TASK QUARTERLY 18 No 4, 379–384

IMPLEMENTATION AND EVALUATION OF NEW PROTOCOL FOR COMPARATIVE MODELING OF PROTEIN STRUCTURES

MARTA STRUMILLO, ALEKSANDRA E. DAWID, AGATA SZCZASIUK AND DOMINIK GRONT

Laboratory of Theory of Biopolymers, Faculty of Chemistry
University of Warsaw
Pasteura 1, 02-093 Warsaw, Poland

(Paper presented at the CBSB14 Conference, May 25–27, 2014, Gdansk, Poland)

Abstract: Template-based modeling (termed also Comparative or Homology Modeling) of a protein structure is one of ubiquitous tasks of structural bioinformatics. The method can deliver model structures important for testing biological hypotheses, virtual docking and drug design. The performance of these methods is evaluated every two years during a Critical Assessment of Protein Structure Prediction (CASP) experiment.
In this contribution we present a new automated protocol for template-based modeling, which combines computational tools recently developed in our laboratory: the database of protein domain structures (BDDB) with one dimensional and three dimensional threading applications. The protocol was tested during a CASP11 experiment.

Keywords: protein structure prediction, comparative modeling, protein threading, protein alignment, BioShell, Rosetta

1. Introduction

Critical Assessment of protein Structure Prediction – CASP [1] – is a worldwide experiment which provides a comparison of methods for prediction of protein three-dimensional structures. The experiment has taken place every two years since 1994, and it takes the form of a competition in which research groups have limited time for predictions. The double-blind fashion of the experiment ensures that neither the predictors nor the organizers nor the assessors know the proper structure of the targeted sequence. Automated servers are taking part in the experiment's category *server* with the time limit of 72 hours, while research groups in the *human* category have 3 weeks for submitting the prediction. Each category allows submitting up to 5 models for each prediction target. The models should be sorted by a participant according to their presumed accuracy. The

structures which are solved experimentally by crystallographic or spectroscopic groups are deposited in the PDB database not sooner than the expiration of the time limit CASP. Finally, independent assessors provide quantitative evaluation of each of the submitted models and overall ranking of all the participants. This ranking should be considered from a proper perspective. For example, in a case of an 'easy' comparative modeling target, the accuracy of nearly all the predicted models will be within an experimental error of the native structure determination. On the other extreme (template-free modeling), all the submitted models may hardly resemble the correct topology, but still the assessors assign their ranks and point out the winner. However, despite its limitations, the CASP experiment has been widely accepted as an unbiased verification of protein structure prediction methods. During the 2014 edition of CASP, the organizers provided 100 target sequences to be solved by 44 *servers* and 123 *human* predictors.

2. BioShell-Server protocol

This year (summer 2014) our group participated in the CASP11 experiment to evaluate our automated method for protein structure prediction. The protocol is based on one-dimensional threading (*i.e.* profile-profile aligner) [2] and three-dimensional threading [3] applications that have been recently introduced to the BioShell package [4, 5]. The protocol was initially optimized based on targets from previous CASP experiments as well as on the MALIDUP [6] database of homologous domains. The scheme of a BioShell-server protocol for template-based modeling is shown in Figure 1 – the left-hand route. For cases where no homolog could be found, the right-hand side path in the Figure 1 was followed by a *human* predictor. For a template-based modeling case, the protocol performs the following steps:

Psiblast search to collect sequences of homologous proteins
The resulting sequences are used to build a sequence profile for a given target sequence. The PsiBlast [7] results are further utilized by three independent secondary structure prediction methods: PsiPred [8], Porter [9] and SpineX [10].

Template selection by Threading 1D algorithm
At this stage, a sequence profile created from the target sequence is aligned to profiles created for template proteins. The set of templates comprises a non-redundant subset of protein chains deposited to the PDB database combined with a set of representative domains (the BDDB database to be published elsewhere). In some sense the set of templates is redundant, as a protein domain taken from the BDDB database may be also included in one or more chains. Our goal at this stage was to maximize the chance for selecting the correct template. Splitting each template protein into domains improves the sensitivity and results in alignments of higher quality. Whole chains were included in the set of templates to simplify the modeling of multi-domain targets – in cases where the domain composition was the same both in the target and the template. Ten best scoring templates were selected to the next stage.

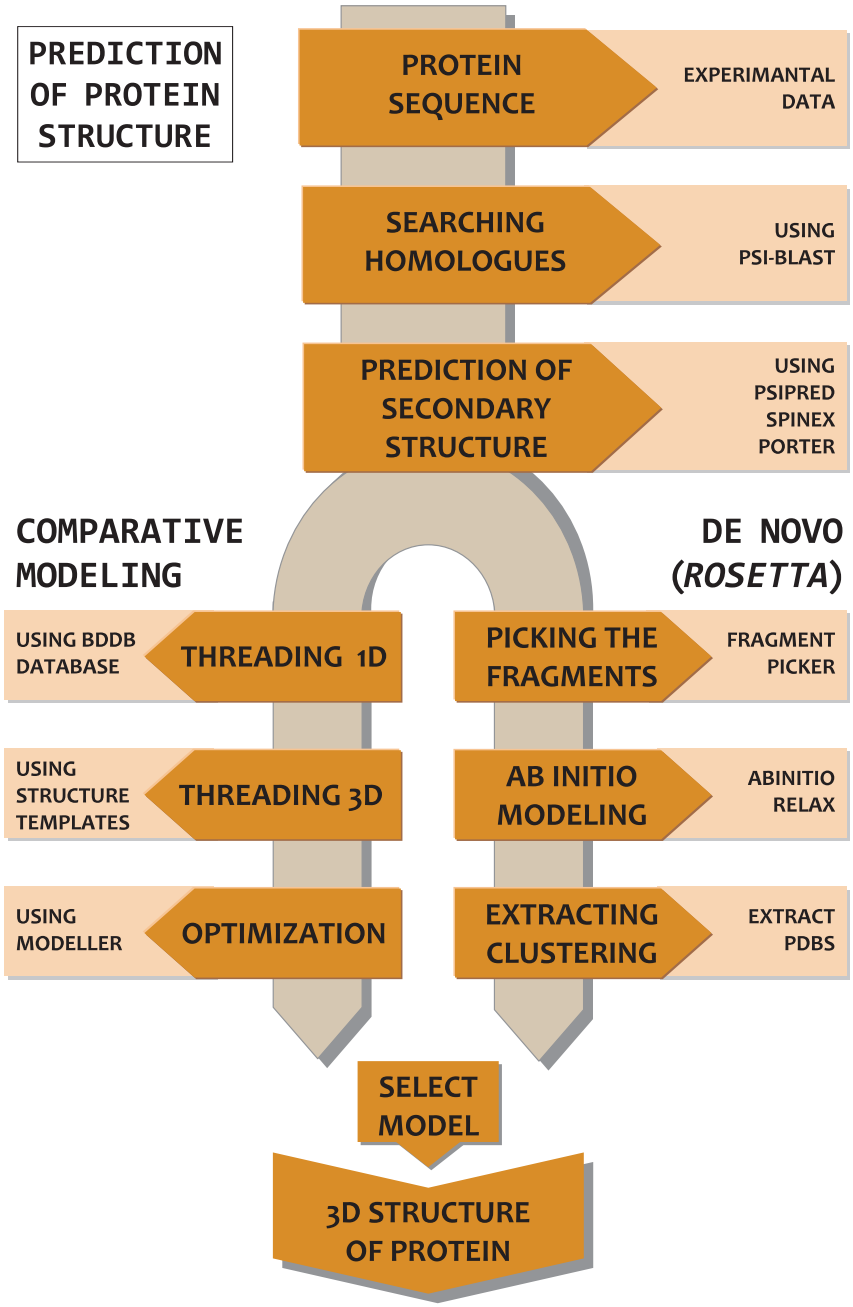

Figure 1. Schematic representation of BioShell modelling protocol

Alignment optimization

The alignment space for each target-template pair was sampled with a Replica Exchange Monte Carlo Scheme [3]. At each Monte Carlo step an alignment modification was proposed and accepted according to the Metropolis criterion.

a custom alignment score comprising several terms (sequence and secondary structure profile match, contact energy and a gap penalty) was used as an energy function. For each of the ten best templates selected by Threading 1D, five best scoring alignments were selected to generate models. Overall from 1000 to 5000 models were built for each target using Modeller [11].

Model scoring

The resulting models were ranked using the DOPE (Discrete Optimized Protein Energy) [12] function. Five best scoring models were submitted to the CASP server.

In the case when no template can be found, the Rosetta [13] *de novo* protocol was employed to predict a target protein structure. Models were continuously computed until the approach of the submission deadline. At the very end, all the models were clustered with `clust` [14] application of the BioShell package. The best models were selected by Rosetta energy, cluster size and visual inspection.

3. Results & Conclusions

In general, the protocol presented in this work is similar to standard approaches in the field (*e.g.* [15, 16]). Following the well established methods, it comprises the four classic steps of comparative modeling: template selection, template-to-target alignment, model building and assessment. Potentially, the most important improvement introduced in this contribution results from the fully three-dimensional approach to the threading problem. Standard approaches in the field do not attempt to optimize the target-to-template alignment *per se*. Any assessment is done after the model structure is calculated. Based on the model (or models) assessment, the alignment is corrected and new models should be calculated. This procedure, repeated several times, often requires human intervention and is difficult to automate. This problem has been already recognized; a standard solution is to generate models based on as many slightly perturbed alignments as possible [17]. In the presented approach, the configurational spaces of structure-to-sequence alignments are sampled with a Monte Carlo algorithm in order to find a population of more accurate alignments.

At the time of preparation of this manuscript, the results of the CASP11 were still unknown. However, preliminary analysis may be done based on the 21 targets whose structures have already been deposited to the PDB database. In Figure 2, a BioShell-server protocol performance was compared to all the models submitted by all the predictors participating in the *server* category. The plot shows TM-score [18] values (defined in the range [0,1] with 1.0 meaning identical structures) on the Y axis for particular targets, marked on the X axis. The range between the best and the worst model submitted for each target by any predictor is marked by a vertical bar while the average TM-score value is denoted by a black square. Pink circles denote the best models found by our method. Overall, the BioShell protocol performed similarly to other approaches. However, the evaluation set (21 models) is too limited to be able to draw major conclusions.

Moreover, the participation in the CASP experiment evaluates the BioShell-server protocol rather as a whole than its particular components. Detailed separate tests for every component of the protocol will be a subject of the forthcoming research.

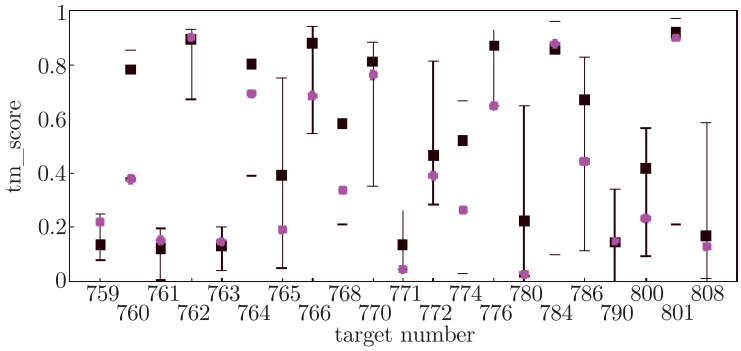

Figure 2. Protocol evaluation on 21 CASP11 targets

Acknowledgements

We would like to acknowledge the support from the Foundation for the Polish Science TEAM project (TEAM/2011-7/6) cofinanced by the European Regional Development Fund operated within the Innovative Economy Operational Program and from the Polish National Science Centre (NCN), Grant No. DEC-2011/01/D/NZ2/07683.

References

[1] Kryshtafovych A, Fidelis K, and Moult J 2009 *Proteins* **77** (S9) 217
[2] Gniewek P, Kolinski A, and Gront D 2012 *J. Comp. Biol.: a journal of computational molecular cell biology* **19** (7) 879
[3] Gniewek P, Kolinski A, Kloczkowski A, and Gront D 2014 *BMC Bioinformatics* **15** (1) 22
[4] Gront D and Kolinski A 2008 *Bioinformatics* **24** (4) 584
[5] Gront D and Kolinski A 2006 *Bioinformatics* **22** (5) 621
[6] Cheng H, Kim B-H H, and Grishin N V 2008 *Proteins* **70** (4) 1162
[7] Altschul S F, Madden T L, Schäffer A A, Zhang J, Zhang Z, Miller W, and Lipman D J 1997 *Nucleic Acids Research* **25** (17) 3389
[8] Jones D T 1999 *Journal of Molecular Biology* **292** (2) 195
[9] Pollastri G and McLysaght A 2005 *Bioinformatics* **21** (8) 1719
[10] Faraggi E, Xue B, and Zhou Y 2009 *Proteins* **74** (4) 847
[11] Šali A and Blundell T L 1993 *J. Mol. Biol.* **234** (3) 779
[12] Shen M-Y Y and Sali A 2006 *Protein Science : a publication of the Protein Society* **15** (11) 2507
[13] Rohl C A, Strauss C E M, Misura K M S and Baker D 2004 *Methods in Enzymology*, Elsevier, **383** 66
[14] Gront D and Kolinski A 2005 *Bioinformatics* **21** (14) 3179
[15] Song Y, DiMaio F, Wang R Y-R Y, Kim D, Miles C, Brunette T, Thompson J and Baker D 2013 *Structure* **21** (10) 1735

[16] Madhusudhan M S, Marti-Renom M A, Eswar N, John B, Pieper U, Karchin R, Shen M Y and Sali A 2005 *Comparative protein structure modeling*, Humana Press Inc, USA 831
[17] Chivian D and Baker D 2006 *Nucleic acids research* **34** (17), e112
[18] Zhang Y and Skolnick J 2004 *Proteins* **57** (4) 702