

## CLB3Group

### **Predictor@home: A Multiscale, Distributed Approach for Protein Structure Prediction**

C. An, M. Taufer and C.L. Brooks III

*The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA 92037, USA*

brooks@scripps.edu

#### **Motivation**

In the previous CASP exercises we focused our efforts on addressing basic algorithmic and/or scientific questions related to the scoring of predicted protein structures and their refinement via all atom models. Retrospective analysis of our approaches and methods from these experiences suggested that when native-like protein conformations were sampled they could be identified with all atom physics-based force fields including implicit solvation<sup>1</sup>. During CASP6, we focused more directly on the question of conformational sampling, and whether, by augmentation of our earlier methods and algorithms by orders of magnitude more computational power, we could significantly improve our ability to predict protein structure. To achieve this objective we assembled a "structure prediction supercomputer" based on volunteered resources and a distributed computing platform using the world-wide-web in a project called Predictor@home.

#### **Protocol for Protein Structure Prediction**

Predictor@home approaches structure prediction through a multi-step pipeline that is similar to protocols that have led to successful prediction in the past<sup>1</sup>. In the first step of this pipeline, homology modeling and fold recognition templates are identified as significant hits from the BLAST and SAM-T02 servers. In addition, secondary structure is predicted by the PSIPRED server. The results from template recognition are used to generate restraints for aligned residues during lattice-based MFold simulations; untemplated regions are sampled by a Monte Carlo conformational search with the MONSSTER<sup>2</sup> force field using any available secondary structure information from PSIPRED. Secondary structure is the only information used to guide folding "new fold" prediction targets by MFold. In order to sample viable folded conformations, 5-10 thousand simulated annealing MFold tasks were distributed for each target, thereby increasing our sampling by 1-2.5 orders of magnitude over our past studies<sup>1</sup>. In the refinement step, each sampled structure is subjected to all-atom simulated annealing between 1000K and 300K using the molecular simulation package CHARMM and an intermediate accuracy all-atom force field. The lattice-based predictions provide inter-residue restraints implemented as NOE-like restraints based on side chain - side chain centers of mass

contacts. Minimization is performed in the presence of the GBMV<sup>3</sup> solvent model to produce the final structure and energy value to be used in scoring. Scoring and ranking proceed via hierarchical clustering of the all-atom results based on the side chain contact-map.

#### **The Architecture of Predictor@home**

Predictor@home is built on top of the Berkeley Open Infrastructure for Network Computing (BOINC)<sup>4</sup>. BOINC is a well-known desktop grid framework that provides built-in support for distributed computing on heterogeneous PCs connected to Internet or Intranet networks. It currently supports a wide range of PC platforms (i.e., Linux, Windows, Mac, and Solaris). Protein structure prediction was achieved through two computationally intensive phases accomplished by two different codes:

1. MFold for protein structure assembly based on a low-resolution modeling method that uses a lattice representation;
2. CHARMM for protein refinement with an all-atom modeling method.

Predictor@home is a client-server based parallel computation paradigm. For each target, the server continuously generates MFold and CHARMM workunits (independent computations on a given target). The results from MFold are redirected by the server to CHARMM. Clients apply for computation and receive several workunits at a time. Client failures may occur and the returned results may be affected by hardware malfunctions or malicious attacks. Predictor@home addresses the integrity of the returned result using replicated computing and homogeneous redundancy (redundant instances of a computation are dispatched to numerically identical computers).

Over the course of the CASP6 season, we sampled over 430 thousand protein structures for 65 targets, each validated as the result of at least three replicas. In total nearly 7 thousand users registered for Predictor@home, with over 14 thousand machines.

1. Feig, M. & Brooks III, C. L. (2002). Evaluating CASP4 Predictions With Physical Energy Functions. *Proteins* **49**, 232-245.
2. Skolnick, J., Kolinski, A. & Ortiz, A. R. (1997). MONSSTER: A Method for Folding Globular Proteins with a Small Number of Distance Restraints. *J. Mol. Biol* **265**, 217-241.
3. Lee, M. S., Feig, M., Salsbury, F. R., Jr. & Brooks, C. L., III. (2003). New analytic approximation to the standard molecular volume definition and its application to generalized Born calculations. *J. Comput. Chem.* **24**, 1348-1356.
4. Anderson, D. P. (2003). BOINC: Berkeley Open Infrastructure for Network Computing. [http:// boinc.berkeley.edu](http://boinc.berkeley.edu).