Ab Initio Prediction of the Three-Dimensional Structure of a De Novo Designed Protein: A Double-Blind Case Study

John L. Klepeis,[1†] Yinan Wei,[2‡] Michael H. Hecht,[2] and Christodoulos A. Floudas[1*]
[1]*Department of Chemical Engineering, Princeton University, Princeton, New Jersey*
[2]*Department of Chemistry, Princeton University, Princeton, New Jersey*

ABSTRACT **Ab initio structure prediction and de novo protein design are two problems at the forefront of research in the fields of structural biology and chemistry. The goal of ab initio structure prediction of proteins is to correctly characterize the 3D structure of a protein using only the amino acid sequence as input. De novo protein design involves the production of novel protein sequences that adopt a desired fold. In this work, the results of a double-blind study are presented in which a new ab initio method was successfully used to predict the 3D structure of a protein designed through an experimental approach using binary patterned combinatorial libraries of de novo sequences. The predicted structure, which was produced before the experimental structure was known and without consideration of the design goals, and the final NMR analysis both characterize this protein as a 4-helix bundle. The similarity of these structures is evidenced by both small RMSD values between the coordinates of the two structures and a detailed analysis of the helical packing. Proteins 2005;58:560–570.** © 2004 Wiley-Liss, Inc.

Key words: structure prediction; protein design; protein folding; binary patterning; four-helix bundle; optimization

INTRODUCTION

The idea that proteins spontaneously fold into their native, compact conformations is a fundamental tenet of research in the area of structural biology and chemistry. This observation was first established through the pioneering work of Anfinsen,[1] who showed that the same stable, or native, state of a protein could be attained, even after denaturation, through reintroduction of the protein into the original environment. These results led to the development of Anfinsen's thermodynamic hypothesis, which states that the amino acid sequence alone provides adequate information for finding the native conformation of a protein, since a protein in its surrounding environment attempts to minimize the free energy of the system. Therefore, the protein exists at the global minimum free energy state given a set of environmental conditions. In spite of several decades of research, the ability to fully explain and understand the mechanisms by which this protein folding occurs remains incomplete.

Knowledge of a protein's 3D structure has become even more important with the recent completion of various genome projects, including the elucidation of the human genome. The goal of research in the area of structural genomics is to provide the means to characterize and identify the large number of protein sequences that are being discovered. Although the structures of approximately 20,000 proteins have been determined by the experimental techniques of NMR and X-ray crystallography, and are catalogued in the PDB, there are thousands more to be discovered, each with a unique structure and special properties. Therefore, there is a great interest in developing computational approaches to correctly predict the 3D structure of proteins.[2] These approaches can be classified as (1) homology or comparative modeling methods, (2) fold recognition or threading methods, (3) ab initio methods that utilize knowledge-based information from structural databases (e.g., secondary and/or tertiary structure restraints), and (4) ab initio methods without the aid of knowledge-based information. The first 3 types of approaches rely on the use of databases to exploit information regarding secondary structure, distance constraints, and conformational preferences taken from sequence and structural alignments. True ab initio methods represent the most challenging, but also the most promising, approaches to solve the problem of predicting the structure of proteins, because these approaches do not depend on typical knowledge-based assumptions. Instead, complex atomic interactions are modeled by a semiempirical force

Abbreviations: 3D, three-dimensional; CASP, Critical Assessment of Techniques for Protein Structure Prediction; DSSP, Definition of Secondary Structure of Proteins given a set of 3D coordinates; MD, molecular dynamics; PDB, Protein Data Bank; PSI-BLAST, position-specific iterated basic local alignment search tool; RMSD, root-mean-square deviation.

The Supplementary Materials referred to in this article can be found at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/index.html

†Current address: D. E. Shaw Research and Development, L.L.C., 120 West Forty-Fifth Street, New York, NY 10036.

‡Current address: Department of Biology, Brookhaven National Laboratory, Upton, NY 11973.

Grant sponsor: National Science Foundation (to C. A. Floudas). Grant sponsor: National Institutes of Health; Grant number: R01 GM52032 (to C. A. Floudas). Grant sponsor: National Institutes of Health; Grant number: R01 GM062869 (to M. H. Hecht).

*Correspondence to: Christodoulos A. Floudas, Department of Chemical Engineering, Princeton University, Princeton, NJ 08544-5263. E-mail: floudas@titan.princeton.edu

Received 20 March 2004; Accepted 9 July 2004

Published online 17 December 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20338

field and the conformation of the system is optimized in order to locate the lowest energy, and thus the most stable structure for the protein.[3–6] The main obstacles for ab initio protein structure prediction include the validity of the available molecular models and the complexity of the search space due to the immense number of possible conformations accessible to the protein. These challenges can only be met through the application of powerful algorithms, along with experimentally accurate models.

Recently, a novel 4-stage ab initio approach, ASTRO-FOLD, was introduced for the structure prediction of single-chain polypeptides.[3,4] The methodology combines the classical and new views of protein folding, while using free energy calculations and integer linear optimization to predict helical and β-sheet structures, respectively. Detailed atomistic modeling and the deterministic global optimization method, αBB (a Branch-and-Bound deterministic global optimization method), coupled with torsion angle dynamics, form the basis for the final tertiary structure prediction. The agreement between experimental and predicted structures for a variety of benchmark and blind studies highlight the excellent performance of the ASTRO-FOLD approach for generic protein structure prediction. In particular, the method was found to be especially accurate in the ab initio prediction of secondary structure, in addition to the novel identification of β-sheet topologies.[7] Detailed performance analysis for the CASP experiment can be found elsewhere.[7]

It must be emphasized that ASTRO-FOLD is an entirely ab initio method. Thus, for the current study, neither the experimental structure nor any statistical information was incorporated into the prediction. Moreover, there was no attempt to bias the prediction by classification of the target; the prediction method was utilized without incorporation of any knowledge about whether the protein was designed to be α-helix or β-sheet.

A related problem to that of protein structure prediction is the design of de novo proteins. The ability to successfully design proteins tests the capacity to understand the relationship between the amino acid sequence of a protein and its 3D structure. The techniques used to discover these de novo proteins can be divided into experimental- and computational-based approaches.[8–21] Computational approaches resemble an inverse protein folding calculation in which the goal is to search sequence space by correctly modeling and determining the atomic interactions that best stabilize the structural features or properties for a given protein template. On the other hand, experimental approaches use the principles of rational design and/or combinatorial methods to produce proteins with the desired fold or properties. Experimentalists have applied the techniques of mutagenesis, directed evolution, and combinatorial and rational design to discover de novo proteins, although obtaining high-quality functional proteins remains a challenge.[22]

A method to enhance the success of combinatorial libraries has been described.[10,23–25] The basic premise of the approach is to produce focused libraries of novel proteins by integrating rational design and combinatorial methods. The libraries are generated such that the exact identities of polar and nonpolar residues are varied combinatorially, although the binary patterning of polar and nonpolar residues is designed rationally. This binary coding strategy helps to focus libraries of novel sequences, thereby favoring the formation of well-folded protein structures. We previously reported the design and construction of binary patterned libraries of both α-helical and β-sheet proteins.[24–26] Recently, we described a second-generation library of sequences designed to fold into 4-helix bundles of 102 residues.[23] Five proteins from this library were chosen for biophysical characterization. All 5 were found to be α-helical and stable; and 4 of the 5 formed structures that were well-ordered and/or nativelike.[23] This is particularly significant, since the 5 proteins were chosen arbitrarily from a naive library that had not been subjected to genetic selections or high throughput screens. Therefore, we presume that stably folded protein structures occur quite frequently in this library. More recently, the first high-resolution structure of a protein from this library was determined by NMR–spectroscopy.[10] The experimentally determined structure matches that expected from design: It is a 4-helix bundle with nonpolar side-chains buried in the protein interior and polar side-chains exposed to solvent.

In this article, we present the ab initio prediction of the structure of a de novo designed protein. The prediction was done as a double-blind study: ASTRO-FOLD was used to predict the structure of protein S-824 prior to the experimental determination of the actual solution structure. The prediction was done without knowledge of the experimental data and without consideration of the design goals. To ensure that the prediction was "blind," the sequence of S-824 was provided to Klepeis and Floudas in January 2001—before any structural information was available from NMR studies. The NMR structure was solved 2 years later, in January 2003, by Wei, Hecht, and coworkers.[10] The coordinates of the NMR structure were made available to Klepeis and Floudas only after they sent the coordinates of the predicted structure to Wei and Hecht.

This report describes the first ab initio prediction of the structure of a de novo protein from a designed combinatorial library. This is also the first application of the ASTRO-FOLD methodology to an α-helical structure.

Protein S-824 was ultimately found to be a 4-helix bundle. The ab initio prediction for the S-824 protein closely matches the experimental structure. In this article, we provide brief descriptions of the ASTRO-FOLD approach and of the de novo design strategy, followed by a detailed analysis of the ab initio prediction of the structure of protein S-824.

THEORY AND MODELING

ASTRO-FOLD: Ab Initio Structure Prediction of Proteins

ASTRO-FOLD, a 4-stage hierarchical approach for the ab initio prediction of the 3D structures of proteins, employs modeling and optimization techniques to reconcile competing explanations of protein folding.[27–29] The

classical view regards folding as hierarchical, implying that the process is initiated by rapid formation of secondary structural elements, followed by the slower arrangement of the tertiary fold. The opposing perspective is based on the idea of a hydrophobic collapse, and suggests that tertiary and secondary features form concurrently. Two important components of the approach are the ideas that helix nucleation is controlled by local interactions, while nonlocal hydrophobic forces drive the formation of β structure. A formulation that combines both concepts is used to predict the overall tertiary structure.

Helix Prediction

The first stage of the ASTRO-FOLD approach involves the prediction of helical segments and is accomplished by partitioning the overall target sequence into oligopeptides such that consecutive oligopeptides possess an overlap of $N - 1$ amino acids (where N is the length of the oligopeptide); atomistic-level modeling using the selected force field; generating an ensemble of low-energy conformations; calculating free energies that include entropic, cavity formation, polarization, and ionization contributions for each oligopeptide; and calculating helix propensities for each residue using equilibrium occupational probabilities of helical clusters. The concept of partitioning the protein sequence into overlapping oligopeptides is based on the idea that helix nucleation relies on local interactions and positioning within the overall sequence. The explicit consideration of local interactions through overlapping oligopeptides allows for detection of cases in which identical amino acid sequences adopt different conformations in different proteins.[30] This is consistent with the observation that local interactions extending beyond the boundaries of the helical segment retain information regarding conformational preferences.[29] The partitioning pattern is generalizable and can be extended to oligopeptides of any length, although typically pentapeptides are preferred, since they are the smallest systems that still capture the i to $i + 4$ hydrogen-bonding pattern in helices. Specifically, for a protein of length N, there will be $N - 4$ overlapping pentapeptides, $N - 6$ heptapeptides, and so forth.

The overall methodology for the ab initio prediction of helical segments encompasses the following steps:

1. The overlapping oligopeptides are modeled as neutral peptides surrounded by a vacuum environment using the ECEPP/3 force field.[31] Side-chains are modeled explicitly (i.e., all-atom). An ensemble of low potential energy pentapeptide conformations, along with the global minimum potential energy conformation, is identified using a modification of the αBB global optimization approach[32] and the conformational space annealing approach.[33] For the set of unique conformers, $\mathcal{X}$, free energies ($(F_{\text{vac}}^{\text{har}})$ are calculated using the harmonic approximation for vibrational entropy.[32] Although not a free energy in the sense of classical MD, this approximation provides a means for calculating a relative free energy based on the unique clustered conformers in the ensemble.
2. The energy for cavity formation in an aqueous environment is modeled using a solvent-accessible surface area expression, $F_{\text{cavity}} = \gamma A + b$, where A is the surface area of the protein exposed to the solvent.
3. For the set of unique conformers, $\mathcal{X}$, the total free energy is calculated from

$$F_{\text{total}} = F_{\text{vac}}^{\text{har}} + F_{\text{cavity}} + F_{\text{solv}} + F_{\text{ionize}}. \quad (1)$$

Here F_{solv} represents the difference in polarization energies caused by the transition from a vacuum to a solvated environment, and F_{ionize} represents the ionization energy. These energies are calculated through the solution of the Poisson–Boltzmann equation[34] for all unique conformers. The set of unique conformers ($\mathcal{X}$) is determined by removing all duplicate and symmetric minima, as well as those that do not differ by more than 50° for at least one dihedral angle (disregarding the first and last backbone dihedral angles and the last dihedral angle in each side-chain).

4. For each oligopeptide, total free energy values (F_{total}) are used to evaluate the equilibrium occupational probability for conformers having 3 central residues within the helical region of the ϕ–ψ space. Helix propensities for each residue are determined from the average probability of those systems in which the residue constitutes a core position.

β-Strand and β-Sheet Topology Prediction

In the second stage, β-strands, β-sheets, and disulfide bridges are identified through a novel superstructure-based mathematical framework, a concept originally employed for the solution of chemical process synthesis problems.[35] Two types of superstructure have been introduced, both of which emanate from the principle that hydrophobic interactions drive the formation of β structure. The first one, denoted as *hydrophobic residue-based superstructure*, encompasses all potential contacts between pairs of hydrophobic residues (i.e., a contact between 2 hydrophobic residues may or may not exist) that are not contained in helices (except cystines, which are allowed to have cystine–cystine contacts even though they may be in helices). The second one, denoted as β-*strand-based superstructure*, includes all possible β-strand arrangements of interest (i.e., a β-strand may or may not exist) in addition to the potential contacts between hydrophobic residues. Implementation of ASTRO-FOLD predicted that protein S-824 did not contain any β structure. Therefore, we will not describe those parts of ASTRO-FOLD that pertain to β structure. For a full description of this approach, the reader is referred to earlier studies.[3]

Tertiary Structure Prediction

The third stage of the approach serves as a preparative phase for the atomistic-level tertiary structure prediction by deriving appropriate constraints based on the results of the previous two stages. Specifically, this involves the introduction of lower and upper bounds on dihedral angles of residues belonging to predicted helices or β-strands, as well as restraints between the C^{α} atoms for residues of the

selected β-sheet and disulfide bridge configuration. Furthermore, free energy runs of overlapping oligopeptides are performed when possible to developed tighter bounds on the conformations of the loop residues that connect the elements of predicted secondary structure.

The fourth and final stage of the ASTRO-FOLD approach involves the prediction of the tertiary structure of the full protein sequence. The problem formulation, which relies on dihedral angle and atomic distance restraints acquired from the previous stage, is

$$\min_{\phi} E_{\text{ECEPP/3}},$$

$$\text{subject to } E_l^{\text{distance}}(\phi) \le E_l^{\text{ref}}\ l = 1,\ldots, N_{\text{CON}},$$

$$\phi_i^L \le \phi_i \le \phi_i^U,\ i = 1,\ldots, N_\phi.$$

Here $i = 1,\ldots, N_\phi$ refers to the set of dihedral angles, ϕ_i, with ϕ_i^L and ϕ_i^U representing lower and upper bounds on these variables that are used to generate a 3D conformation of the protein. The total violations of the $l = 1,\ldots,N_{\text{CON}}$ distance constraints are controlled by the parameters E_l^{ref}.[36] To overcome the multiple minima difficulty, the search is conducted using the αBB global optimization approach, which offers theoretical guarantee of convergence to an ϵ global minimum for nonlinear optimization problems with twice-differentiable functions.[37,38] This global optimization approach effectively brackets the global minimum by developing converging sequences of lower and upper bounds, which are refined by iteratively partitioning the initial domain. Upper bounds correspond to local minima of the original nonconvex problem, while lower bounds belong to the set of solutions of convex lower bounding problems, which are constructed by augmenting the objective and constraint functions by separable quadratic terms. To ensure nondecreasing lower bounds, the prospective region to be bisected is required to contain the infimum of the minima of lower bounds. A nonincreasing sequence for the upper bound is maintained by selecting the minimum over all the previously recorded upper bounds. The generation of low energy starting points for constrained minimization is enhanced by introducing torsion angle dynamics[39] within the context of the αBB global optimization framework. The αBB has been successfully applied to computational chemistry problems, including microclusters, small acyclic molecules, and isolated and solvated oligopeptides.[38]

De Novo Protein Design Strategy

The binary code strategy for protein design has been described in detail elsewhere.[10,23–25] In essence, this strategy represents a fusion of combinatorial methods with rational protein design. Large combinatorial libraries are produced; however, the sequences are not generated randomly. Instead, the libraries are focused into productive regions of "sequence space" by designing the binary patterning of polar and nonpolar residues to ensure that all sequences in the library are consistent with the formation of specified amphiphilic secondary structural elements.[24,25] Since α-helices and β-strands each possess a characteristic structural periodicity, it is necessary to rationally design the polar and nonpolar patterns in the linear sequences of the library to match the periodicity that codes for the desired amphiphilic secondary structural elements. For α-helices, a helical turn repeats every 3.6 residues, which translates to a binary code that places a nonpolar (*N*) residue every 3 or 4 positions. For example, *PNPPNNPPNPPNNP* represents a design pattern for an amphiphilic α-helix. On the other hand, the design of amphiphilic β-strands requires only a 2-residue pattern, with an alternating code of polar and nonpolar residues like *PNPNPNP*.

With these rational design goals in place, the binary polar–nonpolar code of the designed sequences is specified; however, the actual identities of the corresponding residues are not restricted. In other words, the patterns are combinatorially complex, and many sequences are compatible with the particular design goals. Incorporation of this diversity into actual libraries of sequences is made possible by the organization of the genetic code. Specifically, 5 nonpolar amino acids (Met, Leu, Ile, Val, and Phe) can be encoded by the degenerate codon NTN, while the degenerate codon VAN provides 6 polar amino acids (Lys, His, Glu, Gln, Asp, and Asn). N represents the DNA bases A, G, C, and T, while V represents A, G, or C. Such binary patterning of polar and nonpolar amino acids has been successfully employed in the design of a number of focused libraries of both α-helical and β-sheet proteins.[23–26] Although characterization of these libraries has qualitatively verified the achievement of the prescribed design goals, only recently has validation been obtained at high resolution through the determination of the solution structure for S-824, a 4-helix bundle protein.[10]

RESULTS AND DISCUSSION

Qualitative Structural Analysis for Protein S-824

As described previously, the solution structure of S-824 was determined by NMR and shown to be an up-down-up-down 4-helix bundle.[10] The experimentally determined structure is extremely well ordered—even by the standards of natural proteins.[10] Protein S-824 was chosen from a library with the potential for enormous combinatorial diversity. Because S-824 was chosen arbitrarily, without high throughput screens or selections, we presume that S-824 is not a rare "needle in a haystack" but rather is fairly typical of the kind of proteins present in the library as a whole. Therefore, it seems likely that the binary code strategy used to produce S-824 is sufficient to generate an enormous number of well-ordered and nativelike de novo proteins.

Prior to our having any knowledge of the experimental results, ASTRO-FOLD was used to predict the structure of protein S-824. The first stage of this ab initio approach involves the prediction of whether or not helical segments exist, and their initiation and termination sites if they do exist, and requires the partitioning of the 102-residue sequence into a set of 98 overlapping pentapeptides. For each pentapeptide, detailed free energy calculations were performed, and the ensembles of low-energy conformers

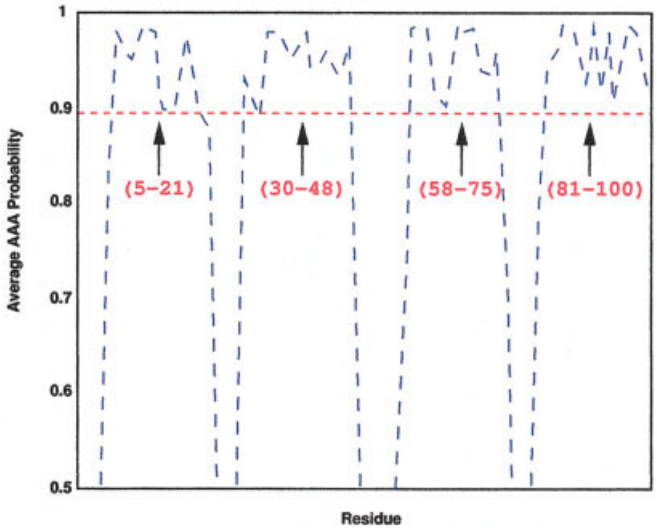

Fig. 1. Average probability of central 3 residues of cascading pentapeptide as a function of residue number. The identification of a helical segment corresponds to average helical propabilities exceeding 90% for more than 3 consecutive residues. For S-824, helical segments are predicted between residues 5–21, 30–48, 58–75, and 81–100. The 102 amino acid S-824 sequence is as follows: MYGKLNDLLED-LQEVLKNLHKNWHGGKDNLHDVDNHLQNVIEDIHDFMQGGGSGGKL-QEMMKEFQQVLDELNNHLQGGKHTVHHIEQNIKEIFHHLEELVHR.

were used to calculate helix propensities for each residue. The total computational effort for this stage of the approach corresponds to approximately 2 wallclock days on a fully utilized cluster of 80 CPUs (running Linux on Pentium III 600 Mhz processors). The final assessment is made according to average probabilities, and the results are depicted in Figure 1. For the S-824 sequence, helical segments were predicted to occur between residues 5–21, 30–48, 58–75, and 81–100. These initial predictions, which provide information on the location of helices in the S-824 sequence, agree with the de novo design goals, and clearly define a system that comprises 4 helical segments.

Because 4 helices were predicted strongly and the segments between the predicted helices are devoid of hydrophobic residues, the β-strand and β-sheet protocol was not applicable. In addition, these loop segments are relatively short and glycine-rich, a characteristic that promotes conformational flexibility. As a result, only the α-helix prediction results were used to constrain the system for tertiary structure prediction. The variable domains for those dihedral angle of residues predicted to be helical were bounded between ([−85, −55] for ϕ, [−50, −10] for ψ). Distance restraints included 58 lower and upper C^{α}–C^{α} (5.5–6.5 Å) bounds to enforce the hydrogen-bonding network within these α-helices. As the predictions did not include any β-sheet structure, the tertiary structure prediction was not constrained by any long-range contacts.

During the course of the global optimization search, the branch-and-bound search tree was formed by partitioning domains belonging to selected backbone variables of the loop segments, while the remaining variables were treated locally. A significant sample of low-energy structures was

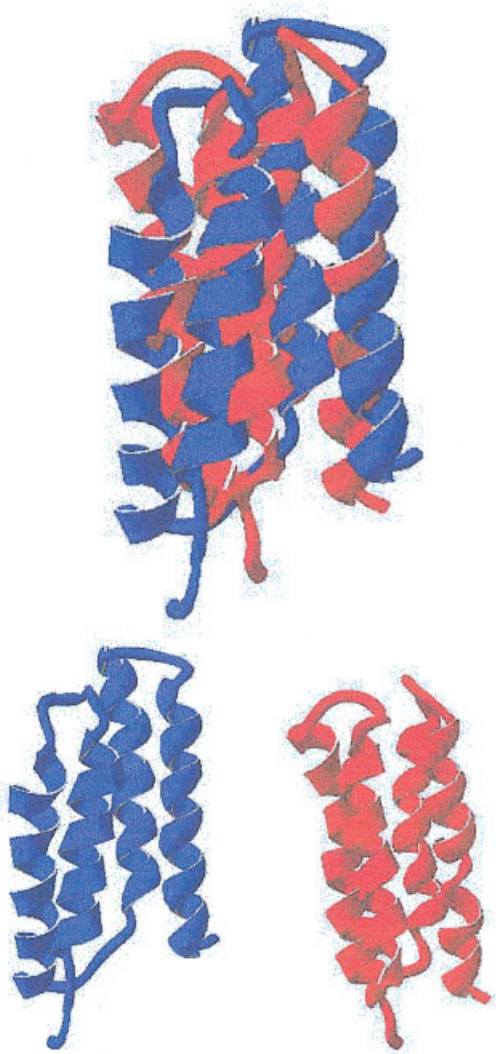

Fig. 2. Superposition of predicted and experimental structures for protein S-824. The experimental structure corresponds to the lowest energy structure from the 10 low-energy structures calculated from the NMR data.[10] The predicted structure is shown in blue. Ribbon diagrams created with SwissPDBViewer 3.7.[40]

identified (the total computational effort for this stage of the approach corresponds to approximately 4 wallclock days on a fully utilized cluster of 80 CPUs running Linux on Pentium III 600 Mhz processors). Using only the criterion of lowest energy, the predicted native structure provided by ASTRO-FOLD is that of an up-down-up-down 4-helix bundles. Qualitatively, this result agrees with both the design goals[23] and the experimental structure,[10] as shown in Figure 2. The final locations of the 4 helices in the predicted structure changed slightly when compared to the results of the α-helix prediction stage. To some extent, the determination of final helix content is dependent on the definitions used, and different methods based on either dihedral angle value or backbone hydrogen bonding may provide different results. A number of methods were used, and the consensus of these results (consistent with DSSP analysis when using 3D) are presented in Table I, along

TABLE I. Location of the Four Helices According to Experimental Results, Helix Prediction Results, and Final Tertiary Structure Prediction Results

Helix	Experiment	Helix Prediction	Tertiary Prediction	PSIPRED	SAM-T02
1	5–20	5–21	5–21	5–22	5–21
2	28–48	30–48	30–49	30–49	30–49
3	56–72	58–75	56–75	57–76	55–75
4	80–99	81–100	81–100	80–100	83–100

with the locations of the helices in the experimental structure. The main observation is that the helices are identical, with only slight differences in the initiation and termination of the 4 helices. Table I also presents the location of helices predicted by other statistical methods. Both the PSIPRED[41,42] and SAM-T02[43] predictions identify 4 helices in good agreement with the ASTRO-FOLD physics-based predictions, as well as the experimental structure. Overall, these results suggest that accurate secondary structure for S-824 should be relatively easy to obtain. However, these results are not necessarily trivial, as other prediction methods are less sucessful. For example, Robetta-JUFO-3D,[44] a secondary structure predictor that uses Robetta de novo decoys and comparative models in addition to PSI-BLAST multiple sequence information and an amino acid property profile to produce 3-state predictions, substantially underpredicts (by more than 3 residue on the N- and C-termini) the lengths of the second and third helices. PROFsec,[45] an improved version of PHDsec, predicts only 2 long helices, by merging the first helix with the second, and the third helix with the fourth. The best statistically based secondary structure predictions methods do approximately as well as the ASTRO-FOLD physics-based predictions; however, it is difficult to choose the best among these predictions a priori.

Examination of the overall 3D structure reveals other features that are consistent with the experimental structure. These results are especially interesting, because some of these features were not part of the original goals of the design, and in fact could not have been designed a priori using a strategy that incorporates combinatorial methods. For example, the overall topology of the bundle for both the predicted and experimental structures is left-turning. In general, right-turning topologies are more abundant in natural 4-helix bundle proteins, although neither topology was explicitly specified in the design of S-824. The orientations of the helices were also not part of the binary code design, but the values for the angles between the helices are quite similar for the predicted and experimental structures. In particular, helices 1 and 2, and helices 3 and 4 are roughly antiparallel in both the experimental and predicted structures. On the other hand, the angles between helices 1 and 4 and between helices 2 and 3 are approximately 20°, a characteristic of the packing of natural α-helical proteins. This combination of packing angles has also been observed in other 4-helix bundle proteins. Finally, as expected, the burial of hydrophobic side-chains is a common feature of both structures. The experimental structure exhibits tight packing of nonpolar side-chains, whereas the predicted structure is somewhat more loosely packed.

Comparison to Other Methods

As previously described, the ASTRO-FOLD approach relies on ab initio principles in the prediction of protein structures, making it applicable to any protein regardless of representation by sequence or structural homologs. However, many methods attempt to exploit such database information and, for completeness, homology and threading analyses were performed for the S-824 sequence. First, sequence homology was tested against the PDB using GenTHREADER.[41] No high-confidence predictions were found to exist, and the best alignments did not exceed 20% sequence identity. Next, a more rigorous fold recognition check was conducted using multiple profiles and predicted secondary structure.[46] Although a number of predictions could be classified with medium confidence, none of the most probable topologies correspond to structures with 4-helix topologies.

More recently, the applicability of fold recognition methods has been pushed even further through the use of metaservers, which combine the results of individual server predictions to arrive at a consensus ranking of the most likely predictions. The performance of these metaservers, as well as the individual servers, has been monitored through the LiveBench experiments.[47] A particularly consistent method [available online (http://bioinfo.pl) and similar to Pcons[48] and 3D-Shotgun[49] servers] is the 3D-Jury server,[50] which can actually perform its analysis as a meta-metaserver. The 3D-Jury method assigns scores based on a similarity measure that counts α-carbon pairs within 3.5 Å deviations after alignment; a score greater than 50 is generally considered meaningful. The highest scoring prediction for S-824 generates a score less than 35, and although many of the matches correspond to all helical structures, none of the top 20 matches exhibit the correct 4-helix bundle topology. In fact, many of the predicted structures have only 2 or 3 packed helices. Although the prediction with the best 3D structural alignment (after knowing the 3D structure) has 4 helices, only 3 of these helices are in a parallel–antiparallel packed arrangement. These results suggest that ASTRO-FOLD, as an ab initio method, can be important for the prediction of protein structures (including the 4-helix bundle of S-824).

The results presented above confirm the applicability of ASTRO-FOLD in a qualitative sense. As shown in the

following section, quantitative analyses show that the ASTRO-FOLD results are also highly significant.

Quantitative Structural Analysis for Protein S-824

RMSDs can be calculated to give a more rigorous quantification of the similarity between the experimental and predicted structures of protein S-824. When considering only backbone atoms, the RMSD over all 102 residues between the lowest energy predicted and experimental structures is 4.94 Å. In general, backbone RMSDs below 6 Å constitute very good predictions, especially for protein systems with more than 100 residues.[51–53]

It is also interesting to place these results in the context of related work. Often helix-bundle proteins are used as test systems because of their substantially large hydrophobic cores; in particular, they are useful for testing reduced model calculations. In this vein, Friesner and coworkers used a reduced model potential and a branch and bound algorithm (actually based on the principles of the αBB) to successfully predict the structure of a 4-helix bundle protein to within 4 Å backbone deviation.[54] There are several important differences in the work presented here. First, the 4-helix bundle protein sequence of Friesner and coworkers is three-fourths the length of S-824. In addition, the authors completely freeze secondary structure (based on the experimental structure) and represent the sidechains of residues by single points. Finally, although the lowest energy cluster correlates well to the native topology, the authors note that this is not trivial, as other topologies may also have competitive energies. Another branch-and-bound algorithm, again based on the same αBB principles, was also successfully used to predict structures of larger proteins to within 6-Å RMSDs, while again maintaining secondary structure fixed.[55] These calculations differ further in that NMR data and some sparse tertiary contact data were used to further restrain the system.

More recently, a method for packing helices using reduced models and global optimization was presented.[56] These calculations rely on a very coarse-grained contact potential, while fixing the secondary structure content. Again, nativelike folds are reproduced with reasonable accuracy; for systems with 4 helices, the best RMSD values are typically between 4 Å and 5 Å, although values are somewhat higher in certain cases. However, the analysis does not necessarily use the lowest energy criterion, so there is certainly room for improvement in going to all-atom physics based models such as ASTRO-FOLD.

Deviations in individual components of the overall structure can also be computed. The results of this analysis, which are reported in Table II, reveal several important facts. First, all α-helical segments exhibit good RMSD values, with no segment above 2.5 Å for backbone deviations. In fact, helix 1 (H1) and helix 4 (H4) have small deviations (1.14 Å and 0.84 Å, respectively). Conversely, the loop segments, which are much shorter than the helical segments, exhibit backbone RMSD values between 2 Å and 3 Å. Furthermore, when considering all atoms, only one loop (L2), which connects H2 and H3, gives a reasonable RMSD value. These observations highlight the fact that loop prediction remains a bottleneck in the accurate prediction of 3D structures of proteins. Note, however, that very few experimental restraints were available for the NMR structure refinement of the loops. Therefore, the exact structures of the loops in the experimental structure are known with far less certainty than the overall structure.

TABLE II. RMSDs Between Experimental and Predicted Structures of S-824

ID	Position	Length	BB	C^{α}
N-term	1–4	4	1.39	1.21
H1	5–20	16	1.14	1.29
L1	21–27	7	2.87	3.22
H2	28–48	21	2.37	2.57
L2	49–55	7	2.32	2.43
H3	56–72	17	2.06	2.30
L3	73–79	7	3.00	3.45
H4	80–99	20	0.84	0.89
C-term	100–102	3	0.97	0.12
All	1–102	102	4.94	5.18

Following the first column, which provides an identifier of the segment, columns 2 and 3 provide the location of the segment in the sequence and the number of residues, respectively. RMSDs are provided for backbone (BB) and C^{α} atoms.

TABLE III. RMSDs Between Experimental and Predicted Structures of S- 824

ID	Length	BB	C^{α}
H1, H2	37	3.57	4.15
H2, H3	38	4.84	5.17
H3, H4	37	2.25	2.41
H1, H4	36	4.29	4.54
H1, H3	33	3.98	4.34
H2, H4	41	5.05	5.33
H1, H2, H3	54	4.47	4.85
H1, H3, H4	53	4.08	4.31
H2, H3, H4	58	4.76	5.00
L1, L2, L3	21	4.85	5.11
H1, H2, H3, H4	74	4.79	5.07

Following the first column, which provides an identifier of the combination of segments, column 2 indicates the number of residues. RMSDs are provided for backbone (BB) and C^{α} atoms.

RMSD analyses can also be performed for combinations of structural segments, as shown in Table III. For pairwise combinations of helices, it is evident that the combination of H3 and H4 provides the lowest RMSD value, implying that the relative orientation between these 2 helices closely matches that of the experimental structure. On the other hand, helices 2 and 3 combine to give the worst RMSD value for consecutive helix pairs, followed only by the nonconsecutive combination of H2 and H4, which have a backbone RMSD value of 5.05 Å. As expected, for combinations of 3 helices, the pair, H3 and H4, in addition to H1, combine to give the best RMSD value for 3 helices. Finally, it is interesting to note that the RMSD value for all 4 helices is lower than that of the 3 loops, even though

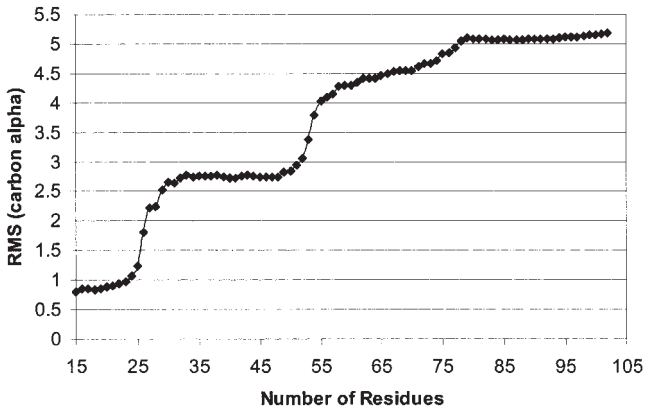

Fig. 3. Smallest RMSD for the longest continuous segment (LCS) between predicted and experimental structures. The RMSD values are plotted versus the length of the sequence for the segment providing that minimum RMSD value; that is, the segments displaying the minimum RMSD for a given length are continuous segments. However, the segments are not ordered or even overlapping; for example, the segment for one length (say, $N = 26$ residues) may be from a different part of the protein than for the segment of the next longer length (say, $N = 27$ residues).

the loop residues constitute only one-fifth of the sequence, as compared to three-fourths for the helices.

A summary of the RMSD analysis is seen in Figure 3. (Note: A full contour mapping of the RMSD values of C^{α} coordinates for all possible residue segments is provided in Fig. 5 in the Supplementary Material).

Figure 3 presents a trace of the smallest RMSD values as a function of longest continuous segment length. For segment lengths less than 25 residues, the longest continuous segments have RMSD values below 1 Å, and these segments correspond to the part of the sequence coding for H4. There is a substantial step change when considering the addition of the next helix, which is H3, although for these segments, the RMSD values remain below 3 Å. For segments longer than half the length of the overall sequence, there is a second step change, after which the smallest RMSD values increase linearly, as different combinations of helical segments contribute to this lowest RMSD envelope.

Energetic Analysis for Protein S-824

A detailed energetic analysis is also an important component when evaluating the quality of the predicted structure. In particular, this analysis can be used to validate the ability of the search technique to find low energy structures by comparing the force field energy of the predicted structure to the force field energy of the experimental structure. Since the premise of the structure prediction methodology is to locate the global minimum free energy state of the protein, the difference in these energies is a gauge on the performance of the approach. Comparison of these energies allows one to assess whether the method has failed to locate a lower energy cluster represented by the experimental structure (and therefore a more accurate prediction).

A variety of force fields exist for evaluating the energy of a given protein's structure. In particular, the GROMOS[57] and EEF1[58] (CHARMM-based[59]) force field were selected for this analysis, in addition to the ECEPP/3 force field,[31] which was used in the ASTRO-FOLD ab initio prediction approach. GROMOS is a standard force field used in protein simulation and design, and EEF1, which is based on the CHARMM force field, has been found to be especially efficient in discerning the native state of a protein among many other low-energy decoys.[60]

The detailed results for the GROMOS and EEF1 analyses are provided as Table IV in the Supplementary Material. For both scenarios, the initial energies of the predicted structure were substantially lower than those of the experimental structure. This is because the exact location of the atoms can greatly influence the energy values based on the particular parameterization of atomic interactions, and the predicted structure is a better representative of virtual systems. In order to provide a more accurate picture, we allowed the systems to equilibrate to a metastable state by performing 300 steps of energy minimization. For the GROMOS force field, 150 steps of steepest descent were followed by 150 steps of conjugate gradient minimization. In the case of EEF1, the minimization protocol entailed 300 steps of an adopted-basis Newton–Raphson method. Although the energy gaps between the experimental and predicted structures were reduced, the predicted structure still retained lower energies for both force fields. In addition, the structures exhibited only minor conformational changes upon minimization, as evidenced by the similar RMSD values between the predicted and experimental structures before and after minimization (shown in Table V of Supplementary Material). These results validate the performance of the search techniques employed in the ASTRO-FOLD approach.

A more elaborate protocol is needed to evaluate the ECEPP/3 energy of the experimental structure. The energy of the predicted structure, which corresponds to the putative global minimum energy value, is −846.4 kcal/mol. However, because the ECEPP/3 force field relies on an internal coordinate system rather than the Cartesian coordinate system, a direct evaluation of the energy of the experimental structure is not possible. The main problem is that the fixed bond length and bond angles assumptions used in the dihedral angle coordinate system of the ECEPP/3 force field do not exactly match those bond lengths and angles in the experimental structure. To overcome this difficulty, the back-calculated dihedral angles were subjected to parametric variation through the solution of a nonlinear minimization in which the RMSD between the experimental and the ECEPP/3-generated structure was minimized. This process was repeated until a structure with an all-atom RMSD of 0.57 Å was obtained (see Table VI in Supplementary Material). The goal was to minimize this deviation, and the nonlinear objective function corresponds simply to the RMSD between the ECEPP/3-generated structure and the experimental structure; that is, no additional energy function was used. Although a discrete rotamer library was used to reduce atomic clashes, the starting energy was still rather high. Following minimization using the ECEPP/3 force field, the energy was

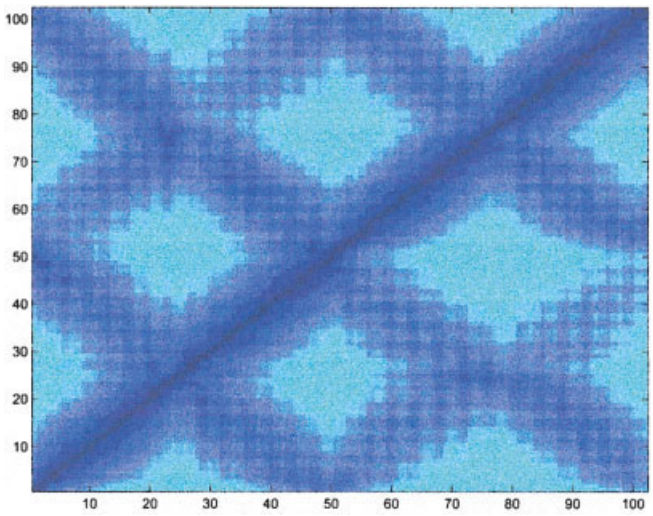

Fig. 4. Contact map comparison between C^{α} coordinates. The upper left triangle corresponds to interatomic distances calculated from the experimental structure, while the lower right triangle corresponds to those derived from the predicted structure. An upper distance cutoff of 30 Å was used in order to emphasize small interatomic distances. The progression from small to large distances follow the dark to light shading.

reduced to −755.2 kcal/mol, although the structure had shifted away from the experimental structure by 2.0 Å. The energy gap between the experimental and predicted structures was 100 kcal/mol (see Table VI in Supplementary Material).

Distance Analysis for Protein S-824

The final set of analyses involves the comparison of intraprotein distances in the experimental and predicted structures. This analysis is particularly important for understanding the packing of the individual secondary structure elements. A qualitative summary of the analysis is given in Figure 4, which provides a comparative distance map for C^{α}–C^{α} distances (a C^{β}–C^{β} distance map is provided as Figure 6 in the Supplementary Material). The distance values are shaded such that small distance are dark, and the left side of the diagonal corresponds to distances within the experimental structure, and the right side of the diagonal corresponds to distances within the predicted structure.

The contact maps verify the overall agreement in the tertiary arrangements of the experimental and predicted structures. In particular, the dark shaded regions perpendicular to the diagonal represent the formation of antiparallel helical matches (H1 to H2, H2 to H3, H3 to H4, and H1 to H4), while the shaded regions parallel to the diagonal represents the parallel interactions between H1 and H3, and H2 and H4. Based on the darker shading for the antiparallel contacts, the antiparallel packing of helices is tighter for S-824, although the lines parallel to the diagonal also indicate that some of the cross-parallel helical interactions are also relatively close. In addition, some subtle differences can be identified through closer examination of these distance maps. For example, the match between H2 and H3, and the parallel cross-helical contacts are represented by much more narrowly and lightly shaded regions for the predicted structure. Generally, the matches in the predicted structure are indicated by slightly lighter shadings, which verifies the looser tertiary packing in the predicted structure.

A quantitative analysis of the packing of these helices was performed (detailed results are given in Tables VII and VIII in the Supplementary Material). In this analysis, the number of hydrophobic to hydrophobic contacts within certain distance ranges were counted, and the percentage of these contacts coming from a particular helix-to-helix packing were calculated. These percentages can be used to discern which of the helices are most tightly packed. As expected, the antiparallel hydrophobic-to-hydrophobic matches (H1 to H2, H2 to H3, H3 to H4, and H1 to H4) dominate the contacts in both the predicted and experimental structures. However, the relative rankings of these matches are not identical. In the experimental structure the C^{α} interactions are overrepresented by contacts between H1 and H4, and H2 and H3, especially at small distances. When considering C^{β} distances, the distribution of helical contacts is almost uniform between all antiparallel helical matches in the experimental structure. On the other hand, in the predicted structure, the antiparallel helical contacts between H1 and H2, and H3 and H4 each account for one-third of the hydrophobic-to-hydrophobic interactions for both the C^{α} and C^{β} distances. The contacts between H1 and H4 are also well represented; however, the last antiparallel match, between H2 and H3, only contributes a small number of hydrophobic-to-hydrophobic interactions. These differences possibly reflect the lack of explicitly imposed tertiary contacts for purely helical proteins, and the ability to correctly predict such matches may enhance the performance of the ASTRO-FOLD approach.

CONCLUSIONS

Two important problems in the field of protein science are structure prediction and de novo protein design. Recently, a de novo design approach based on the integration of concepts from both rational design and combinatorial methods was used to successfully design sequences for 4-helix bundle proteins. The design involves the binary coding of nonpolar and polar residues in order to favor certain secondary structural elements. Prior to knowledge of the actual experimental structure, and without incorporation of either experimental or statistical-based structural information, a double-blind study was conducted, in which the ASTRO-FOLD ab initio approach was used to predict the 3D protein structure for S-824.

The ASTRO-FOLD method bridges the gap between the competing explanations for protein folding by employing a framework in which helix nucleation is controlled by local interactions, while nonlocal hydrophobic forces drive the formation of tertiary contacts, such as in the formation of β structure. Qualitative agreement between the experimental and predicted structure for protein S-824 is impressive, and features not expressly incorporated into the design were correctly predicted. Rigorous quantitative analyses

validate the extent of this agreement, and energy analyses verify the performance of the underlying modeling and search protocols. The current work relies on all-atom physics-based prediction of both secondary and tertiary structure, and so differs from the reduced model calculations that have been used with some success in the structure prediction of helical proteins. Furthermore, although 4-helix bundle proteins are typically well behaved systems, because of their large hydrophobic cores (and sometimes even well-predicted by reduced models), it is also shown that the structure prediction of S-824 by statistical methods is not trivial. The ASTRO-FOLD results therefore indicate that accurate ab initio prediction of designed proteins may be quite successful, because the modeling methodology employed in ab initio (all-atom physics-based) methods are somehow complementary to the concepts employed in the rational design approach. Furthermore, detailed analyses indicate that the ability to accurately predict hydrophobic contacts between helices may be helpful in further improving the accuracy of the ASTRO-FOLD approach for structure prediction of helical proteins, and research in this area is currently being pursued.

REFERENCES

1. Anfinsen CB. Principles that govern the folding of protein chains. Science 1973;181:223–230.
2. Moult J, Fidelis K, Zemla A, Hubbard T. Critical assessment of methods of protein structure prediction (CASP)—round 4. Proteins 2001;Suppl 5:2–7.
3. Klepeis JL, Floudas CA. Ab initio tertiary structure prediction of proteins. J Global Optim 2003;25:113–140.
4. Klepeis JL, Floudas CA. ASTRO-FOLD: a combinatorial and global optimization framework for ab initio prediction of three-dimensional structures of proteins from the amino-acid sequence. Biophysical J 2003;85:2119–2143.
5. Liwo A, Lee J, Ripoll D, Pillardy J, Scheraga H. Surmounting the multiple minima problem in protein folding. Proc Natl Acad Sci USA 1999;96:5482–5485.
6. Pillardy J, Czaplewski C, Liwo A, Wedemeyer W, Lee J, Ripoll D, Arlukowicz P, Oldziej S, Arnautova E, Scheraga H. Development of physics-based energy functions that predict medium resolution structure for proteins of α, β and α/β structural classes. J Phys Chem B 2001;105:7299—7311.
7. Klepeis JL, Floudas CA. Prediction of beta-sheet topology and disulfide bridges in polypeptides. J Comp Chem 2003;24:191–208.
8. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D. Design of a novel globular protein fold with atomic-level accuracy. Science 2003;302:1364–1368.
9. Klepeis JL, Floudas CA, Morikis D, Tsokos CG, Argyropoulos E, Spruce L, Lambris JD. Integrated computational and experimental approach for lead optimization and design of compstatin variants with improved activity. J Am Chem Soc 2003;125:8422–8423.
10. Wei Y, Kim S, Fela D, Baum J, Hecht M. Solution structure of a protein from a designed combinatorial. Proc Natl Acad Sci USA 2003;100(23):13270–13273.
11. Hill RB, Raleigh DP, Lombardi A, DeGrado WF. De novo design of helical bundles as models for understanding protein folding and function. Acc Chem Res 2000;33:745–754.
12. Hecht MH, Richardson DS, Richardson DC, Ogden RC. De novo design, expression and chracterisation of felix: a four helix bundle protein with native-like sequence. Science 1990;249:884–891.
13. Bowie JU, Reidhaar-Olson JF, Lim WA, Sauer RT. Deciphering the message in protein sequences: tolerance to amino acid substitutions. Science 1990;247:1306–1310.
14. Moore JC, Arnold FH. Directed evolution of a para-nitrobenzyl esterase for aqueous-organic solvents. Nat Biotechnol 1996;14:458–467.
15. Dahiyat BI, Gordon DB, Mayo SL. Automated design of the surface positions of protein helices. Protein Sci 1997;6:1333–1337.
16. Dahiyat BI, Mayo SL. De novo protein design: fully automated sequence selection. Science 1997;278:82–87.
17. Skalicky JJ, Gibney BR, Rabanal F, Urbauer RJB, Dutton PL, Wand AJ. Solution structure of a designed four-r-helix bundle maquette scaffold. J Am Chem Soc 1999;121:4941–4951.
18. Baltzer L, Broo KS. De novo designed polypeptide catalysts with adopted folded structures. Biopolymers 1998;1:31–40.
19. Raha K, Wollacott AM, Italia MJ, Desjarlais JR. Prediction of amino acid sequence from structure. Protein Sci 2000;9:1106–1119.
20. Street AG, Mayo SL. Pairwise calculation of protein solvent-accessible surface areas. Fold Des 1998;3:253–258.
21. Nohaile MJ, Hendsch ZS, Tidor B, Sauer RT. Altering dimerization specificity by changes in surface electrostatics. Proc Natl Acad Sci USA 2001;98:3109–3114.
22. Wei Y, Hecht MH. Enzyme-like proteins from an unselected library of de novo sequences. Protein Engineering, Design and Selection (PEDS) 2004;17:65–75.
23. Wei Y, Liu T, Sazinsky SL, Moffet DA, Pelczer I, Hecht MH. Stably folded de novo proteins from a designed combinatorial library. Protein Sci 2003;12:92–102.
24. Kamtekar S, Schiffer JM, Xiong HY, Babik JM, Hecht MH. Protein design by binary patterning of polar and nonpolar amino-acids. Science 1993;262:1680–1685.
25. West MW, Wang WX, Patterson J, Mancias JD, Beasley JR, Hecht MH. De novo amyloid proteins from designed combinatorial libraries. Proc Natl Acad Sci USA 1999;96:11211–11216.
26. Wang W, Hecht MH. Rationally designed mutations convert de novo amyloid-like fibrils into soluble monomeric β-sheet proteins. Proc Natl Acad Sci USA 2002;99:2760–2765.
27. Dill KA. Polymer principles and protein folding. Prot Sci 1999;8:1166–1180.
28. Baldwin RL, Rose GD. Is protein folding hierarchic?: I. Local structure and peptide folding. TIBS 1999;24:26–33.
29. Baldwin RL, Rose GD. Is protein folding hierarchic?: II. Folding intermediates and transition states. TIBS 1999;24:77–83.
30. Minor D, Kim P. Content dependent secondary structure formation of a designed peptide sequence. Nature 1996;380:730–734.
31. Némethy G, Gibson KD, Palmer KA, Yoon CN, Paterlini G, Zagari A, Rumsey S, Scheraga HA. Energy parameters in polypeptides: 10. J Phys Chem 1992;96:6472–6484.
32. Klepeis JL, Floudas CA. Free energy calculations for peptides via deterministic global optimization. J Chem Phys 1999;110:7491–7512.
33. Lee J, Scheraga H, Rackovsky S. Conformational analysis of the 20-residue membrane bound portion of mellitin by conformational space annealing. Biopolymers 1998;46:103–115.
34. Gilson M, Sharp K, Honig B. Calculating electrostatic interactions in biomolecules: method and error assessment. J Comp Chem 1987;9:327–335.
35. Floudas CA. Nonlinear and mixed-integer optimization: fundamentals and applications. New York and Oxford: Oxford University Press; 1995.
36. Klepeis JL, Floudas CA, Morikis D, Lambris JD. Predicting peptide structures using NMR data and deterministic global optimization. J Comput Chem 1999;20:1354–1370.
37. Androulakis IP, Maranas CD, Floudas CA. αBB: a global optimization method for general constrained nonconvex problems. J Global Optim 1995;7:337–363.
38. Floudas CA. Deterministic global optimization: theory, methods and applications. Kluwer Academic; 2000.
39. Güntert P, Mumenthaler C, Wüthrich K. Torsion angle dynamics for NMR structure calculation with the new program dyana. J Mol Biol 1997;273:283–298.
40. Guex N, Peitsch MC. Swiss-model and the Swiss-pdbviewer: an environment for comparative protein modeling. Electrophoresis 1997;18:2714–2723.
41. Jones DT. Protein secondary structure prediction based on position specific scoring matrices. J Mol Biol 1999;292:195–202.
42. McGuffin LJ, Bryson K, Jones DT. The psipred protein structure prediction server. Bioinformatics 2000;16:404–405.
43. Karplus K, Karchin R, Draper J, Casper J, Mandel-Gutfreund Y, Diekhans M, Hughey R. Combining local-structure, fold-recognition and new-fold methods for protein structure prediction. Proteins 2003;53:491–496.

44. Simons K, Kooperberg C, Huang C, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 1997;268:209–225.
45. Rost B, Liu JF. The predictionprotein server. Nucleic Acids Res 2003;31:3300–3304.
46. McGuffin LJ, Jones DT. Targeting novel folds for structural genomics. Proteins 2002;48:44–52.
47. Rychlewski L, Fischer D, Elofsson A. Livebench-6: large scale automated evaluation of protein structure prediction servers. Proteins 2003;53:542–547.
48. Lundstrom J, Rychlewski L, Bujnicki J, Elofsson A. Pcons: a neural network based consensus predictor that improves fold recognition. Protein Sci 2001;10:2354–2362.
49. Fischer D. 3ds3 and 3ds5 3d-shotgun meta predictors in cafasp3. Proteins 2003;53:517–523.
50. Ginalski K, Elofsson A, Fischer D, Rychlewski L. 3d-jury: a simple approach to improve protein structure predictions. Bioinformatics 2003;19:1015–1018.
51. Koehl P, Levitt M. De novo protein design: I. In search of stability and specificity. J Mol Biol 1999;293:1161–1181.
52. Skolnick J, Kolinski A, Ortiz A. Monsster: a method for folding globular proteins with a small number of distance restraints. J Mol Biol 1997;265:217–241.
53. Skolnick J, Kolinski A, Kihara D, Betancourt M, Rotkiewicz P, Boniecki M. Ab initio protein structure prediction via a combination of threading lattice folding, clustering and structure refinement. Proteins 2001;Suppl 5:149–156.
54. Eyrich V, Standley DM, Felts AK, Friesner RA. Protein tertiary structure prediction using a branch and bound algorithm. Proteins 1999;35:41–57.
55. Standley DM, Eyrich VA, Felts AK, Friesner RA, McDermott AE. A branch and bound algorithm for protein structure refinement from sparse NMR data sets. J Mol Biol 1999;285:1691–1710.
56. Nanias M, Chinchio M, Pillardy J, Ripoll DR, Scheraga HA. Packing helices in proteins by global optimization of a potential energy function. Proc Natl Acad Sci USA 2003;100:1706–1710.
57. van Gunsteren WF, Berendsen HJC. GROMOS. Groningen, the Netherlands: Groningen Molecular Simulation; 1987.
58. Lazaridis T, Karplus M. Effective energy function for proteins in solution. Proteins 1999;35:133–152.
59. Brooks B, Bruccoleri R, Olafson B, States D, Swaminathan S, Karplus M. CHARMM: a program for macromolecular energy minimization and dynamics calculations. J Comp Chem 1983;4: 187–217.
60. 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.