Bioinformatics, 32(3), 2016, 462–464
doi: 10.1093/bioinformatics/btv581
Advance Access Publication Date: 10 October 2015
Applications Note

OXFORD

Structural bioinformatics

FALCON@home: a high-throughput protein structure prediction server based on remote homologue recognition

Chao Wang[1,2,†], Haicang Zhang[1,2,†], Wei-Mou Zheng[3], Dong Xu[4], Jianwei Zhu[1], Bing Wang[1], Kang Ning[5], Shiwei Sun[1], Shuai Cheng Li[6,*] and Dongbo Bu[1,*]

[1]Key Lab of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, [2]University of Chinese Academy of Sciences, Beijing, China, [3]Institute of Theoretical Physics, Chinese Academy of Sciences, Beijing, China, [4]Department of Computer Science and Christopher S. Bond Life Sciences Center, University of Missouri, MO 65211, USA, [5]College of Life Science, Huazhong University of Science and Technology, Wuhan, China and [6]Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Anna Tramontano

Received on July 20, 2015; revised on September 15, 2015; accepted on October 4, 2015

Abstract

Summary: The protein structure prediction approaches can be categorized into template-based modeling (including homology modeling and threading) and free modeling. However, the existing threading tools perform poorly on remote homologous proteins. Thus, improving fold recognition for remote homologous proteins remains a challenge. Besides, the proteome-wide structure prediction poses another challenge of increasing prediction throughput.

In this study, we presented FALCON@home as a protein structure prediction server focusing on remote homologue identification. The design of FALCON@home is based on the observation that a structural template, especially for remote homologous proteins, consists of conserved regions interweaved with highly variable regions. The highly variable regions lead to vague alignments in threading approaches. Thus, FALCON@home first extracts conserved regions from each template and then aligns a query protein with conserved regions only rather than the full-length template directly. This helps avoid the vague alignments rooted in highly variable regions, improving remote homologue identification.

We implemented FALCON@home using the Berkeley Open Infrastructure of Network Computing (BOINC) volunteer computing protocol. With computation power donated from over 20 000 volunteer CPUs, FALCON@home shows a throughput as high as processing of over 1000 proteins per day. In the Critical Assessment of protein Structure Prediction (CASP11), the FALCON@home-based prediction was ranked the 12th in the template-based modeling category. As an application, the structures of 880 mouse mitochondria proteins were predicted, which revealed the significant correlation between protein half-lives and protein structural factors.

Availability and implementation: FALCON@home is freely available at http://protein.ict.ac.cn/FALCON/.

Contact: shuaicli@cityu.edu.hk, dbu@ict.ac.cn
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The protein structure prediction problem is concerned with the prediction of three-dimensional structure of a protein from its primary sequence of residues. Based on whether homologous templates exist for the query protein, the prediction approaches can be categorized into template-based modeling (TBM, including homology modeling and threading) and free modeling (FM). Specifically, a threading approach recognizes the most likely fold of the query protein and reports its alignments to known structural templates as output. This strategy has been relatively successful for various proteins; however, the existing threading approaches have performed poorly for remote homologous proteins. Thus, improving the fold recognition for remote homologous proteins remains a challenge to protein structure prediction (Liu *et al.*, 2014). In addition, currently available prediction servers, such as ROSETTA, I-TASSER (Yang *et al.*, 2015) and RaptorX (Peng and Xu, 2011) are far from meeting the needs of high-throughput structure prediction from the community. More servers supporting proteome-wide structure prediction are in high demand.

In general, the sequences of remote homologous proteins show a relatively weak signal of native structures. However, this does not indicate a lack of sequence conservation features for structure. In fact, during the evolutionary process of homologous proteins, some regions are conserved while some regions are relatively variable in term of structure and sequence. The conserved regions are known as common structural frameworks that are shared by homologous proteins. Note that Xu *et al.* (2003) successfully utilized the concept of conserved region in the design of RAPTOR. This observation is also supported by the success of multiple-templates threading strategy (Peng and Xu, 2011). Thus, based on this observation, we have proposed a novel threading approach to improving remote homologue identification and further developed a high-throughput protein structure prediction server called FALCON@home.

2 Implementation

The FALCON@home server consists of two major modules, namely, TBM and *ab initio* modules (Supplementary Fig. S1). The TBM module was designed to recognize the homologous templates for the query protein sequence and to report alignments with identified homologous templates. Specifically, for each template, all of its homologous proteins were first identified based on sequence and structure similarity. Next we employed an integer linear program to calculate the common structural frameworks shared by these homologous proteins (Zhu *et al.*, 2015). Then we aligned the query protein against the common structural frameworks, which helps to avoid vague alignments rooted in the structurally variable segments of templates. Finally, the full-length alignments between query and templates were generated by using the *tree conditional random field (Tree-CRF)* model (Zhang *et al.*, 2015) (see SI Section 2 for details).

If the TBM module failed to identify homologues for the query protein, FALCON@home activates the FALCON *ab initio* module (Li *et al.*, 2008) to generate models from the very beginning. Briefly speaking, the FALCON *ab initio* module employs a statistical model to describe the local structural preference for each residue and uses a position-specific hidden Markov model to describe the dependencies of neighboring residues (see SI Section 3 for details).

To accelerate the computational process, we deployed the TBM modules onto the Berkeley Open Infrastructure of Network Computing (BOINC) volunteer computing platform. In particular, a total of about 50 000 common frameworks were split into 32 clusters. Each BOINC client aligns the query protein against the common frameworks in a certain cluster and returns the alignments to the BOINC server for further selection. Currently, FALCON@home has over 20 000 volunteer CPUs from all around the world and can process over 1200 proteins per day.

3 Results

3.1 CASP11 evaluation

In the CASP11 competition, the FALCON@home-based server, FALCON_TOPO, was ranked 12th over the TBM category and 17th over the FM category. An enhanced version of FALCON@home, FALCON_EnvFold, was ranked 9th over the TBM category (see SI Section 4). The results indicated the advantage of FALCON@home in remote homologue identification. Take the CASP11 target T0678 as an example. The challenge was to determine how to align the three N-terminal strands. Using the pre-calculated common frameworks, FALCON@home successfully identified the most similar template as `4gt6_A` and finally generated a high-quality prediction model with a TM-score of 0.84 (see Supplementary Fig. S4).

3.2 Application in the investigation of protein half-life

Recently, we applied FALCON@home to investigate the relationship between protein half-life and protein tertiary structure. In particular, we used FALCON@home to predict structures for a total of 6033 mouse proteins in 6 days, including 442 proteins expressed in mouse liver mitochondria, 438 proteins expressed in mouse brainstem mitochondria, and 5153 nucleic proteins. Then we calculated the *surface residue ratio* as the number of residues on protein surface over the total number of residues in the protein. Analysis of the 880 mitochondrial proteins revealed a strongly negative correlation between the surface residue ratio and protein half-life (correlation coefficient is around −0.81). Thus, in other words, with the increase in the ratio, the protein is more likely to be degraded quickly.

4 Conclusions and discussion

With the rapid growth of protein sequences, the needs for structure prediction have also increased. FALCON@home provides free and high-throughput service for protein structure prediction. The experimental results in CASP11 evaluation have presented the advantages of using FALCON@home in remote homologue identification. In addition, the application in the investigation of protein half-life also demonstrates that FALCON@home can be used in the proteome-wide prediction.

Funding

This study was funded by National Basic Research Program of China under Grant 2012CB316502, National Nature Science Foundation of China under Grants 11175224, 11121403, 31270834, 61272318, 30870572 and 61303161, and National Institutes of Health of USA under Grants R21/R33-GM078601

and R01-GM100701. This work made use of the Infrastructure provided by European Commission co-funded project CHAIN-REDS (306819).

Conflict of Interest: none declared.

References

Li,S.C. *et al*. (2008) Fragment-hmm: a new approach to protein structure prediction. *Protein Sci.*, **17**, 1925–1934.

Liu,B. *et al*. (2014) Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. *Bioinformatics*, **30**, 472–479.

Peng,J. and Xu,J. (2011) Raptorx: exploiting structure information for protein alignment by statistical inference. *Proteins Struct. Funct. Bioinf.*, **79**, 161–171.

Xu,J. *et al*. (2003) Raptor: optimal protein threading by linear programming. *J. Bioinf. Comput. Biol.*, **1**, 95–117.

Yang,J. *et al*. (2015) The i-tasser suite: protein structure and function prediction. *Nat. Methods*, **12**, 7–8.

Zhang,H. *et al*. (2015) Improving protein threading accuracy via combining local and global potential using treecrf model. *arXiv preprint arXiv:1509.03434*.

Zhu,J. *et al*. (2015) Topo: Improving remote homologue recognition via identifying common protein structure framework. *arXiv preprint arXiv:1507.03197*.