Preprint
UCRL-JC-153905

Fold Recognition Using Sequence Fingerprints of Protein Local Substructures

A.Kryshtafovych, T.R. Hvidsten, J. Komorowski, K. Fidelis

This article was submitted to
the IEEE Computer Society Bioinformatics Conference (CSB2003)
Stanford, CA, USA, August 11-14, 2003

U.S. Department of Energy

Lawrence
Livermore
National
Laboratory

June 4, 2003

Fold Recognition Using Sequence Fingerprints of Protein Local Substructures

Andriy Kryshtafovych[1], Torgeir R. Hvidsten[2], Jan Komorowski[3] and Krzysztof Fidelis[4]
[1, 4] *Lawrence Livermore National Laboratory, Livermore, CA, USA*
[2, 3] *The Linnaeus Centre for Bioinformatics, Uppsala University, Uppsala, Sweden*
E-mails: [1] andriy@llnl.gov, [2] Torgeir.Hvidsten@lcb.uu.se, [3] janko@lcb.uu.se, [4] fidelis@llnl.gov

Abstract

A protein local substructure (descriptor) is a set of several short non-overlapping fragments of the polypeptide chain. Each descriptor describes local environment of a particular residue and includes only those segments that are located in the proximity of this residue. Similar descriptors from the representative set of proteins were analyzed to reveal links between the substructures and sequences of their segments. Using detected sequence-based fingerprints specific geometrical conformations are assigned to new sequences. The ability of the approach to recognize correct SCOP folds was tested on 273 sequences from the 49 most popular folds. Good predictions were obtained in 85% of cases. No performance drop was observed with decreasing sequence similarity between target sequences and sequences from the training set of proteins.

1. Introduction

Modern structure prediction methods can consistently produce reliable structural models for protein sequences with more than 25% sequence identity to proteins with known structures. But even if no protein with significant similarity can be detected for the protein of interest, there is still a chance that it can be assembled with local blocks existing in structural archives.

We have developed the method of local descriptors of protein structure that detect common local structural environments in proteins and organize them into a limited number of shape similarity classes [1]. Representatives from these classes can be used as elementary building blocks to reconstruct native protein structures or model unknown folds. Here we discuss application of the library of building blocks to fold recognition problem.

2. Descriptors and their similarity groups

A local descriptor of protein structure encompasses short segments of a protein chain that are located around the selected amino acid residue. To build a descriptor for a particular residue, we check distances between this residue and all other residues in the protein. The residues closer than 6.5 Å to the descriptor origin are added to the descriptor together with their four closest sequence neighbors. Assembled in such a manner, descriptors consist of several continuous segments, five or more residues long each (see Figure 1a). Number and length of fragments in the descriptor depend on local conformation of its backbone and packing of amino acid side chains.

We calculated descriptors for 4006 SCOP domains [2] from ASTRAL's [3] 40% sequence identity list (release 1.57). All individual descriptors were compared basing on the number of fragments, their length, shapes and packing schemes. If 7 or more descriptors from the dataset were found similar to some particular descriptor, we created a structural similarity group for that descriptor (Figure 1b). If two descriptors are very close structurally, their groups contain a big portion of the same descriptors. Joining overlapping groups into a bigger group we build a library of substructures representing relatively different local geometrical conformations from the considered dataset of proteins.

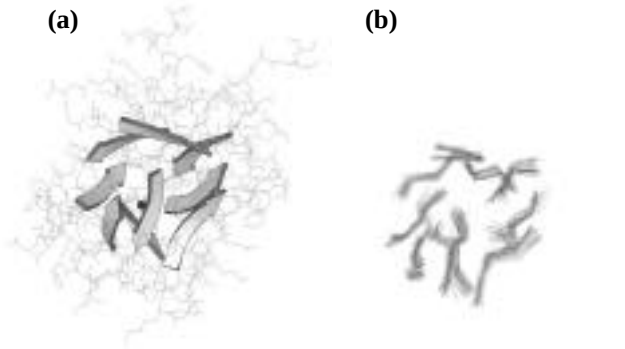

Figure 1. (a) Cartoons illustrate the descriptor for residue #164 from the fibroblast growth factor 9 (PDB code 1ihk, chain A); the descriptor's center is shown as a dark ball. (b) Structural superposition of 34 descriptors included into the group for descriptor 1ihka_#164. The images were created using visualization software [4].

3. Fold recognition

Having structurally corresponding segments of descriptors in the groups sequentially aligned as a result of comparison procedure, we extract a sequence-derived signal from each descriptor group. Using these sequence fingerprints we try to assign groups to sequences from outside the training dataset and subsequently determine their fold using voting procedure.Probability estimates calculated by counting occurrences of sequence-based features are used as a basis for extracting signals. These probabilities are determined for each of the 258 features [6] in all positions associated with the location of the residue on the segment of the group. Using these values we assign a signal vector to each set of aligned segments from the group. The probabilities for insignificant features are set to 0. (The feature is considered to be significant for the particular group if its probability is very unlikely to be observed in random data, i.e. it falls outside the 99% confidence interval). The significant features are used to capture the uniqueness of the group and henceforth to discriminate the proteins containing the corresponding local structure from the proteins that don't. The match between the target sequence and the group is the sum of the optimal individual assignments of all signal vectors.

Given a method for extracting signal vectors from groups and for matching these vectors to a protein sequence, we optimize the discriminatory power of the signal by comparing matching scores for the proteins that are represented in a particular group and those that are not. A threshold minimizing the error rate is selected for each group so that only the scores above this threshold allow the group to be assigned to the sequence. A greedy boosting algorithm is used to extract several sets of signal vectors responsible for different sequence patterns of the segments in the group. A genetic algorithm is used in each boosting cycle to select a subset of the features that minimize the error rate for the optimal threshold.

To predict a fold for a sequence, all groups are matched to this sequence using their signal vectors. The groups with the scores higher then the corresponding acceptance threshold cast votes to the SCOP folds represented in them proportionally to the fold popularity. The folds that received votes are considered predictions with a certainty given by their normalized vote-fractions.

The approach was tested on 273 target proteins from the 49 most popular SCOP folds. The correct fold was among the five best predictions in 85% of cases. We also used this test set to compare our performance with the fold recognition results obtained through purely sequence-based methods. Not surprisingly, PSI-BLAST [6] performed slightly better in cases where a good sequence homologue could be found and worse otherwise, i.e. if sequence identity of the target sequence to the closest sequence from the training set was below 25% (see Figure 2). One can also notice that our fold recognition capability is insensitive to the sequence identity level. This fact shows that we are able to capture general, amino acid-independent properties of local structures.

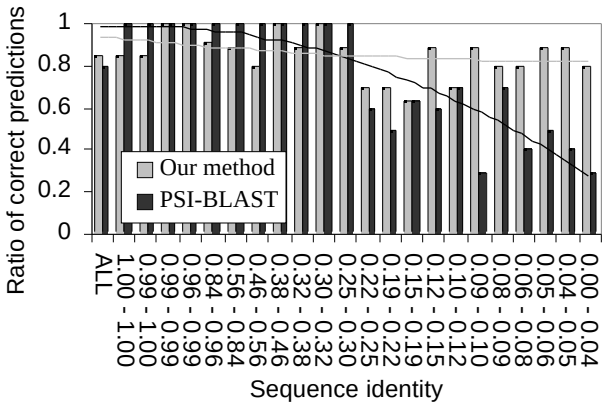

Figure 2. The fraction of test domains with the correct fold in the top five predictions distributed over different bins of sequence identity. The test set includes 273 domains that are present in ASTRAL 1.59 but not in ASTRAL 1.57.

This work was performed in part under the auspices of the U.S. Department of Energy by the University of California, Lawren-ce Livermore National Laboratory (contract W-7405-Eng-48).

4. References

[1] A. Kryshtafovych and K. Fidelis, "Local descriptors of protein structure. Part I. General approach and classification of local 3D regions in proteins," *In preparation*, 2003.

[2] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, "SCOP: a structural classification of proteins database for the investigation of sequences and structures," *J Mol Biol*, vol. 247, 1995, pp. 536-40.

[3] S. E. Brenner, P. Koehl, and M. Levitt, "The ASTRAL compendium for protein structure and sequence analysis," *Nucleic Acids Res*, vol. 28, 2000, pp. 254-256.

[4] P. J. Kraulis, "MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures," *Journal of Applied Crystallography*, vol. 24, 1991, pp. 946-950.

[5] K. Yu, "Theoretical determination of amino acid substitution groups based on qualitative physicochemical properties,"
http://cmgm.stanford.edu/biochem218/Projects%202001/Yu.pdf

[6] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Res*, vol. 25, 1997, pp. 3389-402.