John von Neumann Institute for Computing

Semiautomatic Workflow for Fold Recognition – Results from the CASP 2006 Competition

F. Fredel, J. Meinke, S. Mohanty, O. Zimmermann,
U. H. E. Hansmann

published in

From Computational Biophysics to Systems Biology (CBSB07),
Proceedings of the NIC Workshop 2007,
Ulrich H. E. Hansmann, Jan Meinke, Sandipan Mohanty,
Olav Zimmermann (Editors),
John von Neumann Institute for Computing, Jülich,
NIC Series, Vol. **36**, ISBN 978-3-9810843-2-0, pp. 113-116, 2007.

http://www.fz-juelich.de/nic-series/volume36

Semiautomatic Workflow for Fold Recognition – Results from the CASP 2006 Competition

Fabian Fredel[1], **Jan Meinke**[1], **Sandipan Mohanty**[1], **Olav Zimmermann**[1], **and Ulrich H. E. Hansmann**[1,2]

[1] John von Neumann Institute for Computing,
Research Centre Jülich, 52425 Jülich, Germany
E-mail: {*f.fredel, j.meinke, s.mohanty, olav.zimmermann, u.hansmann*}*@fz-juelich.de*

[2] Department of Physics, Michigan Technological University,
1400 Townsend Drive,
Houghton, MI 49931, USA
E-mail: hansmann@mtu.edu

We outline a semi-automatic procedure for structure prediction of proteins. A first analysis of the performance of this procedure in the CASP 2006 competition is presented.

1 Introduction

In order to test combinations of physics-based simulation techniques and sequence-based prediction methods, our group participated in the "Critical Assessment of Techniques for Protein Structure Prediction" (CASP) competition in the summer of 2006. As a first-time participant our goal was to establish a semi-automatic workflow. We combined existing methods for fold recognition with our refinement algorithms testing heuristics for the selection at each step. In this article, we give an overview of the workflow and the results of an in-depth statistical analysis of our results. In particular, we assess the significance of measured performance differences between the prediction methods. Analyzing our workflow, we try to find the critical points where alternative decisions lead to a significant change in the results. Our aim is to obtain rules that guide the decision process in the workflow to optimize our predictions.

2 Workflow

The first step in our workflow is the manual selection of templates from 3D-Jury[1] predictions. Preference is given to high 3D-Jury-scores and agreement between the secondary structure of the template and the predicted secondary structure of the target sequence. For targets that were obviously not comparative modeling targets, 3D-Jury predictions from fold recognition servers are preferred.

We search the fold space[2] using CABS[3]. This parallel tempering Monte Carlo program uses constraints from the respective 3D-Jury templates and secondary structure prediction by PSIpred 2.5[4]. We use 32 replicas for sequences with less than 200 residues and 64 replicas for proteins with longer sequences. The statistics was between 15,000 sweeps for long sequences and 100,000 sweeps for short sequences.

Clustering is performed using hierarchical clustering with HPCM[5] using a fixed RMSD of 2.5 Å as clustering radius. Structure clusters are selected based on cluster averages of CABS energy and structure similarity (TM-score[6]) to the PDB structure on which the 3D-Jury template was based.

Averaged structures from the selected clusters are subject to regularization by SMMP[7]. Regularized structures are ranked according to the total and partial energies of the structures in SMMP. In ambiguous cases, the consistency of this ranking with a similar ranking based on energy terms of PROFASI[8] is checked.

The 5 to 10 structures ranked best are selected for refinement. For most structures, refinement involves a set of constrained simulated annealing runs with SMMP, starting from very high temperatures. Most structures dissolve and re-form into local minima of the potential that are close to the input structures of the refinement procedure. The final structures from different annealing trajectories are once again ranked following a similar procedure as for the initial selection for the refinement runs. Final selection and ranking is based on several energy terms, secondary structure content and visual inspection.

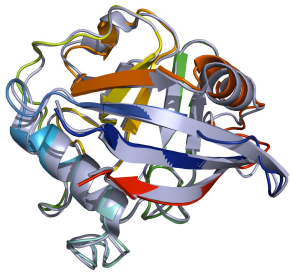

Figure 1. Grey is the experimental structure (2HE9). Colored is our best structure for T0346.

3 Comparing Prediction Methods

Not all of our predictions are as good as the one in figure 1. In order to assess the significance of measured performance differences between us and other prediction methods we use a nonparametric statistical test, the Friedman test[9]. It has a simple two-way layout for k treatments (groups) and n blocks (target).

TARGET	GROUP A_1 A_2 A_3 ... A_k	$R_i = \sum_{j=1}^{k} r_{ij}$
#1	r_{11} r_{12} r_{13} ... r_{1k}	R_1
#2	r_{21} r_{22} r_{23} ... r_{2k}	R_2
#3	r_{31} r_{32} r_{33} ... r_{3k}	R_3
... #n	r_{n1} r_{n2} r_{n3} ... r_{nk}	R_n
$R_j = \sum_{i=1}^{n} r_{ij}$	R_1 R_2 R_3 ... R_n	$\sum_{j=1}^{k} R_j = \sum_{i=1}^{n} R_i$

For each of the n experiments the k results are ranked from 1 for the best to k for the worst result. The ranked results are based on TM-score and RMSD for the predictions to the experimental structure. The Null-hypothesis of the test is: $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ (no treatment differs). The rank of group j at experiment i is given by r_{ij}. R_j is the sum of the n ranks of group j. $\bar{R_j}$ is the mean of the n ranks of group j and $R.. = \frac{k+1}{2}$.
Compute

$$S = \frac{12\,n}{k\,(k+1)} \sum_{j=1}^{k} (\bar{R_j} - R..)^2.$$

S is an approximation for the χ^2 distribution.

The Nullhypothesis is rejected in all test cases. Across the field we find significant differences between all servers. The best servers are *Zhang-Server*, *MetaTasser*, *Pmodeller 6* and *BayesHH* outperforming our procedure and demonstrating the need for further improvement of our workflow.

4 Workflow Analysis

For this reason, we have decided to search the workflow for critical points where alternative decisions lead to significant changes and improvements in our results. We asked ourselves the following questions:

- Do we select the best template?
- Do we trust PSIpred for the secondary structure prediction?
- Which structures should be used for clustering?
- Which clusters are the best?
- Is the final ranking of energy terms ideal to find the best structure?

As an example, we show the analysis of the workflow for target T0354 (130 residues):

best template selected ?		best template	our template
	RMSD	3.98	3.89
	TM-score	0.515	0.395
do we trust PSIpred ?	good secondary structure prediction by PSIpred		
which structures for clustering ?	good predictions often at the lower replica numbers		
best clusters are?	average of CABS energy not always the best		
ranking of energy terms ideal ?	structure ranked 1st	TM-score 0.3323	
	best structure ranked 12th	TM-score 0.5085	

5 Conclusion

We have described a method for structure prediction of proteins. While currently not competitive with other approaches, we have shown a way to analyze its performance and to explore possible improvements.

Acknowledgment

The calculations were done on computers of the John von Neumann Institute for Computing, Forschungszentrum Jülich, Jülich, Germany.

Ulrich H. E. Hansmann is supported by a research grant (GM62838) of the National Institute of Health (USA).

References

1. K. Ginalski, A. Elofsson, D. Fischer and L. Rychlewski, Bioinformatics **19**, 1015-1018, 2003.
2. U. H. E. Hansmann, Chem. Phys. Lett. **281**, 140, 1997.
3. A. Kolinski and J. M. Bojnicki, Proteins **61 Suppl. 7**, 84-90, 2005.
4. D. T. Jones, J. Mol. Biol. **292**, 195-202, 1999.
5. D. Gront and A. Kolinski, Bioinformatics **21**, 3179-3180, 2005.
6. Y. Zhang and J. Skolnick, Proteins **57**, 702-710, 2004.
7. F. Eisenmenger, U. H. E. Hansmann, S. Hayryan and C. K. Hu, Comp. Phys. Comm. **174**, 422, 2006.
8. A. Irbaeck and S. Mohanty, J. Comp. Chem. **27**, 1548, 2006.
9. Myles Hollander, Douglas A. Wolfe (1973), *Nonparmetric Statistical Methods*, John Wiley & Sons (ISBN 0-471-40635-X).