Protein Structure Prediction: Dreams and Reality Postprint

Authors: Wei Yi, Yang Jishuang, Yuan Xiongying, Shao Mingfu, Wang Chao, Bu Dongbo

Date: 2017-03-09T00:00:00+00:00

Abstract

The contributions of bioinformatics can be measured from two perspectives: one is the contribution to biology, namely whether it can assist biologists (or independently) in making new biological discoveries (discovery) through mathematical and physical means; the other is the contribution to computer science, namely that practical problems are the driving force and source of algorithmic research, and whether we can test existing algorithms and develop new algorithms (algorithm) in the process of solving practical problems. The purpose of this paper is to elaborate on the contributions at these two levels, using the FALCON method for protein structure prediction as a case study. In short, from the perspective of biological discovery, the results of FALCON provide quantitative support for the assertion that "the number of protein structural conformations is limited" ; from the algorithmic perspective, FALCON is essentially a novel optimization framework that can significantly reduce the size of the search space, whereas classical Monte Carlo and Local search maintain a relatively large search space throughout. Experimental results demonstrate that this technique of reducing search space size can effectively improve the likelihood of successful search.

Full Text

Preamble

Vol.8 No.1 Information Technology Letters
Protein Structure Prediction: Dreams and Reality
Wei Yi, Yang Jishuang, Yuan Xiongying, Shao Mingfu, Wang Chao, Bu Dongbo

Abstract

The contributions of bioinformatics can be measured at two distinct levels. At the biological level, we ask whether mathematical and computational methods can assist biologists (or independently) in making new biological discoveries. At the computer science level, practical problems serve as the driving force and source of algorithmic research, prompting us to evaluate existing algorithms and develop new ones in the process of solving real-world challenges. This paper uses the FALCON method for protein structure prediction as a case study to illustrate contributions at both levels. Briefly, from a biological discovery perspective, FALCON' s results provide quantitative support for the hypothesis that the number of protein conformational states is limited. From an algorithmic perspective, FALCON represents a novel optimization framework that dramatically reduces the size of the search space, whereas classical Monte Carlo and local search methods maintain a relatively large search space throughout. Experimental results demonstrate that this search space reduction technique effectively improves the probability of successful structure prediction.

Keywords: protein structure prediction; optimization; primal-dual; sampling

Proteins are long amino acid chains linked by peptide bonds that only acquire specific biological functions upon folding into particular three-dimensional shapes. For instance, mad cow disease arises from a structural mutation in a brain protein called the prion protein (PrP), which transforms from its normal water-soluble -helical structure into an insoluble -sheet structure that deposits in brain tissue, causing neurodegeneration and spongiform encephalopathy. Therefore, understanding protein structure is crucial for elucidating protein function.

Experimental methods for determining protein tertiary structure primarily include X-ray crystallography and nuclear magnetic resonance (NMR). However, the speed of these structural determination methods lags far behind DNA sequencing and gene prediction, making them inadequate for proteome-scale structure prediction demands. For example, determining a single protein structure using NMR typically costs $150,000 and requires six months. Consequently, computational prediction is needed to bridge the gap between structure determination and sequence determination speeds. Moreover, advances in prediction methods contribute to understanding protein folding mechanisms and hold significant theoretical value. Furthermore, structure prediction is fundamentally important for novel protein design—predicting structures is essential for efficiently designing proteins with specific architectures. For these three reasons, accurate protein structure prediction from sequence has become an urgent requirement.

Is computational prediction of structure from sequence feasible?

In 1965, Anfinsen proposed the "self-assembly hypothesis" based on experiments showing that reduced and denatured bovine pancreatic RNase1 could refold into

its native structure simply by removing denaturants and reducing agents, without any additional factors. This hypothesis states that an amino acid sequence contains all the information necessary to form its thermodynamically stable native conformation. Subsequent work has supplemented this theory, establishing that amino acid sequences determine spatial conformations and thereby providing the theoretical foundation for protein structure prediction.

To objectively and fairly evaluate the performance of various prediction methods, John Moult and colleagues have organized the Critical Assessment of Techniques for Protein Structure Prediction (CASP) competition series since 1994. Unlike other evaluation methods such as Livebench, CASP employs a blind test approach where target protein structures are either undetermined or, if determined, not yet publicly released at the time of the competition. CASP-7 in 2006 utilized over 100 test cases, providing a relatively fair benchmark dataset for algorithm design and evaluation.

It is worth noting that CASP' s purpose is to stimulate the generation of new ideas rather than simply ranking existing methods or implementations. This represents perhaps the best perspective from which to view such international competitions.

Classic protein structure prediction methods can be divided into three categories: homology modeling, threading, and *ab initio* methods.

Homology modeling infers a target protein's three-dimensional structure through its homologous proteins. The key step involves sequence-sequence similarity comparison to establish evolutionary relationships. For cases with high similarity, homology modeling can predict tertiary structures with high accuracy, but it often fails when sequence similarity is low.

Threading seeks proteins that share the same structural fold type with the target sequence despite lacking significant sequence homology. Its key step involves sequence-structure alignment calculations to obtain the most probable alignment. Compared with homology modeling, threading better utilizes structural information from template libraries, such as amino acid interactions, thereby achieving more precise predictions than homology modeling.

Ab initio methods start from first principles to find the conformation with minimum energy for the target protein. IBM' s supercomputer BlueGene-L was developed specifically for this simulation, yet it can currently only calculate the folding process for a few amino acids. Using a Monte Carlo strategy, Duan and Kollman simulated a one-millisecond folding process of a 36-residue protein on a Cray supercomputer with 256 processors for two months.

After years of effort, structure prediction can now be considered solved for homologous proteins with sequence similarity greater than 30%; threading achieves approximately two-thirds accuracy in fold recognition; while *ab initio* methods require substantial effort and novel ideas to achieve breakthroughs.

In recent years, *ab initio* methods have gained increasing attention because they

offer unique advantages over homology modeling and threading, such as helping to reveal protein folding mechanisms and enabling structure prediction without known homologs. However, these methods also have limitations. Researchers typically describe candidate conformations of local structures using simple enumerated "discrete" approaches rather than characterizing the distribution of "continuous" conformational space. This leads to each candidate being similar yet still substantially different from the true structure, with errors that cannot be eliminated. *Ab initio* methods are often powerless against such local structure discretization issues. Additionally, the enormous search space directly reduces the probability of finding the native structure. These shortcomings have hindered practical applications of *ab initio* methods, motivating the need for new algorithmic frameworks.

As pioneering work, Li Shuaicheng, Bu Dongbo, Xu Jinbo, and Li Ming proposed a novel prediction algorithm called FALCON3 based on Fragment-HMM2, which reduces the protein conformational space from ROSETTA' s (200)^n to O(1.66^n), bringing it closer to Dill' s estimate of O(1.6^n).

The biological rationale for our approach is that protein structure results from the combined effects of short-range and long-range interactions. Local structures are primarily influenced by short-range interactions, while long-range interactions determine the placement of local structures to minimize free energy and produce stable conformations. Therefore, we must address two fundamental questions:

1. How to characterize local structural propensities?
2. How to characterize correlations arising from long-range interactions?

Our FALCON algorithm employs the following techniques:

1. Local Structure Prediction Algorithm

We have implemented the aforementioned algorithm in a software package called FRazor (Fragment Razor) and obtained preliminary experimental results. In our experiments, we randomly extracted 9,338 fragments from the PDB template library, using half for training and half for testing. Initial results show that with a local structure candidate set size of 25, our integer linear programming model achieves prediction accuracies of 98.6% for alpha-helix regions, 89.6% for beta-strand regions, and 78.1% for loop regions—substantial improvements over ROSETTA. When the candidate set size is increased to 40, the accuracies reach 99%, 92.9%, and 82.4%, respectively. These results demonstrate that the integer linear programming model can effectively predict local structures.

2. Dihedral Angle Distribution Characterization and Progressive Refinement

We have implemented an iterative refinement strategy, with preliminary experiments indicating its effectiveness. Figure 1 [Figure 1: see original paper] illus-

trates how the dihedral angle estimates for Residue 41 of protein 2CRO (Cro Repressor) progressively improve through iterations. For this residue, the initial estimates from the local structure prediction step formed two clusters: one in the alpha-helix region and one in the beta-strand region, both substantially different from the true values (=1.44, =-0.63). After one iteration, both clusters weakened while a new concentrated region emerged (centered at =-0.07). After the second iteration, the incorrect beta-strand cluster disappeared completely while the alpha-helix cluster continued to weaken. After the third iteration, the alpha-helix cluster also vanished, and the new cluster gradually strengthened, eventually stabilizing at the center (=-1.82, =-0.13). This center is close to the true values and corresponds to a loop structure.

3. Position-Specific Hidden Markov Model Sampling Algorithm

Unlike existing algorithms such as FB5-HMM, our FALCON design employs a position-specific hidden Markov model (Position-specific HMM), where each position has a different number of hidden nodes and transition probabilities. Experimental results demonstrate that this position-specific HMM effectively reduces the search space.

We have implemented the above model in a preliminary FALCON software prototype. Experimental results show that even without iterative techniques, the algorithm already exhibits advantages over ROSETTA.

Table: ROSETTA vs FALCON Performance (% of structures <6.0Å)

Target Protein	ROSETTA	FALCON
Protein A, 1FC2		
Homeodomain, 1ENH		
Protein G, 2GB1		
Cro repressor, 2CRO		
Protein L7/L12, 1CTF		
Calbidin, 4ICB		

When iterative techniques are employed, FALCON' s results improve dramatically. After five rounds of iteration, the proportion of "good structures" for all six test proteins gradually increases to 100%.

Table: FALCON Performance After 5 Iterations (% of structures <6.0Å)

Target Protein	FALCON
Protein A, 1FC2	100%
Homeodomain, 1ENH	100%
Protein G, 2GB1	100%

Target Protein	FALCON
Cro repressor, 2CRO	100%
Protein L7/L12, 1CTF	100%
Calbidin, 4ICB	100%

In CASP-8, FALCON achieved third place in the Fold Recognition Hard category. Figure 2 [Figure 2: see original paper] shows the predicted structure (left) versus the native structure (right) for protein 1CTF, with an error of 0.557 Å.

Essentially, FALCON transforms the traditional discrete-domain optimization problem into the following continuous-domain optimization problem:

ssss t ss i i

Here, _i and _i are angle variables representing the two dihedral angles of the i-th amino acid in the protein. Determining the dihedral angles for all positions allows precise reconstruction of the overall three-dimensional structure. p_i(_i, _i) denotes the distribution over continuous space, while the objective function E represents the energy of the protein conformation determined by the current dihedral angles. The overall goal of the optimization problem is to solve the above formulation using sampling techniques.

It is worth noting that in classic Monte Carlo or local search methods, the search space remains unchanged, whereas our algorithm can dramatically reduce the search space. Our experience suggests that for optimization problems, problem transformation is a crucial technique for algorithmic improvement—either by changing the search space or by modifying the energy landscape.

Although CASP-8 results demonstrate that FALCON has achieved initial success as a prototype system, many theoretical and practical challenges remain before reaching the ideal goal of "seamlessly integrating homology modeling, threading, and first-principles techniques." We continue to work toward this objective.

Author Biographies

Wei Yi: Master' s student, 2006 cohort, Institute of Computing Technology, Chinese Academy of Sciences
Yang Jishuang: Master' s student, 2006 cohort, Institute of Computing Technology, Chinese Academy of Sciences
Yuan Xiongying: Master's student, 2007 cohort, Institute of Computing Technology, Chinese Academy of Sciences
Shao Mingfu: Master' s student, 2008 cohort, Institute of Computing Technology, Chinese Academy of Sciences
Wang Chao: Master' s-PhD combined program student, 2008 cohort, Institute

of Computing Technology, Chinese Academy of Sciences
Bu Dongbo: Associate Professor, Institute of Computing Technology, Chinese Academy of Sciences, dbu@ict.ac.cn

Note: Figure translations are in progress. See original paper for figures.