

Enhancing gapless threading

Tobias Schmidt-Goenner¹

tsg@chemie.fu-berlin.de

Ernst Walter Knapp¹

knapp@chemie.fu-berlin.de

¹ Institute of chemistry, Free University Berlin, Berlin, Germany.

Introduction

Threading is a well known method to generate decoys. Decoys are computer generated artificial conformations of proteins that possess characteristics of the native structure of a protein sequence. The primary use of decoys is to test scoring and energy functions. In 2001 Bastolla presented a simple energy function, which could recognize most native structures among decoys generated by gapless threading.[1] To make progress in this approach we need to enhance the optimization method. Here, we present methods to generate decoys, which are more similar to the native structures such that a refinement of the energy function is required.

Method and Results

Contact Maps: Contact maps are a simple representation of protein structures. A protein with N residues can be represented by a $N \times N$ matrix C . The elements C_{ij} are defined as: $C_{ij} = 1$ if i and j are in contact and $C_{ij} = 0$ otherwise. Two residues i and j are in contact, if the distance between the C_α -atoms is below a given threshold value.[3]

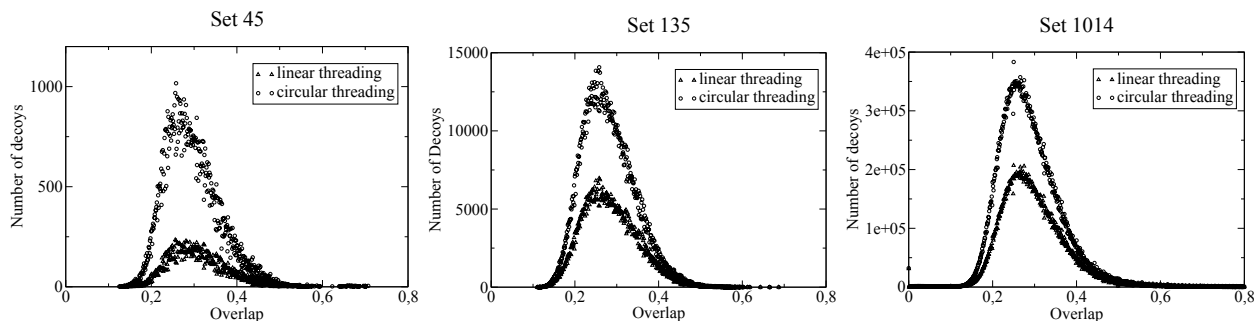
Overlap: To compare different contact maps of the same size, we define the *overlap* q as a measure of similarity. If C and D are 2 contact maps of the same size, the overlap is defined as:

$$q(C, D) = \frac{1}{Z} * \sum_{i,j} C_{i,j} D_{i,j}; \quad Z = \max \left[\sum_{i,j} C_{i,j}, \sum_{i,j} D_{i,j} \right] \quad (1)$$

Linear threading: Gapless threading is a method to produce a large number of artificial sequence-structure pairs (decoys). A decoy combines a given sequence of n_p residues with the native protein structure of an other sequence of the same length. This structure can be obtained by cutting a window out of the structure of a larger protein with n_d residues. ($n_d \geq n_p$). Moving this window along the structure results in $N_d = (n_d - n_p) + 1$ different decoys. Protein structures generated this way, are often not suitable to represent low-energy conformations for a reasonable energy function. But, as has been shown there is currently no simple energy function, that can guarantee recognition of the native state for all proteins simultaneously, if a large structural database is used for threading.

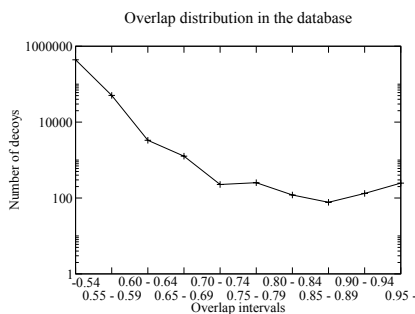
Datasets: Three different subsets of the PDB select database of protein structures with less than 25% sequence homology, were used.[2] **Set A** contains 45 single chain proteins with less than 200 residues each. **Set B** contains 135 single chain proteins, 82 with chains that are shorter than 200 residues. **Set C** contains 1014 proteins, 202 are single chain proteins of less than 200 residues. The sets are embedded according to **Set A** \subset **Set B** \subset **Set C**. For threading, sequences were taken from single chain proteins of less than 200 amino acids only, while all structures were used to generate decoys. Over 30,000 different decoys can be obtained by gapless threading using **Set A** for structure generation. For **Set B** over 970,000 and for **Set C** over 24,000,000 decoys were generated.

Circular threading: Connecting C- and N-terminus of a protein on the level of contact-map, we are able to thread over the equivalent of a circularized protein. Threading a sequence against a sufficient large protein now gives not $(n_d - n_p + 1)$ decoys, but n_d decoys. It also allows to thread against it's own native structure, resulting in $n_d - 1$ new decoys.



Distribution of decoys generated by linear and circular threading, lower and upper datapoints respectively.

Database approach: Circular threading yields mostly decoys in the low overlap range. Hence, there is still a lack of decoys with large overlaps. To address this problem we now decided to use the whole PDB as structure database. This leads to the problem that the number of decoys increases too much to be handable. Therefore we used a method to preselect large overlaps.



These structures are stored in a database, which allows generation of decoys in the interesting range of large overlaps without considering too many structures of small overlap. Since overlaps below 0.5 can be easily obtained by threading inside the given sets, we took this value as lower limit. The resulting distribution of overlaps obtained with linear threading of sequences from **Set A** and structures from the whole PDB, starting at an overlap of 0.5 is shown in the figure on the left.

Conclusions

Circularizing allows to produce more decoys. But similar to the conventional threading, most decoys have overlaps in the range 0.2 to 0.4. Nevertheless there are interesting new decoys.

Even using the full PDB and preselecting structures with large overlaps, the number of decoys with an overlap larger than 0.8 is still very small. However the number of overlaps between 0.5 to 0.7 obtained by these methods is now hopefully large enough to require a refinement of the energy function.

References

- [1] Bastolla, U., Farwer, J., Knapp, E.W., and Vendruscolo, M. How to Guarantee Optimal Stability for most Representative Structures in Protein Data Bank, *Prot. Struct. Funct. Genet.*, 44:79–96, 2001.
- [2] Hobohm, U. and Sander, C. Enlarged representative set of protein structure, *Protein Science*, 3:522–524, 1994.
- [3] Vendruscolo, M., Kussel, E., and Domany, E., Recovery of protein struture from contact maps *Fold. Des.*, 2:295–306, 1997.