Kapitel 5

Abstract

Contact energy functions can be used for protein structure prediction. An important point when using such functions is how to compare different structures. Different distance and similarity criteria are compared regarding their ability to reproduce \( C_{\alpha}, C_{\alpha} \) distance distributions of proteins. The \( C_{\alpha} \) distance, which takes into account the internatomic distances, gives a very good description of protein structures. In order to apply this criterion the internatomic distances are used. A computational less demanding criterion is the overlap \( g \). This similarity relates the number of common contacts of two structures with the maximum number of contacts of the two structures. The contact distance \( D_{\text{contact}} \) which relates to the overlap by \( D_{\text{contact}} = 1 - g \), is inferior to the \( C_{\alpha} \) distance when reproducing the distance distributions, but much faster to apply. This property makes it suitable for the training of energy functions with large sets of structures, where a large number of comparisons has to be made.

Different methods for generating such functions are compared by looking at the following criteria:

- The ability to distinguish between native and non-native protein structures.
- The ability to recognize structures similar to the native one as being similar with respect to the overlap.
- The stability of native protein structures in Monte Carlo simulations. As low temperatures a native protein structure should remain native like in such a simulation.
- The calculation of native protein structures.

The following methods for generating energy functions are used:
A maximization of the Boltzmann-weighted overlap between alloy structures and experimental structures.

A linear optimization, in which a set of linear equations is solved.

A joint optimization method, in which the contact energy parameters are assigned by counting the different types of contact in native and non-native protein structures.

Several versions of the different methods are tested. Different parameter sets are applied to show the capability of the energy functions to assign native and non-native protein structures correctly. The performance of the functions when used together with methods for calculating native protein structures is tested. The following methods are applied to generate protein structures:

Threading: A very fast and effective method for generating structures. The secondary structure is generated (the target sequence) and the structure of a native protein as a combination is then identified. Using several target sequences together with a large number of native protein structures yields a high number of such decoys.

For this work a computational method which takes into account that the number of decoys for different target sequences differs in general is most useful in assigning such structures native structures correctly. For example, it is most successful in being transferable: the training with a very small set of structures yields an energy function which is successful in assigning structures of much larger sets correctly. Furthermore, it is unique in the training of the function only within the most desirable structures. For the small sets of structures 90% of the decoys can be excluded from the learning procedure. The inclusion of using all structures remains the same, whereas the error.

Monte Carlo Simulations: The structures from the Monte Carlo trajectory can be used as decoys for the training of the energy functions. For example, the native structure of a given sequence can be used as starting point for a such a simulation. Structures over a wide range of similarity can be generated in this way by varying the temperature of the simulation and the total energy function. In this work folding simulations carried out with energy functions trained in this way do not give better results than folding simulations carried out using energy functions trained with threading structures.

Permutations: Monte Carlo simulations are used for predicting native protein structures. This is done using different types of energy functions. When doing a folding simulation on the native protein structure with an energy function optimized without inclusion
this provides a structure with an overlap of 0.56 and a RMSD of 6.6 Å can be obtained. Therefore the simulation ends up in a state not being native but having stabilization in the native structure.

The energy function is in the simulation contains one of (1) contact energy parameters (two fixed types of amino acid pair). There are several possibilities for extending this type of function. For example, one can distinguish between different distances along the sequence of the two amino acids in contact. With looking at the residues (and the energy parameter for the given types of amino acids is chosen with respect of the distance) $\rho_1$. For example, using threshold and two different distances (that means the number of contact energy parameters is doubled), improves the recognition for a set of 131 proteins (from which 82 are used as target sequences, as well as proteins which have to be recognized from 95% to 85%). For a set of 420 proteins with 140 target sequences the recognition is improved from 52% to 85%.