494e Distance-Dependent Force Field Using a High Resolution Decoy Set

Rohit Rajgaria, Scott R. McAllister, and Christodoulos A. Floudas

With the explosive growth in the number of sequenced genomes, the demand for fast and reliable protein structure prediction techniques has increased tremendously. A very important component of the protein structure prediction approach is to devise an energy function that is capable of recognizing the native structure among myriads of similar low energy, non-native structures. Ideally, this energy function should include every type of interaction present in a detailed atomic model of a protein. However, such a rigorous approach increases the computational cost enormously. An alternative approach is to use a discrete, distance-dependent force field. These force fields have been used by various researchers and have shown tremendous potential ([1],[2],[3],[4],[5],[6],[7]).

As we go from medium resolution structure prediction to high resolution structure prediction we need a force field which can distinguish between structures with very low RMSDs. This requires using a diverse and high quality decoy set to generate the force field. The generation of decoy structures is split into two stages. The first stage identifies the hydrophobic core of the protein and uses a set of tolerances to establish a varying degree of protein flexibility within the bounds. Then an ensemble of decoy structures can be created via DYANA ([8]) which uses simulated annealing with torsion angle dynamics. Using this method, high resolution decoys (i.e. decoy structures with 0.0 A < RMSD < 2.0 A to the native structure) have been generated for a set of 1400 non-homologous proteins that are expected to span the experimentally-determined structures in the Protein Data Bank.

A linear programming based approach ([7]) is used to train these decoys. To incorporate a large number of decoys (> 625000) an iterative dropping scheme based on RMSD and energy criteria was used. Decoys for each protein were ranked based on RMSD and then a force field was generated using the 45 lowest RMSD decoys of each protein. This force field was then used to rank all decoys and the best decoys (ranked by energy) were used to generate the next force field. This iterative scheme was repeated until there was no more change in the ranking of the decoys.

The force field was then tested on a different set of 75000 low RMSD decoys and has shown very promising results. A test set comprising of 500 decoys of 150 proteins was used. This high resolution force field was capable of correctly identifying the native structure of 109 test proteins out of 150. The average ranking and average Zscore for this test set was 1.81 and 2.11 respectively. This force field was also tested on a set of medium resolution test decoys ([7]). It was observed that this force field does equally well on medium resolution decoys. When tested on this decoy set, it ranked native structures of 115 test proteins as rank 1 as compared to 93 in their published work. The average Zscore in this case also increased from 3.08 to 3.78. To further increase the effectiveness of this force field, additional work is being done to refine the bin structure and to include the effect of side chains by considering interaction between centroids of the side chain.

[1] Maiorov, V. N. and Crippen, G. M. (1992). Contact potential that recognizes the correct folding of globular proteins, Journal of Molecular Biology 227: 876-888.

[2] Vendruscolo, M. and Domany, E. (1998). Pairwise contact potentials are unsuitable for protein folding, Journal of Chemical Physics 109: 11101-11108.

[3] Miyazawa, S. and Jernigan, R. L. (1999). An empirical energy potential with a reference state for protein fold and sequence recognition, Proteins: Structure, Function, and Bioinformatics 36: 357-369.

[4] Liwo, A., Oldziej, S., Pincus, M. R., Wawak, R. J., Rackovsky, S. and Scheraga, H. A. (1997a). A united-residue force field for off-lattice protein structure simulations. I. Functional forms and parameters of long-range side-chain interaction potentials from protein crystal data, Journal of Computational Chemistry 18: 849-873.

[5] Tobi, D. and Elber, R. (2000). Distance-dependent, pair potential for protein folding: Results from linear optimization., Proteins: Structure, Function, and Bioinformatics 41: 40-46.

[6] Samudrala, R. and Moult, J. (1998). An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction, Journal of Molecular Biology 275: 895-916.

[7] Loose, C., Klepeis, J. L. and Floudas, C. A. (2004). A new pairwise folding potential based on improved decoy generation and side-chain packing, Proteins: Structure, Function, and Bioinformatics 54: 303-314.

[8] P. Güntert, C. Mumenthaler and K. Wüthrich. Torsion angle dynamics for NMR structure calculation with the new program DYANA. Journal of Molecular Biology, 273 (1997): 283-298.