AB-INITIO PREDICTION OF THE 3-DIMENSIONAL STRUCTURE OF A DE-NOVO DESIGNED PROTEIN: A DOUBLE BLIND CASE STUDY

John L. Klepeis, Christodoulos A. Floudas, Yinan Wei and Michael H. Hecht
Princeton University, Princeton, NJ 08544

Abstract

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 three dimensional 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 the final NMR analysis both characterize this protein as a four helix bundle. The similarity of these structures is evidenced by both small RMS (root mean squared) deviation values between the coordinates of the two structures and a detailed analysis of helix-packing.

Keywords

Protein structure prediction, Protein design, Global optimization, Combinatorial optimization.

Introduction

Recently, a novel four stage ab initio approach, ASTRO-FOLD, was introduced for the structure prediction of single chain polypeptides (Klepeis and Floudas, 2003). The methodology combines the classical and new views of protein folding, while using free energy calculations and integer linear optimization to predict helical and beta-sheet structures, respectively. Detailed atomistic modeling and the deterministic global optimization method, αBB, 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.

A related problem to that of protein structure prediction is the design of de novo proteins. Computational de novo design 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 (Dahiyat and Mayo, 1997; Klepeis et al., 2003). 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 (Hecht et al., 1990; Wei et al., 2003).

A method to enhance the success of combinatorial libraries has been described (Wei et al., 2003; West et al., 1999). 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. Recently, a second generation library of sequences was described, and the first high resolution structure of a protein from this library was determined (Wei et al., 2003). 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 the sequel, a double blind study of the ab initio prediction of a de novo designed protein is presented in which the ASTRO-FOLD ab initio approach was used to predict the three dimensional protein structure. The protein, S-824, was ultimately found to be a 4-helix bundle that conformed to the design goals of the combinatorial library. Agreement between the experimental and

predicted structure for protein S-824 is impressive, and features not expressly incorporated into the design were correctly predicted.

Theory and Modeling

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 posses 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 of helix propensities for each residue using equilibrium occupational probabilities of helical clusters (Klepeis and Floudas, 2003). 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. This is consistent with the observation that local interactions extending beyond the boundaries of the helical segment retain information regarding conformational preferences (Baldwin and Rose, 1999). 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 which still capture the hydrogen bonding pattern.

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 (Floudas, 2000). Two types of superstructure are introduced, both of which emanate from the principle that hydrophobic interactions drive the formation of β-structure. The resulting mathematical models are Integer Linear Programming (ILP) problems which not only can be solved to global optimality, but can also provide a rank ordered list of alternate β-sheet configurations. For the current system, no β-sheet structure is present.

The third stage of the approach serves as a preparative phase for the atomistic-level tertiary structure prediction, and therefore focuses on the determination of pertinent information from the results of the previous two stages. This involves the introduction of lower and upper bounds on dihedral angles of residues belonging to predicted helices or β-strands, as well as possible restraints between the C-α atom distances.

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

$$\begin{aligned} \min_{\phi} \quad & E_{\text{forcefield}} \\ \text{s.t.} \quad & E^{l}_{\text{distance}} \le E^{l}_{\text{ref}} \qquad l = 1,\ldots, N_{CON} \\ & \phi_i^L \le \phi_i \le \phi_i^U \qquad i = 1,\ldots, N_{\phi} \end{aligned} \tag{1}$$

Here N_ϕ 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 three dimensional conformation of the protein. The total violations of the N_{CON} distance constraints are controlled by the parameters E^l_{ref} (Klepeis et al., 1999). 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 (Floudas, 2000).

Experimental De Novo Design

The generation of de novo proteins from designed libraries incorporates both combinatorial and rational design concepts (Hecht et al., 2003; Wei et al., 2003; Kamtekar et al., 1999). A fusion of these approaches is achieved by employing a binary patterning of polar and nonpolar residues, such that all sequences in the library are consistent with the formation of specified 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 helix.

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, five nonpolar amino acids (Met, Leu, Ile, Val and Phe) can be encoded by the degenerate codon NTN, while the degenerate codon VAN provides six 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 (Wei et al., 2003; Kamtekar et al., 1993). 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 four-helix bundle protein (Hecht et al., 2003).

Results and Discussion

As described previously, the overall structure of protein S-824, which was determined from multi-dimensional NMR experiments, is an up-down-up-down 4-helix bundle (Hecht et al., 2003). Among the ensemble of low energy NMR structures the RMS deviation with the mean structure for all heavy atoms is 1.06 +/- 0.10 angstroms. Deviations within α-helices and among nonpolar amino acids in the hydrophobic core are even smaller, indicating that S-824 forms a well defined structure. It is important to note that protein S-824 was chosen from an unselected library with the potential for enormous combinatorial diversity. Of the 102 positions in the sequence of S-824, only 14 were designed as specific amino acids. The other 88 positions were derived from degenerate DNA codons. Thus, these results indicate that a rich sequence space, in combination with the rational design of binary patterns for four helices, can be used to discover 4-helix bundles with well folded structures.

Prior to any knowledge of these results, the ASTRO-FOLD approach was applied to the three dimensional structure prediction of the S-824 sequence. The first stage of this ab initio approach involves the prediction of whether or not helical segments exist and their initiation and termination sites. For each of the 98 pentapeptides, detailed free energy calculations were performed, and the ensembles of low energy conformers were used to calculate helix propensities for each residue. The final assessment is made according to average probabilities and 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 comprised of four helical segments.

Because four 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 are 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 angles). Distance restraints included 58 lower and upper C-α-C-α (5.5-6.5 angstroms) bounds to enforce the hydrogen bonding network within these 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 identified. 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 and the experimental structure, as shown in Figure 1. The final locations of the four 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 angles or hydrogen bonding may provide different results. A number of methods were used, and the concensus is essentially identical, with only slight differences in the initiation/termination of the helices.

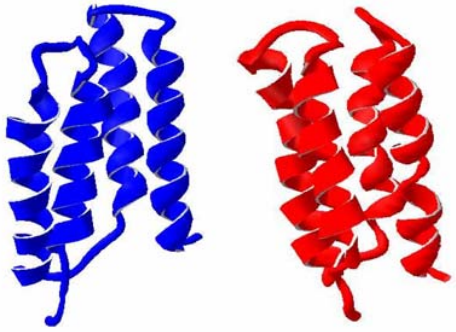

Figure 1. Comparison of predicted (left) and experimental (right) structures for S-824.

Examination of the overall three dimensional structure reveals other features that are consistent with the experimental structure. These results are especially interesting given that these features were not, and could not have been, part of the design goals. 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. 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 is approximately 20 degrees. This combination of packing angles has also been observed in other 4-helix bundle proteins.

Root mean squared (RMS) deviations 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 RMS deviation over all 102 residues between the lowest energy predicted and experimental structures is 4.94 angstroms. In general, backbone RMS deviations below 6 angstroms constitute very good predictions, especially for protein systems with more than 100 residues (Koehl and Levitt, 1999).

Deviations in individual components of the overall structure can also be computed. The results of this analysis reveal several important facts. First, all helical segments exhibit good RMSD values, with no segment above 2.5 angstroms for backbone deviations. In fact, Helix 1 (H1) and Helix 4 (H4) have small deviations (1.14 and 0.84 angstroms, respectively). Conversely, the loop segments, which are much shorter than the helical segments, exhibit backbone RMSD values between 2 - 3 angstroms. These observations emphasize the fact that loop prediction remains a bottleneck in the accurate prediction of protein structure. Note however that because very few experimental restraints were available for the NMR structure refinement, the structures of the loops in the experimental structure are known only approximately.

A quantitative analysis of the packing of these helices is shown in Table 1. 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 (Helices 1 to 2, 2 to 3, 3 to 4 and 1 to 4) 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 helix 1 to helix 4 and helix 2 to helix 3, especially at small distances. When considering C-β the distribution of helical contacts is almost uniform between all antiparallel helix matches in the experimental structure. On the other hand, in the predicted structure the antiparallel helical contacts between helices 1 and 2 and helices 3 and 4 each account for one-third of the hydrophobic-to-hydrophobic interactions for both types of distances. The contacts between helix 1 and helix 4 are also well represented, however the last antiparallel match, between helix 2 and helix 3, 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 correct predict such matches may enhance the performance of the ASTRO-FOLD approach.

Table 1. Percentage of hydrophobic C-α (or C-β) contacts under a 10 angstrom cutoff distance.

System	H1-2	H1-3	H1-4	H2-3	H2-4	H3-4
Pred (α)	37	5	11	0	0	47
Exp (α)	13	10	27	30	4	16

Conclusions

Two important problems in the field of protein science are the structure prediction and de novo protein design. Prior to knowledge of the actual experimental structure of S-824, a protein designed via a de novo approach, a double blind study was conducted in which the ASTRO-FOLD ab initio approach was used to predict the three dimensional protein 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. These results indicate that accurate ab initio prediction of designed proteins may be quite successful because the modeling methodology employed in ab initio methods are complemented by the concepts employed in the rational design approach.

References

Baldwin, R.L. and G.D. Rose (1999). Is protein folding hierarchic? *TIBS*, **24**, 26-33.

Dahiyat, B.I. and S.L. Mayo (1997). De novo protein design: fully automated sequence selection *Science*, **278**, 82-7.

Floudas, C.A. (2000). *Deterministic Global Optimizatoin: Theory, Methods and Applications,* Kluwer Academic Publishers.

Hecht, M.H., D.S. Richardson, D.C. Richardson and R.C. Ogden (1990). De novo design, expression and characterization of felix: a four helix bundle protein with native like sequence. *Science*, **249**, 884-891.

Hill, R.B., D.P. Raleigh, A. Lombardi and W.F. DeGrado (2000). De novo design of helical bundles as models for understanding protein folding and function. *Acc. Chem. Res.*, **33**, 745-754.

Kamtekar, S., J.M. Schiffer, H.Y. Xiong, J.M. Babik and M.H. Hecht (1993). Protein design by binary patterning of polar and nonpolar amino acids. *Science*, **262**, 1680-5.

Klepeis, J.L. C.A. Floudas, D. Morikis and J.D. Lambris (1999). Predicting peptide structures using nmr data and deterministic global optimization. *J. Comput. Chem.*, **20**, 1354-1370.

Klepeis, J.L. and C.A. Floudas (2003). ASTRO-FOLD: A combinatorial and global optimization framework for ab initio prediction of 3-dimensional structures of proteins from the amino-acid sequence. *Biophysical J..*, **85**, 2119-2146.

Klepeis, J.L. C.A. Floudas, D. Morikis, C.G. Tsokos, E. Argyropoulos, L. Spruce and J.D. Lambris (2003). Integrated computational and experimental approach for lead optimization and design of compstatin variants with improved activity. *JACS.*, **125**, 8422-23.

Koehl P., and M. Levitt (1999). De novo protein design. I. In search of stability and specifity. *J. Mol. Biol.*, **293**, 1161-1181.

Wei, Y., S. Kim, D. Fela, J. Baum and M. Hecht. (2003). Solution structure of a protein from a combinatorial library of de novo sequences., *Proc. Natl. Acad. Sci.*, **100**, 231-239.

Wei, Y., T. Liu, S.L. Sazinsky, D.A. Moffet, I. Pelezer and M.H. Hecht. (2003). Stably folded de novo proteins from a designed combinatorial library. *Prot. Sci.*, **12**, 92-102.

West, M.W., W.X. Wang, J. Patterson, J.D. Mancias, J.R. Beasley and M.H. Hecht. (1999). De novo amyloid proteins from designed combinatorial libraries. *Proc. Natl. Acad. Sci.*, **96**, 11211-11216.