

# Screening and Refinement of Protein Structures from Fold Recognition

R. Zhou<sup>1\*</sup>, B. D. Silverman<sup>1</sup>, G. Dent<sup>1</sup>, A. Royyuru<sup>1</sup>,  
A. Curioni<sup>2</sup>, and A. Logen<sup>2</sup>

**Keywords:** Protein structure refinement, structure screening, fold recognition, hydrophobic moment profiling, replica exchange method

## 1 Introduction.

Protein structure prediction has been of great interest recently, as witnessed by the past six worldwide competitions in Critical Assessment of protein Structure Prediction techniques (CASP). Even with enormous efforts from various groups, protein structure prediction remains largely an unsolved problem. Somewhat encouraging though, the Fold Recognition (FR) category is making the most noticeable progress among the three major categories during the latest CASP experiments. However, for a typical FR model, the RMSD from the native structure can be somewhere from 3-4Å as well as greater than 10 . Thus, how to effectively screen and refine the FR models becomes a critical task. In this study, we use a combination of four scoring functions and an extensive sampling technique to systematically screen and refine the protein structures.

## 2 Screening of Models from Fold Recognition

The initial protein structure models are from PROSPECT[1], a collaborative work with the Xu group. PROSPECT employs both sequential and structural information for fold recognition and threading alignment. The evolutionary information is used not only in the profile-profile sequence alignment score, but also in calculating the singleton and pair-wise energies, which greatly improves the performance of both fold recognition and alignment accuracy[1]. The models from PROSPECT are then screened and refined. Four scoring functions are used to screen the initial models. The first scoring function is the OPLSAA/PB energy, which is the minimized total energy of the protein using the OPLSAA force field and the Poisson-Boltzmann (PB) continuum solvent model; the lower the energy, the better the model. The second scoring function is the Hydrophobic Score, or compactness, which is defined as the surface area under the normalized second-order hydrophobic moment profile using an ellipsoidal description of protein shape[2] (see Figure 1). The third scoring function is the Correlation Score, which is defined as the correlation coefficient between the distance of a residue from the center of the protein and its hydrophobicity (also shown in Figure 1). The fourth scoring function is a meta-score, mScore, which combines three scoring functions based on different grounds: a statistics based pairwise Calpha-Calpha distance dependant potential of mean force, a physics based non-bonded interaction energy of the GROMOS force field for the protein, and a phenomenological based Hydrophobic Score as described above. The mScore is expressed as a linear combination of the three normalized individual scores, with the coefficients pre-determined by maximizing the ranking of the native conformations versus decoys on several

---

<sup>1</sup> Computational Biology Center, IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598. <sup>2</sup>Computational Biochemistry and Material Science, IBM Zurich Research Lab, 8003 Rueschlikon, Switzerland. E-mail: ruhongz@us.ibm.com

decoy sets publicly available. For all of the last three scores, the higher the score, the better is the model.

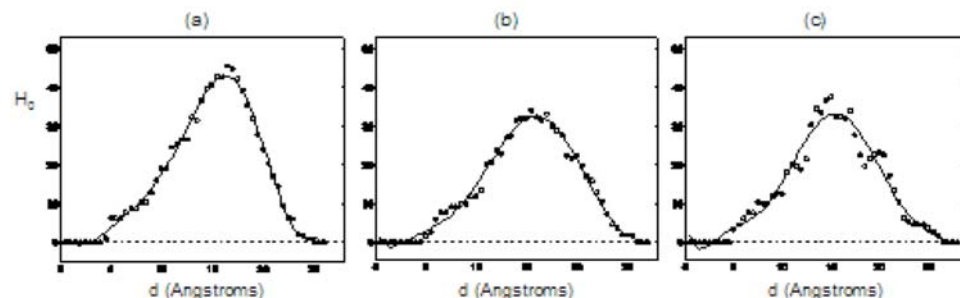


Figure 1. One example of the hydrophobic scores. The RMSD, Raggedness, Compactness and Correlation are (a) 1ash (0.0, 2.15, 2.66, 0.65); (b) 1ash\_1hsy\_r (2.63, 3.45, 2.03, 0.55); (c) 1ash\_hda-a\_r (2.98, 5.62, 2.14, 0.54).

### 3 Refinement of Selective Models

The best models identified by the screening process are then refined with the Replica Exchange Molecular Dynamics method (REMD), which is a powerful tool for efficient sampling of conformational space. The REMD method couples molecular dynamics trajectories with a temperature exchange Monte Carlo process for efficient sampling of the conformational space. In this method, replicas (total 12 in our implementation) are run in parallel at a sequence of temperatures ranging from the desired temperature to a high temperature at which the replica can easily surmount the energy barriers. From time to time the configurations of neighboring replicas are exchanged based on a Metropolis criterion. Because the high temperature replica can traverse high energy barriers, this provides a mechanism for lower temperature replicas to overcome the quasi-ergodicity they would otherwise encounter with a single temperature replica. The sampled conformations are then clustered and ranked based on the minimized total energy. Figure 2 shows one example of the energies/RMSDs of sampled configurations as well as the best refined structure. The force field used in sampling is the OPLSAA force field with continuum solvent models. For further improvement, some biased sampling might be needed to somehow constrain the sampling around the “native-like fragments” based on the confidence level.

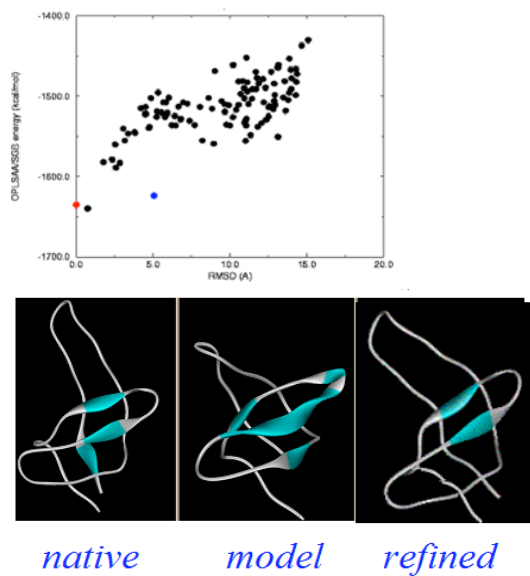


Figure 2. One example of refinement. The energies and RMSDs of sampled configurations are shown on the left (a) (blue: model, red: native), and the native, model and refined structures are shown on the right (b).

## References

1. Guo, J-T, Ellrott, K, Chung, WJ, Xu, D, Passovets, S, Xu, Y. (2004). PROSPECT-PSPP: An Automatic Computational Pipeline for Protein Structure Prediction. *Nucleic Acids Res.*, 32, W522-5.
2. Zhou R, Silverman BD, Royyuru A, Athma P. 2003. Spatial profiling of protein hydrophobicity: Native vs. decoy structures. *Proteins: Struc. Func. & Genetics* 52: 561-572