

Protein structure prediction by combination of fold-recognition with de novo folding

Janusz M. Bujnicki¹, Marcin Feder¹, Michal J. Gajda¹, Jan Kosinski¹, Marcin Pawlowski¹, Michal Boniecki^{1,2}, Dominik Gront², Andrzej Kolinski²

Keywords: homology modeling fold-recognition, threading, protein folding, distance restraints,

1 Introduction.

A new method for protein structure prediction was developed, which allows modeling regardless of the potential homology with any known protein structure. It is based on a combination of the "FRankenstein's Monster" approach for comparative modeling (CM) by recombination of fold-recognition (FR) models [1], with a de novo method for Replica Exchange Monte Carlo protein folding simulation using a new high-resolution lattice representation CABS [2]. This method was validated in the course of the CASP-6 competition (Critical Assessment of protein Structure Prediction, Summer'2004). Among >200 participants, the group BUJNICKI-KOLINSKI obtained the 2nd position in the overall ranking (3rd position in the Comparative Modeling category and 2nd in the New Folds), while GeneSilico, a group of students from IIMCB, who used FRankenstein and ROSETTA [3] for de novo folding, instead of CABS, obtained the 5th position in the overall ranking, with the 2nd position in the Fold Recognition-Homologous category.

2 Methodology

The sequence of a modeled „target” protein is submitted to the GeneSilico structure prediction meta server [4], which is a gateway to a variety of third-party methods for secondary structure prediction and fold-recognition (FR). The structural core is predicted based on the consensus between different FR methods. For each fold, the target-template alignments are used as a starting point for model refinement using the “FRankenstein’s monster” approach [1]. A “monster” is constructed from fragments that either appear in >40% of initial models or are assessed as most “protein-like” according to VERIFY3D [5] (run via COLORADO3D [6]). The hybrid model is superimposed on the previously used template structures to generate a corresponding target-template alignment. Refined models are constructed by iterating the homology modeling procedure, evaluation of the sequence-structure fit, merging of fragments with best scores, and local realignment in poorly scored regions.

Fragments with best scores are used to derive tertiary restraints. Additional tertiary restraints can be derived from methods for de novo structure prediction, such as ROSETTA [3] and/or from experimental data (e.g. inter-residue distances from cross-linking, shape of the molecule determined by electron microscopy etc.). Secondary structure restraints are derived from the consensus of methods implemented in the GeneSilico meta server [4]. Secondary and tertiary restraints are used to guide the REMC/CABS folding simulation [2,3]. The conformations obtained in the course of CABS simulations are subject to the average linkage hierarchical clustering. For a

¹ Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology in Warsaw, Poland, <http://genesilico.pl>. E-mail: iamb@genesilico.pl

² Laboratory of Theory of Biopolymers, Faculty of Chemistry, University of Warsaw, Poland. E-mail: kolinski@ptb1.chem.uw.edu.pl

representative structure from each cluster, a full-atom representation is rebuilt, to produce the final set models, which can be re-evaluated and re-ranked using independent criteria (such as protein-like topology, agreement with experimental data etc.).

3 Strengths and weaknesses of the method:

Our method is fully automatable and scalable. Actually, among the fully-automated or nearly-fully-automated methods, it was ranked as the best in the overall classification (we were exceeded only by Ginalski, who used a lot of manual intervention). Our method is able to build models closer to the native structure than any of the available templates (i.e. much better than practically all homology modeling programs). It is also able to build models de novo, i.e. without any templates and to use restraints from multiple sources (preliminary 3D models obtained by other methods, secondary structure, experimental data). The drawback of the method is that it seeks for thermodynamic minima, so proteins with folds determined by kinetic control may not be modelled correctly. Nonetheless, such proteins are believed to be a minority. Another problem is that despite the structure of domains and subdomains is often predicted correctly, their mutual orientation is not always right. This is, however, a common problem of all modeling methods and in our case it can be easily circumvented by inclusion of sparse experimental data that provide distance restraints for the individual domains. Finally, full-atom refinement and energy evaluation of the resulting models remain to be implemented – we hope that these improvements will allow us to improve the ability to improve the ranking of the final models (i.e. to increase the ability to rank the truly best model at the top position of the ranking).

4 Acknowledgements

The FRankenstein method could not exist without the availability of third-party methods and servers. We would like to thank all developers, in particular: DBaker, GBarton, RDunbrack, D.Eisenberg, A.Elofsson, L.Jaroszewski, D.Jones, A.Godzik, K.Karplus, L.Kelley, J.Meller, J.Meiler, K.Mizuguchi, and B.Rost. The development of our methods for protein structure prediction is supported by the KBN (grant PBZ-KBN-088/P04/2003). JMB was supported by the EMBO/HHMI Young Investigator Award

References

- [2] Kolinski, A. 2004. Protein modeling and structure prediction with a reduced representation. *Acta Biochim Pol*, 51, 349-71.
- [1] Kosinski, J., Cymerman, I.A., Feder, M., Kurowski, M.A., Sasin, J.M., Bujnicki, J.M. 2003. A "FRankenstein's monster" approach to comparative modeling *Proteins*, 53 Suppl 6, 369-79.
- [4] Kurowski, M.A., Bujnicki, J.M. 2003. GeneSilico protein structure prediction meta-server. *Nucleic Acids Res*, 31, 3305-7.
- [5] Luthy, R., Bowie, J.U., Eisenberg, D. 1992. Assessment of protein models with three-dimensional profiles. *Nature*, 356, 83-5.
- [6] Sasin, J.M., Bujnicki, J.M. 2004. COLORADO3D, a web server for the visual analysis of protein structures. *Nucleic Acids Res*, 32, W586-9.
- [3] Simons, K.T., Kooperberg, C., Huang, E., Baker, D. 1997. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J Mol Biol*, 268, 209-25.