Protein Structure Prediction Using CABS – A Consensus Approach

Maciej Blaszczyk, Michal Jamroz, Dominik Gront, and Andrzej Kolinski

Laboratory of Theory of Biopolymers, Faculty of Chemistry, University of Warsaw
02-093 Warsaw, Poland
E-mail: mblaszczyk@chem.uw.edu.pl

We have designed a new pipeline for protein structure prediction based on the CABS engine. The procedure is fully automated and generates consensus models from a set of templates. Restraints derived from the templates define a region of conformational space, which is then sampled by Replica Exchange Monte Carlo algorithm implemented in CABS. Results from CASP9 show, that for great majority of targets this approach leads to better models than the mean quality of templates (in respect to GDT_TS). In five cases the obtained models were the best among all predictions submitted to CASP9 as the first models.

1 Introduction

Knowledge of 3D structures of proteins is a crucial requirement for a progress in many areas of biomedicine, e.g. rational drug design. Due to the complexity and high cost of structure determination by experimental methods (mainly Xray crystallography or NMR), computer-based protein structure prediction methods have been placed in the center of attention of a broad community of molecular and cell biologists[1]. Nowadays, there is a number of publicly available web servers, which provide methods for protein structure prediction[2]. Moreover, thanks to the meta-servers[3,4], which collect data from servers, obtaining the predictions is even easier. However, for most purposes it is necessary to provide one, possibly the best, final model. A common approach to this problem is the use of Model Quality Assessment Programs (MQAPs) which score models according to various criteria[5] and selection of the top scoring one. Obviously, the MQAPs can't propose a model better then the best of input structures. Application of CABS modeling tool[6] with spatial restraints derived from the templates allows for reaching beyond this limit.

2 Methods

The procedure used during CASP9 consisted of several steps (Fig. 1) and was trained on the targets from previous CASPs. The first step was templates selection. As templates we used server predictions submitted to CASP9. The list of the servers from which models were taken, was created on the basis of their performance during the CASP8. To check if the best servers from CASP8 are still the reliable ones, servers predictions from CASP9 were ranked using 3D-jury score[7]. Then, for all selected templates distances between pairs of alpha carbons were extracted[8]. The minimum and the maximum distance between pairs of residues were taken as limits of the ranges of restraints. Using templates as a starting structures we have run two independent Replica Exchange Monte Carlo simulations with CABS[6].

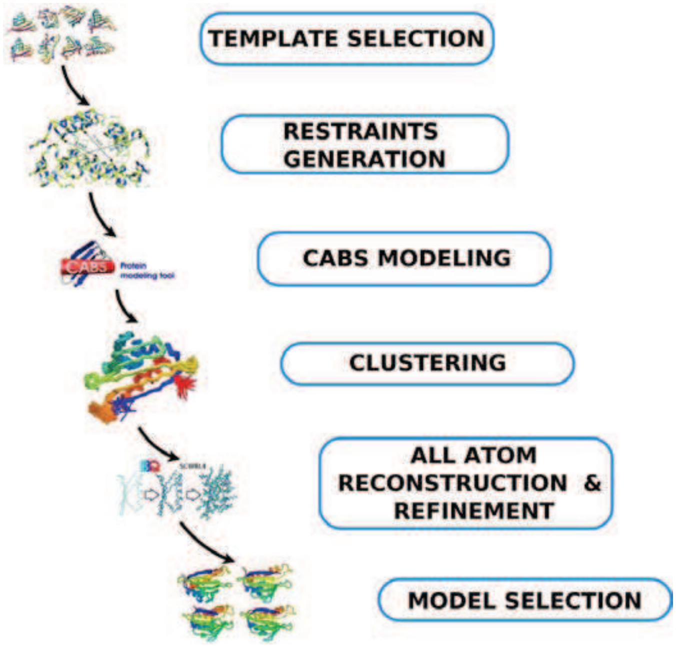

Figure 1. Flowchart of the pipeline used during CASP9. See the text for details.

CABS is a lattice model with a representation reduced to four united atoms per residue: Cα, Cβ, center of mass of a side chain (where applicable) and the center of a virtual Cα – Cα bond. The force field of the model employs knowledge based potentials derived from the statistical analysis of the databases containing known protein structures. Conformational space is sampled using Replica Exchange Monte Carlo method. Application of the restraints reduces conformational space for sampling, which makes modeling faster and more accurate.

The resulted trajectories from CABS were clustered[9], and the clusters' centroids were calculated. Because of reduced representation in CABS, it was necessary to rebuilt the atomistic details of obtained models. Reconstruction of the backbone using BBQ[10] was followed by reconstruction of the side chains with SCWRL4[11]. Next, we performed model refinement, which was also done in two steps. To improve model geometry (e.g. bond length) we employed Modeller[12]. Then, we used GROMACS[13] in order to refine some packing details. Finally, obtained models were ranked on the basis of the clusters' density and the level of similarity of the models from two independent simulations.

3 Results

Since the presented method aims at a consensus prediction from a set of templates it is worth to compare the accuracy of obtained models and the templates used. For great majority of targets GDT_TS of the model was higher then mean GDT_TS of templates. Moreover, in 5 cases the accuracy of the model was better then the accuracy of the best template (see Fig. 2).

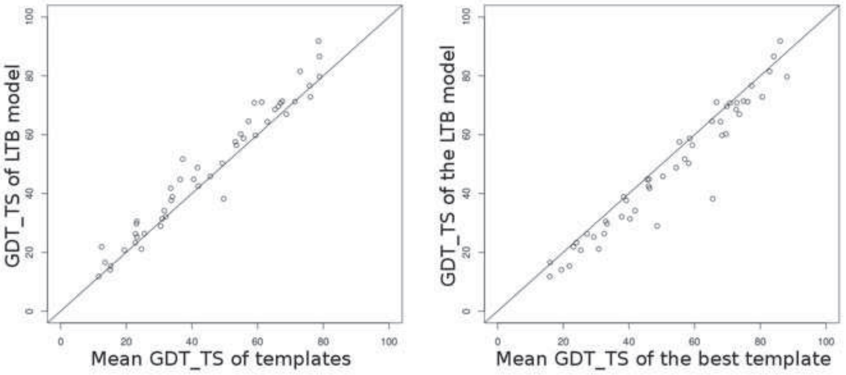

Figure 2. Comparison of GDT_TS of templates and obtained models.

According to the official assessment our models (from Laboratory of Theory of Biopolymers - LTB) for 5 selected domains were the best among all predictions submitted to CASP9 as the first models. As shown in Fig. 3, for great majority of targets, GDT_TS of obtained structure was higher then mean GDT_TS of all models submitted to the CASP. However, there are a few cases with significant losses of accuracy . Most of them are large multi-domain proteins, for which it was necessary to perform domain division, which was not supported in the procedure. This problem is to be solved in a future work.

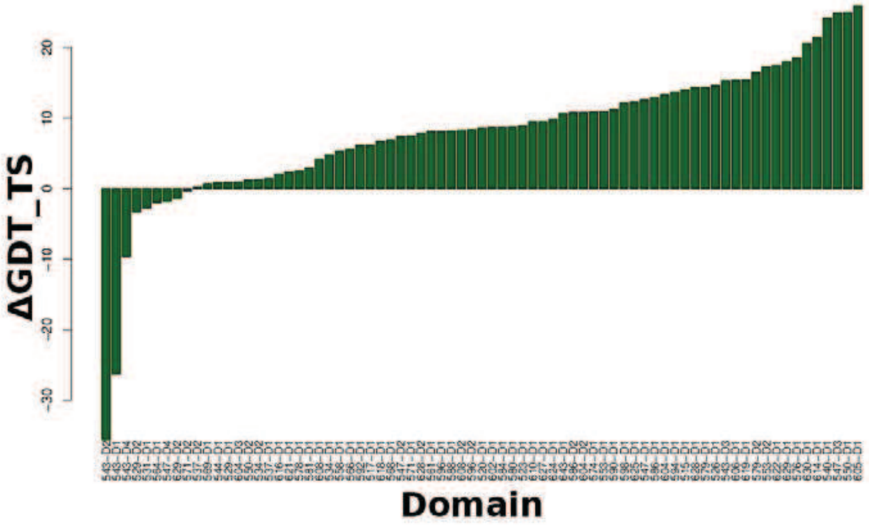

Figure 3. Differences between GDT_TS scores of our models and the mean for all models submitted to CASP.

Acknowledgments

Support from Marie Curie fellowship (FP7-people-IOF) for DG is acknowledged. Computational part of this work was done using the computer cluster at the Computing Center of Faculty of Chemistry, University of Warsaw.

References

1. Yang Zhang, *Protein structure prediction: when is it useful?*, Current opinion in structural biology, **19**, no. 2, 145–155, Apr. 2009.
2. Daniel Fischer, *Servers for protein structure prediction.*, Current opinion in structural biology, Mar. 2006.
3. Krzysztof Ginalski, Arne Elofsson, Daniel Fischer, and Leszek Rychlewski, *3D-Jury: a simple approach to improve protein structure predictions*, Bioinformatics, **19**, no. 8, 1015–1018, May 2003.
4. Jesper Lundström, Leszek Rychlewski, Janusz Bujnicki, and Arne Elofsson, *Pcons: a neural-network-based consensus predictor that improves fold recognition.*, Protein science : a publication of the Protein Society, **10**, no. 11, 2354–2362, Nov. 2001.
5. Andriy Kryshtafovych and Krzysztof Fidelis, *Protein structure prediction and model quality assessment.*, Drug discovery today, **14**, no. 7-8, 386–393, Apr. 2009.
6. Andrzej Kolinski, *Protein modeling and structure prediction with a reduced representation.*, Acta biochimica Polonica, **51**, no. 2, 349–371, 2004.
7. László Kaján and Leszek Rychlewski, *Evaluation of 3D-Jury on CASP7 models*, BMC Bioinformatics, **8**, 304+, Aug. 2007.
8. Dominik Gront and Andrzej Kolinski, *Utility library for structural bioinformatics*, Bioinformatics, **24**, no. 4, 584–585, Feb. 2008.
9. Dominik Gront and Andrzej Kolinski, *HCPM–program for hierarchical clustering of protein models.*, Bioinformatics, **21**, no. 14, 3179–3180, July 2005.
10. Dominik Gront, Sebastian Kmiecik, and Andrzej Kolinski, *Backbone building from quadrilaterals: A fast and accurate algorithm for protein backbone reconstruction from alpha carbon coordinates*, J. Comput. Chem., **28**, no. 9, 1593–1597, July 2007.
11. Adrian A. Canutescu, Andrew A. Shelenkov, and Roland L. Dunbrack, *A graph-theory algorithm for rapid protein side-chain prediction*, Protein Science, **12**, no. 9, 2001–2014, Sept. 2003.
12. Narayanan Eswar, Ben Webb, Marc A. Marti-Renom, M. S. Madhusudhan, David Eramian, Min-yi Shen, Ursula Pieper, and Andrej Sali, *Comparative Protein Structure Modeling Using Modeller*, Current protocols in bioinformatics, **Chapter 5**, Oct. 2002.
13. David Van Der Spoel, Erik Lindahl, Berk Hess, Gerrit Groenhof, Alan E. Mark, and Herman J. Berendsen, *GROMACS: fast, flexible, and free.*, Journal of computational chemistry, **26**, no. 16, 1701–1718, Dec. 2005.