Protein Structure Prediction Using Physical-Based Global Optimization and Knowledge-Guided Fragment Packing

Jinhui Ding
QB3 Institute,
UC Berkeley
jhding@lbl.gov

Elizabeth Eskow
Dept. of Computer Science,
Univ. of Colo., Boulder
eskow@cs.colorado.edu

Nelson Max
Dept. of Computer Science,
UC Davis
max2@llnl.gov

Silvia Crivelli*
QB3 Institute,
UC Berkeley
SNCrivelli@lbl.gov
**Corresponding author*

Abstract

We describe a new method to predict the tertiary structure of new-fold proteins. Our two-phase approach combines the knowledge-based fragment-packing with the minimization of a physics-based energy function. The method is one of the few attempts to use an all-atom physics-based energy function throughout all stages of the optimization. Information from the known proteins is utilized to guide the search through the vast conformational space. We tested this method in CASP6 and it produced the best prediction on one of the new-fold targets – T238, alpha-helical protein. After CASP6, we carried out a series of experiments to test and improve our method and we found that our method performed well on alpha-helical proteins.

1. Introduction

Significant progress has been made in developing methods that have the ability to predict the tertiary structure of proteins that are considered to be "new fold" (i.e., proteins whose tertiary structure has little similarity to any known structure) from the primary sequence alone. Fragment-based methods have been widely applied to the prediction of proteins with new folds [1,2]. The search strategies used by the fragment-based methods vary, but most of the fragment-based methods use statistical-based potential (or scoring) functions [2]. Very few prediction methods utilize physics-based potentials [3]. The limitations of the template-based approaches suggest that fragment-based methods should be combined with *ab initio* methods to improve their performance [4].

We describe a new two-phase method to predict the tertiary structure of new-fold proteins via minimization of a physics-based energy function combined with knowledge-based fragment packing. Our approach is one of the few attempts to use an all-atom physics-based energy function throughout all stages of the optimization [3]. It uses information from known proteins to guide the search through the vast conformational space. There are three crucial components in this novel approach: (1) a technique for packing structural fragments which uses templates from fold-recognition servers in a unique way, (2) a sophisticated global optimization algorithm used to minimize a full-atom potential and (3) an in-house graphical environment created specifically for both manual and automatic manipulation of protein structures.

2. Methodology

Our method has two phases. Phase I is the knowledge-based fragment packing phase, and phase II is the physical-based global optimization phase.

Phase I creates a variety of initial configurations by incorporating knowledge from known proteins in two ways: (1) by using secondary structure predictions and structural templates of known proteins [5], and (2) by using probabilistic results of both protein-fold topology [6] and sequence matching specificity [7]. First, it creates an initial, extended configuration that has alpha-helices and beta strands according to the secondary structure predictions. This extended configuration is split into fragments, each containing a single alpha-helix or beta-strand, and two coils at both ends. The size of the structural fragments is not fixed. The fragments are manually packed using our in-house graphical environment, *ProteinShop* [8], according to templates obtained (if any) from the fold recognition meta-servers using the initial sequence of amino acids as a query. In addition to those manually constructed initial configurations, we also utilize *ProteinShop* to automatically produce a collection of high probability sheet conformations guided by the statistical scoring functions derived from both protein-fold topology and sequence matching specificity. All the starting configurations are local minima before going to the Phase II.

Phase II improves the initial configurations by applying a sophisticated optimization algorithm that optimizes selected subspaces of the predicted coil regions in parallel. The method selects a number of low-energy configurations from the list of initial structures and then selects small subsets of variables for improvement by global minimizations. A subset of variables consists of a number of consecutive dihedral angles picked from the set of amino acids predicted to be coil by the secondary structure predictions. Once the subset is determined, a stochastic global optimization procedure is executed to find the best new positions for the chosen dihedral angles while holding the remaining dihedral angles fixed. A number of those configurations with the lowest energy values are selected for local minimizations in the full-dimensional space. These full-dimensional local minimizations are less likely to produce major structural changes but can cause important, more local refinements throughout the protein structure. The new full-dimensional local minimizers are then merged with those found previously, and the entire process repeats iteratively until the lowest energy configuration does not change substantially after a number of iteration steps.

3. Experiments and Discussion

In CASP6, our method produced the best prediction on one of the targets in the new-fold category, T238, an alpha-helical protein. However, the method did not perform consistently well due to the problems in our implementation of the energy function. After CASP6, we changed the energy function and carried out a series of experiments to test the performance of our methods on the alpha-helical proteins. We “re-predicted” the alpha-helical targets of CASP5 and CASP6. By “re-predict”, we mean that we used the PDB information back to the date before the native structure of the target was released. For each target, we analyzed the results at each phase to track the performance of the method. We calculate the GDT_TS score [9] to quantitatively evaluate the overall similarity between the model and the native structure. Table1 lists the GDT_TS score of the best model generated in phase I and phase II and also lists the GDT_TS scores of the best CASP prediction on the corresponding target. T248_1 and T248_3 are alpha-helical proteins and classified as hard fold-recognition targets, and included here for comparison purposes.

Table 1. GDT_TS Scores Comparisons

Target ID	**Best GDT_TS score**		
	Phase I models	**Phase II models**	**Best CASP prediction**
T129	19.94	24.41	37.94
T170	48.91	55.80	64.85
T172_2	20.79	25.25	31.68
T238	29.87	30.66	29.28
T248_1	34.18	36.08	68.35
T248_2	42.53	43.10	50.00
T248_3	36.21	43.39	50.00

We noticed that the models constructed in the phase I, for most cases, grasp the partially correct folding information, but that phase II did not bring enough improvements on the models of phase I. We also found that inaccuracy of the secondary structure prediction decreased the performance of our method.

Several aspects of our method need improvement. We are trying a new approach to minimizing the errors introduced from inaccurate secondary structure predictions. We are looking into ways of improving phase II by using alternative energy functions with better discriminative ability. In addition to that, more efficient methods for sampling the dihedral angle space are under development. New approaches for clustering and filtering the models generated from phase II are also under test.

References

[1] K. Ginalski, N.V. Grishin, A. Godzik, and L. Rychlewski, *Nucleic Acids Res*., 2005, 33(6), pp. 1874-91.

[2] J. Moult, K. Fidelis, A. Zemla and T. Hubbard, *Proteins*. 2003, 53 Suppl 6, pp.334-9.

[3] C. Hardin, T.V. Pogorelov and Z. Luthy-Schulten. *Curr Opin Struct Biol.* 2002, 12, pp. 176-181.

[4] B. Contreras-Moreira, I. Ezkurdia, M.L. Tress and A. Valencia, *FEBS Lett.* 2005, 14;579(5), pp.1203-7

[5] K. Ginalski, A. Elofsson, D. Fischer, and L. Rychlewski, *Bioinformatics,* 2003 19(8), pp.1015-8.

[6] I. Ruczinski, C. Kooperberg, R. Bonneau and D. Baker, *Proteins*, 2002, 48, pp. 85-97.

[7] H. Zhu and W. Braun, *Protein Sci*, 1999, 8, pp. 326-342.

[8] S. Crivelli., O. Kreylos, B. Hamann, N. Max and W. Bethel, *J Comput Aided Mol Des*, 2004, 18, pp. 271-285.

[9] A. Zemla, *Nucleic Acids Res*, 2003, 31, pp. 3370-3374