# 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 *power distance*, which takes into account the interatomic distances, gives a very good description of protein structures. In order to apply this criterion, the interatomic distances are needed. A computational less demanding criterion is the *overlap* $q$. This similarity relates the number of common contacts of two structures with the maximum number of contacts of the two structures. The contact distance $D_{\text{cont}}$ which relates to the overlap $q$ by $D_{\text{cont}} = 1 - q$ is inferior to the power 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. At 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 maximisation of the Boltzmann-weighted *overlap* between *decoy* structures and experimental structures.
- A linear optimization, in which a set of linear equations is solved.
- A quasi chemical method, in which the contact energy parameters are assigned by counting the different types of contact in native and non native protein structures.

Various versions of the different methods are tested. Different protein sets are applied to check 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 sequence for which structures are generated (the target sequence) and the structure of a native protein are combined to a new sequence/structure pair (a *decoy*). Using several target sequences together with a large number of native protein structures yields a high number of such decoys.
In this work a quasi chemical method which takes into account, that the number of decoys for different target sequences differs in general is most successful in assigning such native/non native structures correctly. This method is also most successful in being transferable: the training with a very small set of structures yields an energy function which is successful also in assigning structures of much larger sets correctly. Furthermore it is enough to train the function only with the most dissimilar structures. For the used sets of structures 90% of the decoys can be excluded from the learning procedure. The recognition (when using all structures) remains the same or becomes even better.
**Monte Carlo Simulations**: The structures from a Monte Carlo trajectory can be used as decoys for the training of the energy function. 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 used 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.
Furthermore Monte Carlo simulations are used for predicting native protein structures. This is done using different types of energy functions. When doing a folding simulation of the 46 residue protein crambin with an energy function optimized without using

this protein a structure with an overlap of $q = 0.56$ and a cRMSD of 6.66Å can be obtained. Therefor the simulation ends up in a structure not being native but having similarities to the native structure.

The energy function in the simplest form consists out of 210 contact energy parameters (one for each type 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. When looking at the residues $i$ and $j$ the energy parameter for the given types of amino acids is chosen with respect of the distance $|j-i|$. For example using threading and two different distances (what means the number of contact energy parameters is doubled) improves the recognition for a set of 135 proteins (from which 82 are used as target sequences, so as proteins which have to be recognized) from 70% to 85%. For a set of 420 proteins (with 186 target sequences) the recognition is improved from 52% to 65%.