ROKKO

Template-free Prediction by Fragment Assembly with SimFold Energy Function at CASP7

S.J. Park[1], N. Hori[2], K. Okazaki[2] and S. Takada[12]
[1] - *Faculty of Sci, Kobe Univ,* [2] - *Grad School, Sci & Tech Kobe Univ*
stakada@kobe-u.ac.jp

Team ROKKO primarily focused on predicting structures that need template-free modeling and could have previously unseen folds. Prediction method, statistics, and short description for each target are available at http://www.proteinsilico.org/ROKKO/casp7/rokko_casp7_strategy.html.

(1) *General Workflow*: All targets automatically stream to the general sequence analysis procedure. BLAST[1] package first searches homologous templates through NRDB, and then mainly PSI-PRED[2] predicts secondary structure elements (SSEs) using filtered NRDB. Some of DBs used are weekly updated. If significant templates for a target are found, all available information on the templates is gathered for selecting high-resolution structural templates (See (2)). When templates do not exist, 3D-Jury[3] templates and alignments are gathered. When we did not get reliable templates, we performed fragment assembly simulations either by MCFA and/or GAFA (See (5) and (6)).

(2) *Template-based Prediction*: If we are satisfied with the quality of a template BLAST found, we sample template-target sequence alignments using the stochastic backtracking procedure[4] (over 100 sub-optimal alignments). When several templates are covering distinct target regions, we randomly pick alignments from each ensemble of the sub-optimal alignments, and input them as initial alignments of the progressive multiple sequence alignment (approximately 1000-3000 alignments). We also use template-target profile alignments when PSI-BLAST found templates with relatively higher E-value (>0.001). We convert the alignments to 3D structures by running MODELLER[5], and evaluate them using both of Verify3D[6] and Prosa[7] to check the initial alignment quality. We iteratively run MODELLER with seemly good alignments, and repeatedly checked SSEs and the quality of local/global structures. After ending this iterative procedure, we select final models from the 2D score distribution generated by Verify3D and Prosa.

(3) *Fragment Library Construction*: For template-free prediction, we first build a set of 10-residue segments by comparing the feature vector of a target sequence with them of library containing 2598 known-structure proteins that share <25% sequence identity. The vector contains PSSM of PSI-BLAST, grouped chemical property of a residue, and a SSE. Two types of fragment libraries are generated. **Type I**; a correlation coefficient scores top 200 segments for each overlapping 10-residue fragment of a target. **Type II**; five scoring functions including the correlation coefficient pick over 200 segments by considering the degree of dominated level (often called "Pareto Frontier" in multi-objective optimization field).

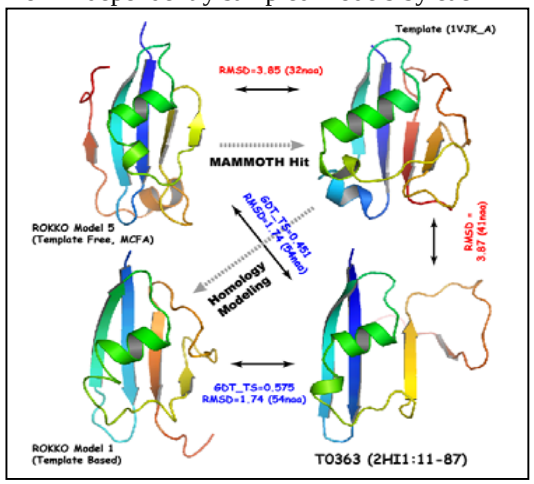
(4) *SimFold Energy Function*: For fragment assembly simulations, we solely used a coarse-grained model, SimFold[8,9], in which side chain atoms are replaced with a center of mass. SimFold contains van der Waals interaction, secondary structure propensity, hydrogen bond interaction, hydrophobic interaction, and pairwise interaction. The latter three terms depend on the degree of burial of interacting atoms. No protein specific potential such as secondary structure prediction based potential is used in the energy function. Parameters in SimFold are optimized by Z-score optimization method.

(5) *Multi-Canonical Ensemble Fragment Assembly (MCFA)*: Using Type I fragment library, we performed the reversible MCFA[10] that fulfills detailed balance condition. The predictive accuracy of our MCFA in *de novo* prediction has been proved in CASP6. On the other hand, to define a reasonable weight function of MCFA is very time-consuming and human-dependant. We applied, therefore, Wang-Landau algorithm[11] to the MCFA with a slight modification. We arranged the reducing schedule of the Wang-Landau factor by using our empirical data, and defined a weight function through approximately 2-3 billion Monte Carlo steps in a MCFA. Independent MCFA for a target sampled conformations as many as possible by the given time limit.

(6) *Genetic Algorithm Fragment Assembly (GAFA)*: Using Type II fragment library, we test a Genetic Algorithm newly developed (its basic code came from the earlier study[12]). With 100 initial random coils, the GA randomly selects a residue as a crossover point. Based on this point, GA shuffles the two parents randomly selected from the current population, and replaces a segment (4-10 residues long) to a fragment coming from the library. After generating 200 offspring, GA updates the parents with the lowest-energy child and random one. Thus, the initial coils hopefully evolve to the lowest-energy conformations

through 5000 update steps. The final conformations of independent GAFA runs are gathered as many as possible, and are analyzed.

(7) *How We did in CASP7*: We performed the template-based procedure for predicting targets that have significant PSI-BLAST E-value (< 0.001) or 3D-jury jscore (>50.0). For all remaining targets, we conducted MCFA and/or GAFA with the different types of fragment libraries. When there exist long alignment gaps (>20 residues) or probably unseen domains (e.g. T0311, T0347, etc.), we first predicted a full-length model with a template and ran FA to predict these broken regions. For a target that is likely to have multiple domains, we parsed it into monomers based on domain DBs, and combined them into a single chain by FA. 13 targets were predicted by the consensus of MCFA and GAFA; by using cluster analysis and visual inspection, we selected five models from independently sampled models by each FA method.



Interestingly, we often found that some of models FA predicted have high structural similarity to known proteins. In such cases, we added a template-based model to the final models if we were confident. For example, in T0363 case, we first selected five models from MCFA samples. MAMMOTH[13] said that all five models are considerably similar to a Beta Grasp Fold. Particularly, model_5 was highly similar to 1MG4_A (z_score=4.92) that is akin to 3D-Jury templates. We believed, consequently, that 1VJK_A (jscore=46.88) is the best template for T0363. Such conspicuous structural similarity with remote homology was found from FA models of several targets (e.g. T0304, T0349, T0353, T0361, T0382, etc.). Surprisingly, a model of SimFold FA for T0383 culled 1QYN_A (jscore=6.25) that is 3.66 Angstroms over 70 residues of the T0383 native.

It is deemed again that SimFold FA is feasible to capture the native-like interactions from high quality fragment library. Therefore, the reliability of structural templates fold recognition servers detected can be confirmed to increase the predictive accuracy. This will be a steppingstone to better prediction of new folds.

1. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.
2. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* ***292***, 195-202.
3. Ginalski,K., Elofsson,A., Fischer,D. & Rychlewski,L. (2003) 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics* **22**, 1015-1018.
4. Muckstein,U., Hofacker,I.L., & Stadler,P.F. (2002) Stochastic pairwise alignments. *Bioinformatics* **18**, S153–S160.
5. Sali,A. & Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779-815.
6. Bowie,J.U., Luthy,R. & Eisenberg,D. (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **253,** 164-170.
7. Sippl,M.J. (1993) Recognition of errors in three-dimensional structures of proteins. *Proteins* **17**, 355-362.
8. Fujitsuka,Y., Takada,S., Luthey-Schulten,Z.A. & Wolynes,P.G. (2004) Optimizing physical energy functions for protein folding. *Proteins* **54**, 88-103.
9. Fujitsuka,Y, Chikenji,G, & Takada,S. (2006) SimFold energy function for de novo protein structure prediction: Consensus with Rosetta. *Proteins* **62**, 381-398.
10. Chikenji,G., Fujitsuka,Y. & Takada,S. (2003) A reversible fragment assembly method for de novo protein structure prediction. *J. Chem. Phys.* **119**, 6895-6903.
11. Wang,F. & Landau,D.P. (2001) Efficient, multiple-range random walk algorithm to calculate the density of states. *Phys. Rev. Lett.* **86**, 2050–2053.
12. Park,S.J. (2005) A study of fragment-based protein structure prediction: biased fragment replacement for searching low-energy conformation. *Genome Informatics* **16**, 104-113.
13. Ortiz,A.R., Strauss,C.E. & Olmea,O. (2002) MAMMOTH (Matching molecular models obtained from theory): An automated method for model comparison. *Protein Science* **11**, 2606-2621.