

Fold Recognition Using Sequence Fingerprints of Protein Local Substructures

Andriy Kryshchak¹, Torgeir R. Hvidsten², Jan Komorowski³ and Krzysztof Fidelis⁴

^{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 substructure describes local environment of a particular residue and includes only those segments of the main chain that are located in the proximity of that residue. Similar descriptors from the representative set of proteins were analyzed to reveal links between the substructures and the sequences of their segments. Using the 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 from local substructures taken from libraries of known folds.

We have developed a method based on descriptors of protein structure to 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 an application of this library of building blocks to the fold recognition problem.

2. Descriptors and their similarity classes

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, each five or more residues long (see Figure 1a). The number and length of fragments in the descriptor depend on local conformation of its backbone and on the packing of amino acid side chains. We have 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 this 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. For the considered dataset of proteins, by joining such overlapping groups into larger groups, we have built a library of substructures representing relatively different local geometrical conformations (descriptor similarity classes).

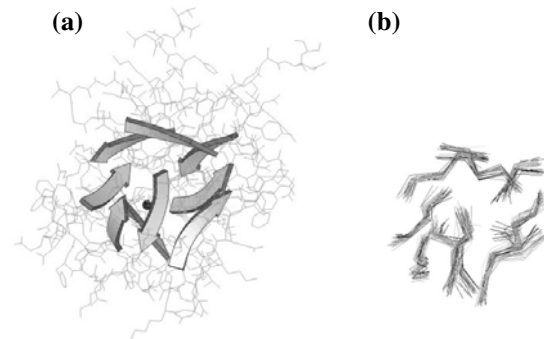


Figure 1. (a) Cartoons illustrate the descriptor for residue #164 from the fibroblast growth factor 9 (PDB code 1ihk, chain A); the descriptor's center is shown as a dark ball. (b) Structural superposition of 34 descriptors included in the similarity class for descriptor 1ihka_#164. The images were created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

3. Fold recognition

Having obtained the descriptor shape similarity classes, we have generated sequence alignments for each of the descriptor segments and extracted a sequence-derived signal for each of the classes. Using these sequence fingerprints we have then tried assigning the similarity classes to sequences outside of the training set and subsequently determining their fold using a voting procedure.

Probability estimates calculated by counting occurrences of sequence-based features were used as a basis for extracting signals. These probabilities were determined for each of the 258 features [6] at all positions of each segment of the similarity class. Using these values we assigned a signal vector to each segment of the similarity class. The probabilities for insignificant features were set to 0. (The feature is considered to be significant for a particular class if its occurrence probability is very unlikely to be observed in random data, i.e. it falls outside of the 99% confidence interval). The significant features are used to capture the uniqueness of the local structure and henceforth to discriminate the proteins containing the corresponding local structure from the proteins that do not. The match between the target sequence and the similarity class is the sum of the optimal and non-conflicting individual assignments of signal vectors for all segments. Given a method for extracting signal vectors from similarity classes and for matching these vectors to a protein sequence, we have optimized the discriminatory power of the extracted signals by comparing scores obtained for proteins from the inside and outside of the similarity class. A threshold minimizing the error rate was selected for each similarity class so that only the scores above this threshold would allow the class to be assigned to the new sequence. A greedy boosting algorithm was used to extract several sets of signal vectors corresponding to different sequence patterns of the segments in the similarity class. A genetic algorithm was then 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 similarity classes were matched to this sequence using their signal vectors. The similarity classes with the scores higher than their acceptance threshold cast votes in favor of the corresponding SCOP folds. The folds that received votes were 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. PSI-BLAST [6] performed slightly better in cases where a good sequence homologue could be found and worse if the 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 sequence-related properties of local structures, rather than similarity based on amino acid identity alone.

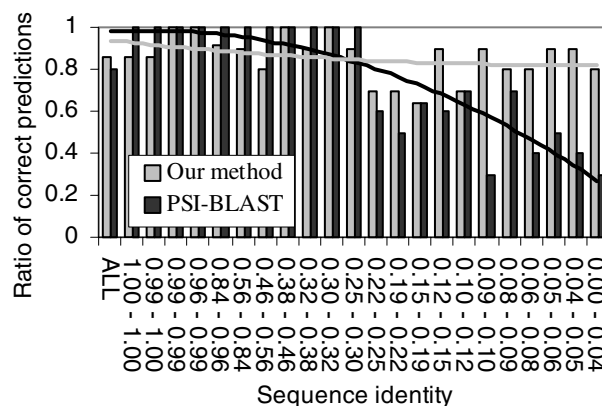


Figure 2. The fraction of test domains with the correct fold in the top five predictions distributed over different bins of sequence identity. The test set includes 273 domains that are present in ASTRAL 1.59 but not in ASTRAL 1.57.

This work was performed in part under the auspices of the U.S. Department of Energy by the University of California, Lawrence 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.