Protein Structure Refinement by Optimization

Proteins are the main active elements of life whose chemical activities regulate cellular activities. A protein is characterized by having a sequence of amino acids and a three-dimensional structure. The three-dimensional structure has only been determined experimentally for 50000 of the seven million sequences that are known. Determining the protein structure from its sequence of amino acids is therefore a major problem in computational structural biology and is referred to as the protein folding problem. The folding problem is solved using de novo methods or comparative methods depending on whether the three-dimensional structure of a homologous sequence is known. Whether or not a protein model can be used for industrial purposes depends on the quality of the predicted structure. A model can be used to design a drug when the quality is high.

The overall goal of this project is to assess and improve the quality of a predicted structure. The starting point of this work is a technique called metric training where a knowledge-based protein potential, for a fixed set of native protein structures and a set of deformed decoys for each native structure, is designed to have native-decoy energy gaps that correlate maximally to a native-decoy distance. The main contribution of this thesis is methods developed for analyzing the performance of metrically trained knowledge-based potentials and for optimizing their performance while making them less dependent on the decoy set used to define them. We focus on using the gradient and the Hessian in the analysis and present a novel smooth solvation potential but otherwise the studied potential is kept close to standard coarse grained potentials.

We analyze the importance of the choice of metric both when used in metric training and when used in the evaluation of the performance of the resulting potential and find a significant improvement by using a metric based on intrinsic geometry. It is well-known that energy minimization of a potential that is efficient in ordering a fixed set of decoys need not bring the decoys closer to the native state. The next part of the work is focused on improving the convergence of decoy structures and we present a method that significantly improves the results of shorter energy minimizations of a metrically trained potential and discuss its limitations. In an ideal potential all near-native decoys will converge toward the native structure being at-least a local minimum of the potential. To address how far the current functional form of the potential is from an ideal potential we present two methods for finding the optimal metrically trained potential that simultaneous has a number of native structures as a local minimum. Our results generally indicate that a more fine-grained potential is needed to meet desired model accuracies but even with our coarse-grained model we obtain good results and there is an unexplored possibility to combine it with comparative modeling.

To allow fast energy minimization in Matlab a new set of more sparse formulas to calculate the first and second derivatives of a molecular potential is derived and implemented.