

Local Descriptors' Library Models the Core of Protein Structures Accurately

Michał Drabikowski^{1 2}, Krzysztof Fidelis³, Andriy Kryshchak³, Jerzy Tiuryn¹

Keywords: protein structure prediction, local structure, structure representation, structure modeling, local descriptors of protein structure

1 Introduction.

The so-called fragment-based methods have a wide variety of applications in analyzing and predicting protein structure ([1]). It was shown that a relatively small library of short backbone fragments (up to 7 residues) allows to accurately model native protein structures ([5], [4]). Here, we present a novel approach based on the library of multi-fragment structure motifs (*local descriptors*) designed to emphasize the long-range contacts between amino acids. To demonstrate the feasibility of our local motif formalism for a wide spectrum of applications, ranging from structure comparison and analysis to prediction, it is critical to show its ability to accurately reproduce protein structure from fragments. We show that our library describes all local 3D structure patterns occurring in the core of proteins and present an algorithm allowing for constructing accurate global 3D structures using these motifs.

2 Methods.

Fundamental to our approach is the concept of *local descriptors of protein structure*, which are units encompassing short continuous backbone *segments* (consisting of at least 5 residues) that are close in 3D space but not necessarily along the protein sequence. We use the ASTRAL 1.63 database ([2]) to select domains having less than 40% sequence identity to one another. For each residue from each considered domain a descriptor is constructed as a set of segments which are located close to the selected residue (a detailed description of descriptor construction may be found in [3]). For further analysis, out of more than one million descriptors constructed in such a way, by removing redundant similarity we select a set of 1753 structurally different representative descriptors. All of these descriptors consist of at least 15 residues and are intended to be taken from the proteins' core, which is more conserved than loops.

In order to explore the limits of our approach we need to know how accurately the set of representative descriptors can model structures of proteins. For any query protein structure we: (1) detect all structurally correct assignments of the representative descriptors (an *assignment* of descriptor d to protein s is defined as an association of all segments belonging to d with non-overlapping fragments of s), which give us a set of the best local-fit approximations, (2) detect a self-consistent set of structurally correct assignments producing the best global-fit approximation.

While the quality of a local-fit approximation t is measured by its RMSD to the query structure s (approximation is said to be correct if $RMSD(s, t) \leq 2.5$), for a global-fit approximation t we introduce a quality measure defined as follows: $score(s, t) = 100 \frac{|t|}{|s|} 2^{\frac{-RMSD(s, t)^2}{16}}$,

¹Institute of Informatics, Warsaw University, Warszawa, Poland

²Corresponding author, e-mail: m.drabikowski@mimuw.edu.pl

³Lawrence Livermore National Laboratory, Livermore, California, USA

where $\frac{|t|}{|s|}$ denotes a fraction of predicted residues (*coverage*). Unfortunately, the problem of constructing a global-fit approximation maximizing *score* is NP-complete. We therefore use a greedy algorithm for finding a good rather than the best model. The main idea of this algorithm is to choose the best subsets (in the sense of *score*) containing two assignments, then attempt to extend these subsets by adding one assignment in all possible ways, choose the best subsets containing three assignments, and so on.

3 Results and Conclusion.

We test the method on a set of all domain structures from the CASP 6 experiment, which covers a wide range of well-known structures and is independent from our training set. We divide this test set into two subsets: the first one (A) consists of 11 domains, which are classified as *new fold*, the second (B) – consists of 89 remaining domains.

	local-fit approximations				global-fit approximations					
	#(assignments)		coverage		<i>score</i>		coverage		RMSD	
	avg	min	avg	min	avg	min	avg	min	avg	max
set A	5.73	0.65	92%	73%	59	36	74%	50%	2.29Å	2.95Å
set B	6.99	0.44	94%	74%	64	37	77%	44%	2.09Å	3.02Å

Table 1: Local-fit and global-fit approximations of testing domains.

Table 1 summarizes the results of running the procedures (1) and (2) on sets A and B. The local-fit approximation section shows the average (and the minimal) number of structurally correct assignments per domain per residue, and the average (and the minimal) coverage of domains by these assignments. The global-fit approximation section presents global results obtained by applying our heuristics as described above. The average RMSD varies from 2.29Å to 2.09Å over about 75% of domain residues (which corresponds to the core of domains and some loops). The obtained results, which are almost as good for set A as for set B, give us reason to expect that the local descriptor-based method of protein structure prediction will be effective also in the most difficult *new fold* category.

This work was supported by the Polish KBN grant 3 T11F 017 26 and U.S. NIH LM007085 (to KF).

References

- [1] Aloy, P., Stark, A., Hadley, C. and Russell R. B. 2003. Predictions Without Templates: New Folds, Secondary Structure, and Contacts in CASP5. *Proteins: Struct. Funct. Genet.* 53, 436–456
- [2] Brenner, S. E., Koehl, P., Levitt, M. 2000. The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res.* 28, 254–256
- [3] Hvidsten T. R. Kryshchovych, A., Komorowski, J. and Fidelis, K. 2003. A novel approach to fold recognition using sequence-derived properties from sets of structurally similar local fragments of proteins. *Bioinformatics.* 19, II81–II91.
- [4] Kolodny, R., Koehl, P., Guibas, L. and Levitt, M. 2002. Small Libraries of Protein Fragments Model Native Protein Structures Accurately. *J. Mol. Biol.* 323, 297–307.
- [5] Unger, R., Harel, D., Wherland, S. and Sussman, J. L. 1989. A 3D building blocks approach to analyzing and predicting structure of proteins. *Proteins: Struct. Funct. Genet.* 5, 355–373.