John von Neumann Institute for Computing

High Throughput Method for Protein Structure Prediction

Dominik Gront, Sebastian Kmiecik, Andrzej Kolinski

published in

NIC Workshop 2006,
From Computational Biophysics to Systems Biology,
Jan Meinke, Olav Zimmermann,
Sandipan Mohanty, Ulrich H.E. Hansmann (Editors)
John von Neumann Institute for Computing, Jülich,
NIC Series, Vol. **34**, ISBN-10: 3-9810843-0-6,
ISBN-13: 978-3-9810843-0-6, pp. 79-82 , 2006.

© 2006 by John von Neumann Institute for Computing
Permission to make digital or hard copies of portions of this work for personal or classroom use is granted provided that the copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise requires prior specific permission by the publisher mentioned above.

http://www.fz-juelich.de/nic-series/volume34

High Throughput Method for Protein Structure Prediction

Dominik Gront, Sebastian Kmiecik, and Andrzej Kolinski

Laboratory of Theory of Biopolymers,
Faculty of Chemistry, Warsaw University
Pasteura 1, 02-093 Warsaw, Poland
E-mail: dgront@chem.uw.edu.pl

Recently several successful methods for protein structure prediction have been proposed. Next step towards modeling on a genomic scale is to combine existing tools into a single automated protocol. Such methods are crucial to fill the gap between the number of currently known protein sequences and structures. Here we utilize a lattice based coarse-grained modeling algorithm together with several accompanying tools to build a generalized pipeline for protein structure prediction. Our strategy was successfully applied during the CASP6 experiment.

1 Introduction

Building accurate 3D structural models for protein sequences of unknown structure is a challenging problem in contemporary computational biology. Large-scale genomic sequencing efforts provide increasing number of sequences. On the contrary the number of experimentally determined structures remains relatively small. Several computational methods that have been recently proposed can help to narrow this gap. The gap however is still getting wider. Therefore, development of high throughput protocol that would be able to predict protein structures with no human intervention becomes an urgent task.

2 Generalized Approach to Structure Prediction in a Reduced Conformational Space

The CABS modeling tool has been designed in a way allowing easy implementation of various restraints. Such restraints could be derived theoretically using various bioinformatics tools and databases of known structures, or experimentally from sparse NMR data, site-directed mutagenesis, etc. The approach is called ,,generalized", since essentially the same strategy was employed to all types of protein targets, from comparative modeling (CM), through the fold recognition category (FR), to the most difficult new fold (NF) cases. This strategy combines FRankenstein (FR - Fold Recognition) method of Bujnicki[12] used for derivation of structural restraints, CABS simulation employed in a consensus model building, clustering and evaluation of models. The ,,FRankenstein Monster" algorithm builds models from large molecular fragments, properly extracted from the databases of known protein structures. These models are often quite accurate in the regions of regular secondary structure, while the loop connections are usually poorly predicted. The general idea was to generate a number of FRankenstein models, extract a large number of distance restraints (often self-contradictory) from these models and to apply them as the set of soft constraining potentials in the CABS simulations. Only the Cα -Cα distance restraints were used. The entire prediction pipeline could be outlined as follows (see flowchart on Figure 1a):

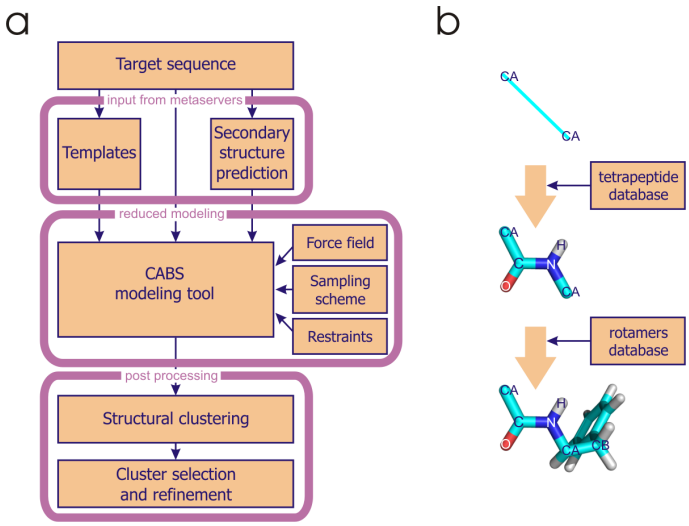

Figure 1. a) Flowchart representing main stages in protein structure prediction by the CABS modeling tool. b) Two-step procedure for building an all-atom model from its reduced representation

- Using GeneSilico metaserver[12] plausible templates (or their fragments) were identified, secondary structure of the target predicted, and FR alignments generated.
- Basing on the data from the metaserver the Frankenstein's all-atom models were built.
- The obtained models were evaluated by VERIFY3D method[2], and the high scoring fragments of each model were used as a source of distance restraints. The poorly scoring fragments of the initial models were left unrestrained. In the cases of the apparent New Folds very weak restraints were derived from fragments of FR models and from few alternative models generated by the Robetta server[3].
- A number of copies of the target structure was generated using the CABS lattice discretization and the initial FRankenstein's models. These constituted the set of the starting replicas for a long simulated tempering Monte Carlo simulations. To obtain a sufficient number of replicas, not only different templates, but also different alignment variants were used in this stage.
- Thousands models in reduced representation from CABS simulations were clustered using HCPM (Hierarchical Clustering of Protein Models)[4,5] algorithm, leading to 5-20 clusters, depending on the degree of convergence of the MC simulations. For a representative alpha-carbon backbone from each cluster a full-atom model was rebuilt.

This procedure has proven to be very efficient. Good predictions were achieved in all categories of the CASP targets and the group Kolinski-Bujnicki[1] using the outlined methodology has been classified as the second best, when averaging scores from all categories. Interestingly, for a number of targets the predicted molecular models were

closer to the true (released after the CASP meeting) structures of the targets than to any of the templates employed in modeling. In some cases of the NF targets reasonable models were obtained that were qualitatively different from all templates used. Apparently, the self-contradictory restraints became „diffused" by the CABS force field, while a small subset (for instance defining a plausible supersecondary element of the fold) of qualitatively correct restraints restricted sufficiently the conformational space during the simulations.

In the high-throughput modeling, especially when many proteins are modeled in a similar way, data processing in an automated fashion is especially important. As a result of our experience gained during the long process of multistage modeling of all CASP targets, we created a BioShell package[6]. BioShell is a way of performing and integrating computations employing UNIX shell as an interface platform. The package itself is composed of several programs, which extend functionality of shell or scripting languages, such as Python or Perl. Most of the examples given in the BioShell website (http://biocomp.chem.uw.edu.pl/) are related to the CABS modeling tool. Nevertheless, owing to standard file formats and algorithms, BioShell programs can be applied to virtually any modeling protocol.

3 Reconstructing the Full-Atom Representation

Simulations of reduced lattice models are usually a couple of orders of magnitude faster than simulations employing equivalent all-atom continuous models[1]. One of the major drawbacks is the necessity to reconstruct all-atom representation from approximate coordinates of Cα atoms. Here we describe a very efficient algorithm that can be applied on a genomic scale (Figure 1b)

For the non-Cα backbone atoms average relative positions were derived in a local coordinate system by statistical analysis of PDB structures. Our method follows previous approaches by Purisima and Scheraga[7] and by Milik[8]. Assuming a constant length of all Cα -Cα vectors, each backbone configuration for a tetrapeptide can be described by three Cα distances: R_{13} (between first and third Cα atoms in a tetrapeptide), R_{24} and R_{14} (defined similarly to R_{13}). The chirality of the backbone is taken into account by applying a sign to R_{14}. Negative values represent left-handed, and positive values represent right-handed conformations. These three distances form a three-dimensional grid in which average positions of C, O, and N atoms are accumulated from the PDB structures according to the local backbone configuration measured by R_{13}, R_{24}, and R_{14} after transformation into the local coordinate system. To make our implementation more accurate the grid spacing was chosen as 0.2 Å , rather than the 0.3 Å used by Milik et al.[8]. In order to reconstruct the C, O, and N backbone atoms, for each tetrapeptide average positions for C, O, and N are taken from the grid described above according to R_{13}, R_{24}, and R_{14} and transformed back into the original coordinate system. In the final step of the modeling, side chain atoms are reconstructed. For this purpose the SCWRL[9] method can be used.
Our approach to backbone reconstruction turned out to be reliable, fast and very accurate. In a test on a set of native structures taken from PDB we compared our algorithm with several existing methods: bb[10], MaxSprout[11], Pulchra[12] and Sybyl (Tripos) program which implements an algorithm by Claessens et al[13]. MaxSprout and S ybyl are the methods that

utilize fragment libraries derived from known structures. In many cases the two methods were not able to find a suitable fragment: MaxSprout succeeded only in 46% and Sybyl in 91% of the native Cα traces in a test set. Among the other three methods our approach was found to be both the fastest and the most accurate. Average crmsd error for 70 chains measured on the fully reconstructed backbone is 0.42Å .

Acknowledgments

This work was partially supported by the grant # PBZ-KBN-088/P04/2003

References

1. A. Kolinski and J. M. M. Bujnicki. Generalized protein structure prediction based on combination of fold-recognition with de novo folding and evaluation of models. *Proteins*, September 2005.
2. D. Eisenberg, R. Luthy, and J. U. Bowie. Verify3d: assessment of protein models with three-dimensional profiles. *Methods Enzymol*, 277:396–404, 1997.
3. D. Chivian, D. E. Kim, L. Malmstrom, P. Bradley, T. Robertson, P. Murphy, C. E. Strauss, R. Bonneau, C. A. Rohl, and D. Baker. Automated prediction of casp-5 structures using the robetta server. *Proteins*, 53 Suppl 6:524–533, 2003.
4. D. Gront and A. Kolinski. HCPM–program for hierarchical clustering of protein models. *Bioinformatics*, 21(14):3179–3180, July 2005.
5. D. Gront, U. H. E. Hansmann, and A. Kolinski. Exploring protein energy landscapes with hierarchical clustering. *Int J Quantum Chem*, 105(6):826–830, 2005.
6. D. Gront and A. Kolinski. Bioshell - a package of tools for structural biology computations. *Bioinformatics*, 22(5):621–622, March 2006.
7. E. O. Purisima and H. A. Scheraga. Conversion from a virtual-bond chain to a complete polypeptide backbone chain. *Biopolymers*, 23(7):1207–1224, July 1984.
8. M. Milik, A. Kolinski, and Jeffrey Skolnick. Algorithm for rapid reconstruction of protein backbone from alpha carbon coordinates. *Journal of Computational Chemistry*, 18(1):80–85, 1996.
9. Jr Dunbrack and M Karplus. Backbone-dependent rotamer library for proteins application to side-chain prediction. *Journal of Molecular Biology*, 230(2):543–574, March 1993.
10. S. A. Adcock. Peptide backbone reconstruction using dead-end elimination and a knowledge-based forcefield. *J Comput Chem*, 25(1):16–27, January 2004.
11. L. Holm and C. Sander. Database algorithm for generating protein backbone and side-chain co-ordinates from a c alpha trace application to model building and detection of co-ordinate errors. *J Mol Biol*, 218(1):183–194, March 1991.
12. M. Feig, P. Rotkiewicz, A. Kolinski, J. Skolnick, and C. L. Brooks. Accurate reconstruction of all-atom protein representations from side-chain-based low-resolution models. *Proteins: Structure, Function, and Genetics*, 41(1):86–97, 2000.
13. M. Claessens, E. van Cutsem, I. Lasters, and S. Wodak. Modelling the polypeptide backbone with 'spare parts' from known protein structures. *Protein Eng.*, 2(5):335–345, January 1989.