBSA implicit solvent model, and a multiple copy MCSA search engine were described. In the prediction protocol, structures are periodically culled, evaluated on the basis of calculated effective free energies, and the lower energy forms are used to spawn new generation of structure prediction on an annealing schedule of progressively lower effective temperatures. The choice of energy function takes advantage of the considerable parameterization efforts originally designed for MD simulations. The GBSA solvent model is simple and economical to implement, represents solvent dielectric polarization, van der Waals and cavitation effects, and is parameterized to give accurate estimates of solvation free energies on prototype cases. The MCSA protocol takes explicit advantage of the energy landscape protein folding, and Monte Carlo Metropolis sampling has the ability to escape to some extent from metastable local minimal on the free energy surface. The results show that the method as implemented to recover the structures of test cases consist of diverse structural motifs within 6.0 Å RMS. Analysis of the contributions of various components of the conformational free energy as function of the folding trajectory shows that prediction follows a landscape model of protein folding, and the contribution of factors such as electrostatics, van der Waals interactions, and the hydrophobic effect fully quantified. The results are sufficiently promising that further studies of this type of protocol are warranted.

ACKNOWLEDGMENTS

This work is supported by NIH GM 37909 (D.L.B., P.I.). We acknowledge in particular the initiatives of Prof. Reinhold Blumel and Thomas Clausen of the Wesleyan University Physics Department that led to the development of the PC cluster on which our *REFOLD* calculations were carried out. A number of analysis tools that were developed in the course of the CASP series of conferences proved useful for systematic and well defined RMS calculations. We thank NCSA and NCI supercomputing center for the computing time for MD simulations, and the *AMBER* collaboratory for the availability of their force fields and other software. Discussions of the protein folding problems with Profs. Barry Honig, Ken Dill and Peter Kollman are gratefully acknowledged.

REFERENCES

1. Anfinsen CB. Principles that govern the folding of protein chains. Science 1973;181:223–230.
2. Privalov PL. Thermodynamic problems of protein structure. Annu Rev Biophys Biophys Chem 1989;18:47–69. Privalov PL. Stability of proteins: small globular prteins. Adv Protein Chem 1979;33:167–241.
3. Dill K. Dominant forces in protein folding. Biochemistry 1990;29:7133–7155.
4. Osguthorpe DJ. Ab initio protein folding. Curr Opin Struct Biol 2000;10:146–152.
5. Sternberg MJ, Bates PA, Kelley LA, MacCallum RM. Progress in protein structure prediction: assessment of CASP3. Curr Opin Struct Biol 1999;9:368–373.
6. McCammon JA, Harvey SC. Dynamics of Proteins and Nucleic Acids, Cambridge University Press; 1987. van Gunsteren, W F & Berendsen, HJC. Computer simulation of molecular dynamics: Methodology, applications and perspectives in chemistry. Agnew Chem Int Ed Engl 1990;29:992–1023. van Gunsteren, WF. Molecular dynamics studies of proteins. Curr Opin Struct Biol 1993;3:277–281. Leach, AR. Molecular Modeling: Principles and Applications, Essex, UK: Addison Wesley Longman Ltd; 1996.
7. Brooks III CL, Karplus M, Pettitt BM. Proteins: a theoretical perspective of dynamics, structure and thermodynamics. In: Prigogine I, Rice SA, editors. Advances in chemical physics. New York: John Wiley and Sons; 1988.
8. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz Jr KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J Am Chem Soc 1995;117:5179–5197.
9. MacKerell AD. Developments in the CHARMM all-atom empirical energy function for biological molecules. Abstr Papers Am Chem Soc 1998;216:042.
10. MacKerell AD Jr, Bashford D, Bellott M, Dunbrack RL, Evanseck J, Field JD, Fischer MJ, Gao S, Guo J, Ha H, Joseph-McCarthy S, Kuchnir D, Kuczera L, Lau K, Mattos FTK, Michnick C, Ngo S, Nguyen T, Prodhom DT, Reiher WBE, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, rkiewicz-Kuczera JW, Yin D, Karplus M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B 1998;102:3586–3616. MacKerell J, AD & Banavali, N. All-atom empirical force field for nucleic acids: II. Application to molecular dynamics simulations of DNA and RNA in solution. J Comput Chem 2000;21:105–120.
11. Caflisch A, Karplus M. Molecular dynamics simulation of protein denaturation: solvation of the hydrophobic cores and secondary structure of barnase. Proc Natl Acad Sci USA 1994;91:1746–1750. Tirado-Rives J, Orozco M, Jorgensen WL. Molecular dynamics of the unfolding of barnase in water and 8M aqueous urea. Biochemistry 1997;36:7313–7329. Daggett V, Levitt M. molecular dynamics simulations of helix denaturation. J Mol Biol 1992;223:1121–1138. Li A, Daggett V. Molecular dynamics simulation of the unfolding of barnase: characterization of the major intermediate. J Mol Biol 1998;275:677–694.
12. Daura X, van Gunsteren WF, Mark AE. Folding-unfolding thermodynamics of a beta-heptapeptide from equilibrium simulations. Proteins 1999;34:269–280. Daura X, Jaun B, Seebach D, van Gunsteren WF, Mark AE. Reversible peptide folding in solution by molecular dynamics simulation. J Mol Biol 1998;280:925–932.
13. Duan Y, Wang L, Kollman PA. The early stage of folding of villin headpiece subdomain observed in a 200-nanosecond fully solvated molecular dynamics simulation. Proc Natl Acad Sci USA 1998;95:9897–9902. Duan Y, Kollman PA. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. Science 1998;282:740–744.
14. Lee MR, Duan Y, Kollman PA. Use of MM-PB/SA in estimating the free energies of proteins: application to native, intermediates, and unfolded villin headpiece. Proteins 2000;39:309–316.
15. Hao MH, Scheraga HA. Designing potential energy functions for protein folding. Curr Opin Struct Biol 1999;9:184–188. Lazaridis T, Karplus M. Effective energy functions for protein structure prediction. Curr Opin Struct Biol 2000;10:139–145.
16. Lazaridis T, Karplus M. Effective energy function for proteins in solution. Proteins 1999;35:133–152.
17. Still WC, Tempczyk A, Hawley RC, Hendrickson T. Semianalytical treatment of solvation for molecular mechanics and dynamics. J Am Chem Soc 1990;112:6127–6129.
18. Cramer CJ, Truhlar DG. Implicit solvation models: equilibria, structure, spectra, and dynamics. Chem Rev 1999;99:2161–2200.
19. Tsui V, Case DA. Molecular dynamics simulations of nucleic acids with a generalized born solvation model. J Am Chem Soc 2000;122:2489–2498.
20. Scarsi M, Apostolakis J, Caflisch A. Comparison of a GB solvation model with explicit solvent simulations: potential of mean force

and conformational preferences of alanine dipeptide and 1,2-dichloroethane. J Phys Chem B 1997;102:3636–3640.
21. Reddy MR, Erion MD, Agarwal A, Viswanadhan WN, McDonald DQ, Still WC. Solvation free energies calculated using the GB/SA model. J Comp Chem 1998;19:769–780.
22. Huber GA, McCammon JA. Weighted-ensemble simulated annealing: faster optimization on hierarchical energy surfaces. Phys Rev E 1997;55:4822–4825.
23. Levinthal C. Are there pathways for protein folding? J Chim Phys 1968;65:44–45. Onuchic JN, Luthey-Schulten Z, Wolynes PG. Theory of protein folding: the energy landscape perspective. Ann Rev Phys Chem 1997;48:545–600.
24. Dill KA, Chan HS. From Levinthal to pathways to funnels. Nat Struct Biol 1997;4:10–19.
25. Karplus M. The Levinthal paradox: yesterday and today. Fold Des 1997;2:S69–75.
26. Simmerling C, Lee M, Ortiz A, Kolinski A, Skolnick J, Kollman P. Combining MONSSTER and LES/PME to predict protein structure from amino acid sequence: Application to small Protein CMTI-1. J Am Chem Soc 2000;122:8392–8402.
27. Vorobjev YN, Hermans J. ES/IS: estimation of conformational free energy by combining dynamics simulations with explicit solvent with an implicit solvent continuum model. Biophys Chem 1999;78: 195–205. Lazaridis T, Karplus M. Discrimination of the native from misfolded protein models with an energy function including implicit solvation. J Mol Biol 1999;288:477–487.
28. Levy R. Scoring decoys using surface generalized born model. Private Communication; 2000.
29. Harrison RW, Chatterjee D, Weber IT. Analysis of six protein structures predicted by comparative modeling techniques. Proteins 1995;23:463–471.
30. Fetrow JS, Bryant SH. New programs for protein tertiary structure prediction. Biotechnology (N Y) 1993;11:479–484.
31. Dill KA, Thomas PD. Statistical potentials extracted from protein structures: how accurate are they? J Mol Biol 1996;257:457–469.
32. Liwo A, Oldziej S, Pincus MR, Wawak RJ, Rackovsky S, Scheraga HA. A united-residue force field for off-lattice protein structure simulations. I. Functional formas and parameters of long-range side-chain interaction potentials from protein crystal data. J Comp Chem 1997;18:849–873. Liwo A, Kazmierkiewicz R, Czaplewski C, Groth M, Oldziej S, Wawak RJ, Rackovsky S, Pinus MR, Scheraga HA. United-reside force field for off-lattice protein-structure simulations: III. Origin of backbone hydrogen-bonding cooperativity in united_residue potentials. J Comp Chem 1998;19: 259–276. Liwo A, Pinus MR, Wawak RJ, Rackovsky S, Oldziej S, Scheraga HA. A united-residue force field for off-lattice protein-structure simulations. II. Parameterization of short-range interactions and determination of weights of energy terms by z-score optimization. J Comp Chem 1997;18:574–887.
33. Hao MH, Scheraga HA. How optimization of potential functions affects protein folding. Proc Natl Acad Sci USA 1996;93:4984–4989.
34. Reva BA, Finkelstein AV, Sanner M, Olson AJ, Skolnick J. Recognition of protein structure on coarse lattices with residue-residue energy functions. Prot Eng 1997;10:1123–1130.
35. Ishikawa K, Yue K, Dill KA. Predicting the structures of 18 peptides using Geocore. Protein Sci 1999;8:716–721.
36. Scott WRP, Hunenberger PH, Tironi IG, Mark AE, Billeter SR, Fennen J, Torda AE, Huber T, Kruger P, van Gunsteren WF. The GROMOS biomolecular simulation program package. J Phys Chem 1999;103:3596–3607.
37. Doniach S, Eastman P. Protein dynamics simulations from nanoseconds to microseconds. Curr Opin Struct Biol 1999;9:157–163. Stocker U, van Gunsteren WF. Molecular dynamics simulation of hen egg white lysozyme: a test of the GROMOS96 force field against nuclear magnetic resonance data. Proteins 2000;40:145–153.
38. Miller JL, Cheatham III TE, Kollman PA. Simulation of nucleic acid structure. In: Neidle S, editor. Oxford handbook of nucleic acid structure. New York: Oxford University Press, 1999. p 95–115. BeveridgeDL, McConnell KJ. Nucleic acids: theory and computer simulation, Y2K. Curr Opin Struct Biol 2000;10:182–196.
39. Honig B, Nicholls A. Classical electrostatics in biology and chemistry. Science 1995;268:1144.
40. Gilson MK, Sharp KA, Honig BH. Calculating the electrostatic potential of molecules in solution: method and error assessment. J Comp Chem 1987;9:327–335.
41. Jayaram B, Liu Y, Beveridge DL. A modification of the generalized born theory for improved estimates of solvation energies and pKa shifts. J Chem Phys 1998;109:1465–1471.
42. Jayaram B, Sprous D, Beveridge DL. Solvation free energy of biomacromolecules: parameters for a modified generalized born model consistent with the AMBER force field. J Phys Chem 1998;102:9571–9576.
43. Rapp CS, Friesner RA. Prediction of loop geometries using a generalized born model of solvation effects. Proteins 1999;35:173–183.
44. Dominy BN, Brooks III CR. Development of a generalized Born model parameterization for proteins and nucleic acids. J Phys Chem B 1999;3765–3773.
45. Levinthal C. Are there pathways for protein folding? Chim Phys 1968;65:44–55.
46. Dill KA. Folding proteins: finding a needle in a haystack. Curr Opin Struct Biol 1993;3:99–103.
47. Honig B. Protein folding: from the levinthal paradox to structure prediction. J Mol Biol 1999;293:283–293.
48. Hansmann UH, Okamoto Y. New Monte Carlo algorithms for protein folding. Curr Opin Struct Biol 1999;9:177–183.
49. Unger R, Moult J. Genetic algorithms for protein folding simulations. J Mol Biol 1993;231:75–81.
50. Lee J, Liwo A, Ripoll DR, Pillardy J, Scheraga HA. Calculation of protein conformation by global optimization of a potential energy function. Proteins 1999;Suppl(3):204–208. Lee J, Liwo A, Scheraga HA. Energy-based de novo protein folding by conformational space annealing and an off-lattice united-residue force field: application to the 10-55 fragment of staphylococcal protein A and to apo calbindin D9K. Proc Natl Acad Sci USA 1999;96:2025–2030.
51. Mathiowetz AM, Jain A, Karasawa N, Goddard WA 3rd. Protein simulations using techniques suitable for very large systems: the cell multipole method for nonbond interactions and the Newton-Euler inverse mass operator method for internal coordinate dynamics. Proteins 1994;20:227–247.
52. Abagyan RA, Totrov M. Ab initio folding of peptides by the optimal-bias Monte Carlo minimization procedure. J Comp Phys 1999;151:402–421.
53. Wu X, Shaomeng W. Self-guided molecular dynamics simulation for efficient conformational search. J Phys Chem B 1998;102:7238–7250.
54. Foreman KW, Phillips AT, Rosen JB, Dill KA. Comparing search strategies for finding global optima on energy lanscapes. J Comp Chem 1999;20:1527–1532.
55. Leopold PE, Montal M, Onuchic JN. Protein folding funnels: a kinetic approach to the sequence–structure relationship. Proc Natl Acad Sci USA 1992;89:8721–8725. Onuchic JN, Socci ND, Luthey-Schulten Z, Wolynes PG. Protein folding funnels: the nature of the transition state ensemble. Fold Des 1996;1:441–450. Osguthorpe DJ. Improved ab initio predictions with a simplified, flexible geometry model. Proteins 1999;Suppl(3):186–193.
56. Simons KT, Bonneau R, Ruczinski I, Baker D. Ab initio protein structure prediction of CASP III targets using ROSETTA. Proteins 1999;37(S3):171–176.
57. Kawai H, Kikuchi T, Okamoto Y. A prediction of tertiary structures of peptide by the monte carlo simulated annealing method. Prot Eng 1989;3:85–94.
58. Karplus M, Kushick JN. Method for estimating the configurational entropy of macromolecules. Macromolecules 1981;14:325–332.
59. Weiner AJ, Kollman P, Case DA, Singh UC, Ghio C, Alagona G, Salvatore Profeta J, Weiner P. A new force field for molecular mechanical simulation of nucleic acids and proteins. J Am Chem Soc 1984;106:765–784.
60. Case DA, Pearlman DA, Caldwell JW, Cheatham III TE, Ross WS, Simmerling C, Darden T, Merz KM, Stanton RV, Cheng A, Vincent JJ, Crowley M, Tsui V, Radmer R, Duan Y, Pitera J, Massova I, Seibel GL, Singh UC, Weiner P, Kollman P. AMBER: Version 6 6.0 edit. San Francisco: University of California; 1999.
61. Hawkins GD, Cramer CJ, Truhlar DG. Pairwise solute descreening of solute charges from a dielectric medium. Chem Phys Lett 1995;246:122–129.
62. Wodak S, Janin J. Analytical approximation to the accessible

surface area of proteins. Proc Natl Acad Sci USA 1980;77:1736–1740.

63. Sharp K, Honig B. DelPhi 2.3 edit. San Diego: Biosym Technologies Inc., 1992.
64. Wang W. Molecular dynamics tool chest: an integrated software package for the analysis of molecular dynamics simulation results of biological macromolecules. MA Thesis, Wesleyan University; 2001.
65. Jayaram B, McConnell K, Dixit SB, Beveridge D L. Free energy analysis of protein-DNA binding: The EcoRI endonuclease-DNA complex. J Comput Phys 1999;151:333–357.
66. Sharp KA, Nicholls A, Friedman R, Honig B. Extracting hydrophobic free energies from experimental data: relationship to protein folding and theoretical models [published erratum appears in Biochemistry 1993 May 25;32(20):5490]. Biochemistry 1991;30: 9686–9697.
67. Spector DH. Building linux clusters. Beijing: O'Reilly; 2000.
68. Zemla A, Venclovas C, Moult J, Fidelis K. Processing and analysis of CASP3 protein structure predictions. Proteins 1999;Suppl(3):22–29.
69. Vorobjev YN, Almagro JC, Hermans J. Discrimination between native and intentionally misfolded conformations of proteins: ES/IS, a new method for calculating conformational free energy that uses both dynamics simulations with an explicit solvent and an implicit solvent continuum model. Proteins 1998;32:399–413. Vendruscolo M, Najmanovich R, Domany E. Can a pairwise contact potential stabilize native protein folds against decoys obtained by threading? Proteins 2000;38:134–148.
70. Samudrala R, Levitt M. Decoys 'R' Us: a database of incorrect conformations to improve protein structure prediction [In Process Citation]. Protein Sci 2000;9:1399–1401.
71. Darden T, York D, Pedersen L. The particle mesh Ewald method. J Chem Phys 1995;98:10089–10092.
72. Dill KA. Theory for the folding and stability of globular proteins. Biochemistry 1985;24:1501–1509.
73. Hill CP, Anderson DH, Wesson L, Degrado N, Eisenberge D. Crystal Structure of alpha1: Implications for protein Design. Science 1990;249:543–546.
74. Scholtz JM, York E, Stewwart J, Baldwin R. A neutral water-soluble, alpha-helical peptide: the effect of ionic strength on the helix-coil equilibrium. J Am Chem Soc 1991;113:5102–5104.
75. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Cryst 1993;26:283–291.
76. Lifson S, Roig AJ. On the theory of helix coil transitions in polypeptides. Chem Phys 1961;34:1963–1974. Qian H, Schellman JAJ. Helix coil theories: A comparative study for finite length polypeptides. Phys Chem 1992;96:3987.
77. Decatur SM, Antonic J. Isotope-edited infrared spectroscopy of helical peptides. J Am Chem Soc 1999;121:11914–11915.
78. Gronenborn AM, Filpula DR, Essig NZ, Achari A, Whitlow M, Wingfield PT, Clore GM. A novel, highly stable fold of the immunoglobulin binding domain of streptococcal protein G [see comments]. Science 1991;253:657–661.
79. McKnight CJ, Matsudaira PT, Kim PS. NMR structure of the 35–residue villin headpiece subdomain [letter]. Nat Struct Biol 1997;4:180–184.
80. Hubbard TJ. RMS/Coverage graphs: a qualitative method for comparing three-dimensional protein structure predictions. Proteins 1999;37(S3):15–21.
81. Onufriev A, Bashford D, Case DA. Modification of the generalized born model suitable for macromolecules. J Phys Chem B 2000;104: 3712–3720. Zhang L, Gallicchio E, Friesner R, Levy RM. Solvent models for protein-ligand binding: comparison of implicit solvent poisson and surface generalized born models with explicit solvent simulations. J Comp Chem 2001;22:591–607.
82. Brass A, Pendleton BJ, Chen Y, Robson B. Hybrid Monte Carlo simulations theory and initial comparison with molecular dynamics. Biopolymers 1993;33:1307–1315.
83. Kazmirski SL, Li A, Daggett V. Analysis methods for comparison of multiple molecular dynamics trajectories: applications to protein unfolding pathways and denatured ensembles. J Mol Biol 1999;290:283–304. McConnell KM, Nirmala R, Young MA, Ravishanker G, Beveridge DL. A nanosecond molecular dynamics trajectory for a B DNA double helix: evidence for substates. J Am Chem Soc 1994;116:4461–4462.