John von Neumann Institute for Computing

Designing an Automatic Pipeline for Protein Structure Prediction

S. Kmiecik, M. Jamroz, A. Zwolinska,
P. Gniewek, A. Kolinski

published in

From Computational Biophysics to Systems Biology (CBSB08),
Proceedings of the NIC Workshop 2008,
Ulrich H. E. Hansmann, Jan H. Meinke, Sandipan Mohanty,
Walter Nadler, Olav Zimmermann (Editors),
John von Neumann Institute for Computing, Jülich,
NIC Series, Vol. **40**, ISBN 978-3-9810843-6-8, pp. 105-108, 2008.

http://www.fz-juelich.de/nic-series/volume40

Designing an Automatic Pipeline for Protein Structure Prediction

Sebastian Kmiecik[1], Michal Jamroz[1], Anna Zwolinska[1], Pawel Gniewek[1], and Andrzej Kolinski[2]

[1] Selvita Sp. z o.o.,
Ostatnia 1c, 31-444 Cracow, Poland
E-mail: sebastian.kmiecik@selvita.com

[2] Laboratory of Theory of Biopolymers,
Faculty of Chemistry, University of Warsaw,
02-093 Warsaw, Poland

Building accurate 3D structural models of proteins and protein assemblies is a challenging task. Our modeling technology is based on the CABS model, extensively tested, state-of-the-art approach to protein structure prediction. The modeling process is divided into two stages: CABS fold assembly followed by the model refinement/selection procedure, using an all-atom representation and a more exact interaction scheme enabling high resolution structure prediction. Fold assembly can be done in a framework of a standard comparative modeling procedure, where spatial restraints are derived from alternative sequence alignments with a template/templates. Preferentially in more difficult modeling cases, a new approach to comparative modeling can be used, which does not require the prior alignment. Selvita's goal is to provide an integrated tool-kit for automated protein structure predictions. However, like blind prediction experiments show, due to high complexity of prediction tasks, fully automated approach often doesn't guarantee the highest possible performance. Therefore, human intervention is made possible at every stage of modeling.

1 Introduction

Thanks to international effort in the genome sequencing projects, enormous library of protein sequences is now available. Despite extensive efforts in structural genomics, the number of experimentally determined protein structures, typically by costly X-ray crystallography or NMR spectroscopy procedures, is lagging far behind the number of known protein sequences. Since proteins are involved in practically all functions performed by a cell, knowledge of protein structures is necessary for understanding and controlling molecular mechanisms of life. Current assumptions are, that for a large fraction of proteins whose structures will not be determined experimentally, computational methods can provide valuable information[1].

2 Multiscale Approach to Structure Prediction: Comparative Modeling and Fold Recognition

During computational protein structure determination the following main challenges can be identified: 1) High accuracy structure prediction, at the resolution comparable to experimental methods, to enable predicted models utilization in a number of protein structure-based approaches (e.g. drug design, protein design, molecular docking, molecular replacement), which is now possible in Comparative Modeling (CM) cases[2], 2) Structure prediction of proteins or protein fragments for which sequence search methods failed to find

unambiguous homologs with known structure (Fold Recognition (FR) and New Fold (NF) prediction)

To meet criteria of both challenges, precise interaction scheme, sensitive to small atomic rearrangement, should be somehow combined with high efficiency in exploring proteins conformational space. That can be achieved by combining all-atom and reduced modeling: the multiscale modeling. Properly designed reduced models make possible very effective search of the protein's conformational space[3] and all-atom modeling enable exact scoring and refinement of the models. Our modeling technology is based on a such hierarchical approach[2]. Reduced-space search of the conformational space by the CABS[3] is followed by a reliable transition into the all-atom resolution and by subsequent fine-tuning and assessment of the final models. Such multiscale approach enable high-resolution protein structure predictions, predictions of protein interactions[4], computer-aided drug design and even study of protein dynamics[5].

CABS computational technology has been rigorously tested during CASP6 (Critical Assessment of Techniques for Protein Structure Prediction) world-wide experiment by the Kolinski-Bujnicki group, which ranked second best among over 200 groups participating, and ranked first when the consistency of the prediction was used as a criterion (the number of CASP targets placed in the top 20 of the best predictions)[6].

The design of CABS model enable easy implementation of spatial restraints. Such restraints can be derived by a large number of bioinformatics tools from appropriate known structures or from experimental sources e.g. from sparse NMR data. Therefore, essentially the same approach is possible at various levels of protein modeling difficulty from CM, to FR and NF cases. For the sake of flexibility two basic modeling pathways were designed and one alternative to make the prediction more effective. The entire prediction pipeline could by briefly outlined as follows (see the flowchart in the Figure 1): 1) Pre-processing: Template identification, secondary structure prediction, target- template alignments, input for more sophisticated user defined FR multiple alignments, 2a) Fast modeling track (easy CM cases) including fast scoring of alternative alignments and generation of spatial restraints, 2b) Rigorous modeling track (hard CM and FR cases) including 3D threading and generation of spatial restraints, 2c) Alternative modeling track by TRACER (hard CM and FR cases) - without prior alignments[7], 3) CABS modeling, 4) Post-processing: trajectory clustering, selection of clusters representatives, rebuilding from reduced to all-atom representation and finally all-atom models refinement and ranking.

Additionally in the most difficult cases (NF) ab initio modeling based only on target sequence can be performed (the accuracy of the resulting models is sometimes sufficient for structure-based protein function identification).

3 Automatic or Human Driven?

As blind structure prediction experiments demonstrated, human expert experience and intuition becomes a key point to the best possible performance, especially in difficult CM and FR[1]. Also in high resolution structure prediction, when a fraction of an Angstrom of the final model resolution matters, human intervention may by helpful by manual insertions of a template structure fragments into the final model. However, our goal is to develop fully automated structure prediction protocol which enable structure prediction on a genomic scale. Considering difficult modeling cases, the modeling approach without prior align-

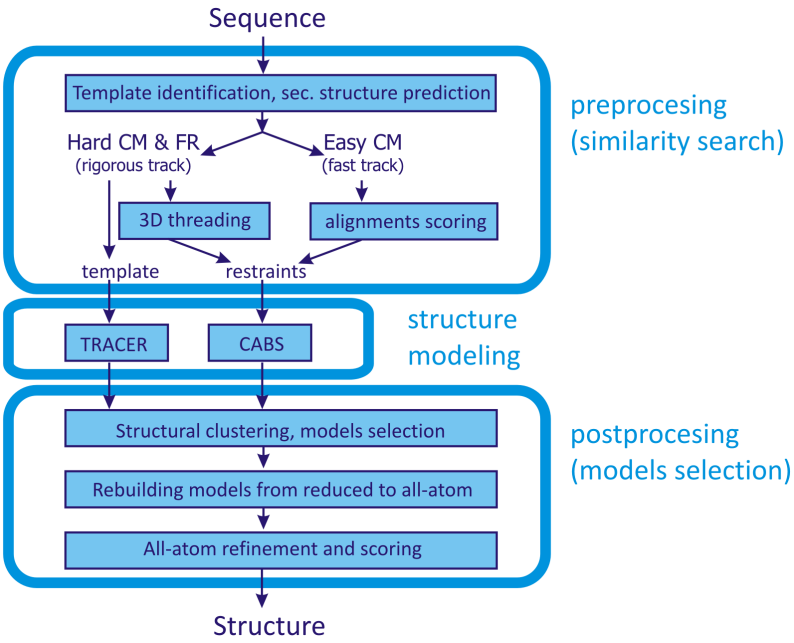

Figure 1. The protein structure prediction flowchart - see the text.

ments[7], included in our pipeline, seems to be an extremely promising step towards fully automated modeling (errors in alignments seem to be the main source of failures in protein structure prediction[1]).

References

1. O. Schueler-Furman, C. Wang, P. Bradley, K. Misura, D. Baker, *Progress in modeling of protein structures and interactions*, Science **310**, 638-42, 2005.
2. S. Kmiecik, D. Gront, A. Kolinski, *Towards high-resolution protein structure prediction. Fast refinement of reduced models with all-atom force field.*, BMC Structural Biology , 7:43, 2007.
3. A. Kolinski, *Protein modeling and structure prediction with a reduced representation*, Acta Biochim. Pol. **51**, 349-371, 2004.
4. M. Kurcinski, A. Kolinski, *Hierarchical modeling of protein interactions*, J Mol Model **13**, 691-8, 2007.
5. DA Debe, JF Danzer, WA Goddard, A. Poleksic, *STRUCTFAST: Protein sequence remote homology detection and alignment using novel dynamic programming and profile-profile scoring*, Proteins **64**, 960-967, 2006.
6. S. Kmiecik, A. Kolinski, *Characterization of protein-folding pathways by reduced-space modeling.*, Proc Natl Acad Sci USA **104**, 12330-5, 2007.
7. A. Kolinski, D. Gront, *Comparative modeling without implicit sequence alignments.*, Bioinformatics **23**, 2522-27, 2007.