Fuli Du
Dr. sc. hum.

Modeling Microbial and Other Protein Structures Applying Reduced Representations and the Genetic Algorithm

Born on 10, 09, 1966 in Shandong, P.R. China.
Doctor degree study in Biochemistry, from SS 1997 to WS 2000.
Graduate study in Applied Physics, 09/1989 - 03/1992, Tianjin University, China.
Undergraduate study in Applied Physics, 08/1985 - 07/1989, Tianjin University, China.

Promotionsfach: Biochemie (EMBL)
Doktorvater: Prof. Dr. Thomas Dandekar

The main purpose of this thesis is modeling and predicting bio-medically relevant protein structures using simplified conformational representations and a genetic algorithm search strategy. The simplified representations used in this thesis comprise a set of standard protein backbone conformations, which efficiently reduce the conformational search space, while maintaining high feasibility to model a real protein structure in acceptable approximations. The genetic algorithm has been shown to be efficient in searching the conformational space of a protein molecule when implemented with basic principles of protein structures. The main work in this thesis includes:

1. Model the electrostatic interaction between two charged particles in aqueous solution. It is generally accepted that the protein-water interactions including hydrophobic effect are a dominant driving force in protein folding; and electrostatic interaction constitutes a large contribution to the solvation energy. Therefore accurate computational methods for electrostatic energy are very essential for any free energy based molecular modeling. However, in most current models, the electrostatic interaction is over simplified with the representation of the electrostatic energy by a sum of pair-wise additive Coulomb type potentials. In this thesis we use a much more accurate biophysics model – the hard sphere model for two overlapping charged particles – to study the electrostatic interaction between two charged ions or groups in protein molecules. We found that although

analytical potentials could be derived, the challenges in mathematics make the simulation too complex and demand formidable computational resources. This is not practical in molecular simulations aiming at primary and medically meaningful structures.

2. Modeling proteins and domains involved in host-parasite interactions using standard conformations and the genetic algorithm. We have reconstructed the observed structures of proteins and protein domains with seven standard conformations and genetic algorithms. We show that the genetic algorithm is an efficient search technique in modeling protein structures. The seven standard conformations are flexible enough to reconstruct most protein structures with sufficient accuracy and thus are promising for structure prediction to achieve a much reduced search space. We also show that the systematic recombination technique proposed by Koenig and Dandekar improves the search efficiency not only in an HP model but also in this grid free conformational model. We have also studied the possibility to improve the modeling results with more and refined standard conformations proposed by other researchers and conclude that they do not show a strong improvement in the modeling simulation while tremendously increasing the computational efforts.

3. Predicting the main chain folding of small proteins. We first have implemented the basic protein structure principles – secondary structure, hydrophobic effect, van der Waals atomic clashes and hydrogen bond etc. as the fitness function of our genetic algorithm. Then the coefficient (weight) of each fitness term has been optimized through systematic simulations using several protein molecules with known structures. We have presented in this thesis the results of predicting the main chain folding of small proteins from sequence and secondary structure using the genetic algorithm.

4. Interplay between secondary and tertiary structure predictions. Secondary structures play an important role in determining the tertiary fold of a protein. Correct and accurate secondary structure assignment, either from experimental results or prediction methods, is important for tertiary fold prediction. In this thesis, we have studied the interplay between secondary and tertiary structure and how to correct insufficient secondary structure prediction for fold prediction. We have focused on the following questions: whether and to what extent (1) a refined combination of several secondary structure prediction

methods could correct mispredictions, (2) the mispredictions could be corrected during protein folding simulations applying a genetic algorithm.