DOMAC: an accurate, hybrid protein domain prediction server

Jianlin Cheng

University of Central Florida

Find similar works at: https://stars.library.ucf.edu/facultybib2000

University of Central Florida Libraries http://library.ucf.edu

This Article is brought to you for free and open access by the Faculty Bibliography at STARS. It has been accepted for inclusion in Faculty Bibliography 2000s by an authorized administrator of STARS. For more information, please contact STARS@ucf.edu.

Recommended Citation
Cheng, Jianlin, "DOMAC: an accurate, hybrid protein domain prediction server" (2007). Faculty Bibliography 2000s. 6950.

https://stars.library.ucf.edu/facultybib2000/6950
DOMAC: an accurate, hybrid protein domain prediction server

Jianlin Cheng*

School of Electrical Engineering and Computer Science, University of Central Florida, Orlando, FL, 32816, USA

Received January 12, 2007; Revised March 28, 2007; Accepted May 1, 2007

ABSTRACT

Protein domain prediction is important for protein structure prediction, structure determination, function annotation, mutagenesis analysis and protein engineering. Here we describe an accurate protein domain prediction server (DOMAC) combining both template-based and ab initio methods. The preliminary version of the server was ranked among the top domain prediction servers in the seventh edition of Critical Assessment of Techniques for Protein Structure Prediction (CASP7), 2006. DOMAC server and datasets are available at: http://www.bioinfotool.org/domac.html.

INTRODUCTION

Protein domains are structural, functional and evolutionary units of proteins. The prediction of domains from sequence information can improve tertiary structure prediction (1), enhance protein function annotation (2), aid structure determination (3) and guide protein engineering (4) and mutagenesis (5).

A number of different methods have been developed to identify domains starting from primary sequences. These methods can be roughly classified into four categories: template-based methods (6–10), ab initio (template-free) methods (11–22), the hybrid approach combining template-based and ab initio methods (23), and meta-domain prediction methods (24).

Here we describe an accurate, hybrid domain prediction server (DOMAC) that integrates homology modeling, domain parsing and ab initio methods together. The preliminary implementation of the server [under the name: FOLDpro (25)] participated in the domain evaluation in the seventh edition of Critical Assessment of Techniques for Protein Structure Prediction (CASP7) (26,27). It was ranked among the top domain prediction servers in CASP7.

IMPLEMENTATION

Our hybrid approach uses the template-based method to predict domains for proteins having homologous template structures in Protein Data Bank (PDB) (28), and the ab initio method based on neural networks (29) to predict domains for de novo proteins. It predicts protein domains in two steps.

First, it uses the PSI-BLAST (30) to search the target sequence against NCBI Non-Redundant sequence database to construct a profile. The profile is used to search a template structure library built from the proteins in PDB to identify templates, similarly as PDB-BLAST approach (31).

Second, if some significant templates are identified (e-value < 0.001), it generates a structure model for the target using Modeller (32) based on the template structures. Multiple significant templates are combined to improve model quality if available. Then it uses an accurate domain parsing tool PDP (33) to parse the model into domains. If the parsed domains do not cover the whole target sequence, DOMAC will assign uncovered regions to adjacent domains.

If no significant homologous template is found, DOMAC will invoke the ab initio domain predictor DOMpro (29) to predict domains. DOMpro uses neural networks in conjunction with sequence profile, predicted secondary structure, and relative solvent accessibility to predict domain boundary. The secondary structure and relative solvent accessibility are predicted by SSpro (34) and ACCpro (35) in the SCRATCH suite (36). DOMpro tries to identify domain boundary positions based on the composition bias of sequence and structural features in domain linker regions.

The preliminary implementation of DOMAC participated in CASP7 and was ranked first among 13 domain prediction servers. Since then, we have significantly speeded up the template identification process without sacrificing accuracy and added a module to update the template library weekly to incorporate the newly released proteins in PDB.

RESULTS

Here we firstly describe the performance of the preliminary implementation of DOMAC in CASP7 (under server name: FOLDpro). We compare it with 12 other server
Table 1. The performance of 13 domain prediction servers in CASP7.

| Method                | Target Num | Domain Num | Domain Acc. (%) | CASP7 Acc. (%) |
|-----------------------|------------|------------|-----------------|---------------|
| FOLDpro (DOMAC)       | 95         | 93.7       | 0.963           |
| Baker-RosettaDom (23) | 94         | 86.2       | 0.940           |
| Ma-OPUS-DOM           | 94         | 87.2       | 0.933           |
| ROBETTA-GINZU (23)    | 94         | 84.0       | 0.932           |
| DomSSEA (7)           | 94         | 78.7       | 0.910           |
| HHpred3 (38)          | 95         | 75.8       | 0.910           |
| Meta-DP (24)          | 95         | 74.7       | 0.907           |
| HHpred1 (38)          | 93         | 75.3       | 0.902           |
| DomFOLD               | 95         | 75.8       | 0.898           |
| DPS(13)               | 93         | 75.3       | 0.889           |
| Chop (22)             | 83         | 56.6       | 0.827           |
| Distill (39)          | 95         | 70.5       | 0.819           |
| NN_PUT-Lab            | 92         | 58.7       | 0.795           |

The second column (target num) lists the number of targets for which a predictor made predictions.

Predictors in CASP7 using two evaluation metrics: CASP evaluation metric (37) and domain number accuracy.

CASP metric (NDO: normalized domain overlap score) is to compute the overlapping score of domains without explicitly checking domain number and domain boundary (37). It computes the numbers of correctly and wrongly overlapped residues between true domains and predicted domains, respectively. It summarizes the numbers of the overlapping residues into a single score to evaluate domain prediction. The best score for a target is 1 and the worst score is 0. The domain number accuracy is defined as the percentage of targets with correct domain number predictions.

Table 1 reports the performance of 13 servers on 95 targets in CASP 7. The CASP score is the average domain overlap score across all predicted targets. The domain number accuracy is computed by comparing the domain number predictions with the official domain definitions released by CASP7. In terms of the two evaluation metrics, the preliminary implementation of DOMAC (FOLDpro) yielded the best performance.

We also evaluate DOMAC on the three categories of CASP7 targets: highly homologous, homologous and analogous/ab initio. The domain number prediction accuracy of DOMAC is 96%, 94% and 88% in the three categories, respectively.

However, because the majority (68 out of 95) of CASP7 targets is single-domain proteins, the domain prediction accuracy is very likely over-estimated.

Thus, we evaluate DOMAC on a larger, balanced, high-quality dataset manually curated by Holland et al. (2). The publicly released version of the Holland’s benchmarks2 dataset has 156 proteins consisting of 54 single-domain proteins, 69 two-domain proteins, 25 three-domain proteins, 4 four-domain proteins, 3 five-domain proteins and 1 six-domain protein. We evaluate both template-based and ab initio methods on the whole dataset, respectively. Table 2 reports the specificity and sensitivity of each method in each category in terms of domain numbers. The overall domain number prediction accuracy of the template-based and ab initio methods is 75% and 46%, respectively.

Moreover, we assess the accuracy of the domain boundary prediction, which is important for generating hypotheses for crystallizing individual protein domains. Following the same convention (7,22), a predicted boundary within 20 residues away from a true domain boundary is considered correct.

The domain boundary specificity and sensitivity is 50% and 76.5% for the template-based method, and 27% and 14% for the ab initio method. Thus, the accuracy are sufficient for guiding the crystallization experiment, whereas the ab initio method is not always reliable enough for the general, practical use.

Table 2. The specificity and sensitivity of domain number prediction on the Holland’s dataset using the template-based and ab initio methods.

| Method | Acc. (%) | 1-dom | 2-dom | 3-dom | 4-dom | 5-dom | 6-dom |
|--------|----------|-------|-------|-------|-------|-------|-------|
|        |          |       |       |       |       |       |       |
| Template | Sens. | 96.1  | 66.7  | 56.0  | 75.0  | 66.7  |       |
|         | Spec.   | 74.2  | 88.0  | 70.0  | 42.9  | 33.3  |       |
| Ab initio | Sens. | 88.5  | 31.3  | 12.0  |       |       |       |
|         | Spec.   | 46.5  | 48.8  | 30.0  |       |       |       |

CONCLUSION AND FUTURE WORK

We have developed a hybrid domain prediction web service integrating template-based and ab initio methods. The template-based method is accurate enough for guiding protein structure prediction, structure determination, function annotation, mutagenesis analysis and protein engineering. However, the ab initio method still needs to be improved for practical use. Since protein domain architecture is largely shaped by gene recombination events, such as gene fusion, fission, domain swapping and exon exchange, leveraging the evolutionary gene recombination signals embedded in the multiple sequence alignment of a protein family and exon boundaries (or splicing sites) in its gene structure, may help improve ab initio domain prediction significantly.

ACKNOWLEDGEMENTS

J.C. is very grateful to Dr Pierre Baldi for the support during his PhD research at University of California Irvine.
The domain prediction is made by the template-based method using the following protein templates in PDB: 2HISZ, 1JUD, 1ZRN, 2GPHA, 2AHS5, 2HD0A, 1008A, 1IZNA, 1AZ08.

Figure 1. Domain prediction result of CASP7 target T0324. The protein is predicted to have two domains. Domain 1 spans from residues 17 to 81. The continuous segments, spanning from residues 1 to 16 and residues 82 to 208, respectively. Domain 2 spans from residues 1 to 16 and residues 82 to 208.

Funding to pay the Open Access publication charges for this article was provided by the New Faculty Start-Up Grant at the University of Central Florida.

Conflict of interest statement. None declared.

REFERENCES

1. Chivian,D., Kim,D.E., Malmstrom, L., Bradley,P., Robertson,T., Murphy,P., Strauss,C.E., Bonneau,R., Rohli,C.A. et al. (2003) Automated prediction of CASP-5 structures using the Robetta server. Proteins, 53(S6), 524–533.
2. Holland,T., Veretnik,S., Shindyalov,I.N. and Bourne,P.E. (2006) A benchmark for domain assignment from protein 3-dimensional structure and it's applications. J. Mol. Biol., 361, 562–590.
3. Campbell,I. and Downing,A. (1994) Building protein structure and function from modular units. Trends Biotechnol., 12, 168–172.
4. Gueiros,R. and Serrano,L. (2001) Protein design based on folding models. Curr. Opin. Struct. Biol., 11, 101–106.
5. Nielsen,P. and Yamada,Y. (2001) Identification of cell-binding sites in the GEANFAMMER package that finds homologous domains in single- and multi-domain proteins. Bioinformatics, 17, 2814–2824.
6. von Ohsen,N., Sommer,I., Zimmer,R. and Lengauer,T. (2004) Prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. Proteins, 59, 627–632.
7. Sim,J., Kim,S.Y. and Lee,J. (2005) PPRODO: prediction of protein domain boundaries using neural networks. Proteins, 61(Suppl. 7), 193–200.
8. Moult,J., Fidelis,K., Zemla,A. and Hubbard,T. (2003) Critical assessment of methods of protein structure prediction (CASP)-round v. Proteins, 53(Suppl. 6), 334–339.
9. Moult,J., Fidelis,K., Rost,B., Hubbard,T. and Tramontano,A. (2005) Critical assessment of methods of protein structure prediction (CASP)-round 6. Proteins, 61(Suppl. 7), 3–7.
10. Cheng,J. and Baldi,P. (2006) DOMpro: Protein domain prediction using profiles, secondary structure, relative solvent accessibility, and recursive neural networks. Data Mining and Knowledge Discovery, 13, 1–10.
11. Altsehul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,A. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res., 25, 3389–3402.
12. Rychlewski,L., Jaroszewski,L., Li,W. and Godzik,A. (2000) Comparison of sequence profiles. Strategies for structural predictions using sequence information. Protein Sci., 9, 232–241.
13. Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of structural restraints. J. Mol. Biol., 234, 799–815.
14. Alexandrov,N. and Shindyalov,I. (2003) PDP: protein domain parser. Bioinformatics, 19, 429–430.
15. Pollastri,G., Przybylski,D., Rost,B. and Baldi,P. (2002) Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. Proteins, 47, 228–235.
16. Cheng,J., Randall,A.Z., Sveredoski,M.J. and Baldi,P. (2005) SCRATCH: a protein structure and structural feature prediction server. Nucleic Acids Res., 33 (web server issue), w72–w76.
17. Tai,C.H., Lee,, W.J. Vincent,J.J. and Lee,B. (2005) Evaluation of domain prediction in CASP6. Proteins, 61(Suppl. 7), 183–192.
18. Söding,J. (2005) Protein homology detection by HMM-HMM comparison. Bioinformatics, 21, 951–960.