On the Effectiveness of Distances Measuring Protein Structure Similarity

Jakub Galgonek, David Hokzsa
Department of Software Engineering
Charles University in Prague, Faculty of Mathematics and Physics
Malostranské nám. 25, 18000 Prague, Czech Republic
Email: {jakub.galgonek, david.hoksza}@mff.cuni.cz

***Abstract*—Determining similarity between two protein structures is one of the most fundamental problems in contemporary structural bioinformatics. With the increasing complexity of the measures, their effectiveness increases as well. However, other important observables, such as the degree of metric properties fulfilment, could rather deteriorate than improve. In this paper we introduce an effective measure and study its degree of metric properties fulfilment.**

***Keywords*-Protein Structure, SCOP, TM-score**

I. Introduction

A protein is a polymer of amino acids. Parts of a nascent protein chain are self-organized into secondary structures (helices, sheets or loops). Finally, the whole protein is formed into a tertiary structure (called *structure* in the following text). The structure of a protein is described by coordinates of its amino acids in the Euclidean space.

Proteins secure many biological functions in living organisms and this function is grossly determined by their structure. Hence, structural similarity is utilized to asses function to a protein with known structure but unknown function.

II. Measures in General

Measure of protein structure similarity consists of three subsequent steps:

1) Finding correspondence between pair of aminoacids in respective proteins. A general algorithm determines similarities (based on local, or possibly global, properties) of all possible amino acid pairs and selects the optimal subset. Output of this step is a vector of amino acid pairs.
2) Transformation of one of the proteins to minimize the distance (see step 3) to the other protein. Since protein structure is not anchored in the Euclidean space, it is difficult to find such a transformation (shift and rotation) that minimizes mutual distance of the respective proteins. Output of this step is a vector of Euclidean distances between pairs of transformed amino acids.
3) Computing a distance for the vector from step 2.

III. Enhancing of Contemporary Solutions

In our paper we focus on enhancing similarity measures (step 2,3) given an alignment (step 1). The resulting alignment is based on Smith-Waterman dynamic programing (DP) algorithm [1] which employs DDPIn's [2] amino acids representation.

We evaluate various similarity measures (within the meaning of step 2 and 3), trying to determine the most effective one according to our alignment algorithm. Hence, we are searching for a measure being robust against small deviations in the alignment.

Measures to be compared are the resulting DP score, root mean square deviation (RMSD), normed RMSD, TM-score[1] [3], and our improved version of TM-score.

One of the improvements in the TM-score stems from reducing number of the initial states. Only the original alignment and states that have identical secondary structure type are considered. In this way we noticeably decrease the runtime whilst keeping the quality of the heuristics. Further modification incorporates iterative modification of the given alignment. There are two types of modifications. First, if there is a segment S_1 missaligned with segment S_2 by a constant value, then S_2 is moved to match S_1. Second, we extend (after the transformation step) the alignment if there are portions of structures that are near each other but not present in the alignment. These are added to the alignment.

The original TM-score uses a scale to normalize distances of amino acid pairs which is parametrized by the length of one of the proteins. Our final modification uses parametrization based on length of both of the proteins. In the experimental evaluation we show only results where minimum of lengths of the proteins is utilized for parametrization, since other ways of parametrization do not score as good.

The effectiveness of the measures (classification accuracy[2]) is evaluated against subset of SCOP database [4], including 4326 database proteins and 979 query proteins having low sequence similarity. Results presented in Tab. 1 demonstrate superiority of the improved TM-score.

[1]Originally, TM-score is dissimilarity measure having maximum value 1, but we transform it to similarity measure by subtracting the actual TM-score from 1.

[2]The percentage of correctly classified proteins according to SCOP superfamilies.

meassure	effectiveness
DP	23.08%
RMSD	75.18%
normed RMSD	88.86%
TM-score	93.36%
iTM-score	93.97%

Table I
EFFECTIVENESS OF MEASURES WITH OUR ALIGNMENT ALGORITHM.

meassure	dimension	T-error	BOF
iTM	131.2	0.000005%	96.8%
$iTM^{2.5}$	24.3	0.04%	58.1%
iTM^{3}	17.5	0.10%	44.5%
-log(1-iTM)	6.9	0.15%	44.4%

Table II
PROPERTIES OF MEASURES ON THE RANDOM SUBSPACE.

IV. METRIC PROPERTIES

The improved TM-score (*iTM*) shows highest effectiveness, hence here we study its fulfilment of metric properties.

iTM is reflexive but it is not non-negative and symmetric. To fulfil this properties we modify the measure as $max(iTM(p_i, p_j), iTM(p_j, p_i))$ (further we use this modified measure). However, effectiveness of this measure decreases to 93.05%. To acquire the former effectiveness, we use top N most similar protein for the symmetric *iTM* and re-sort them according to the original non-symmetric *iTM*. In such a way, the effectiveness increases to 94.38%.

To evaluate the degree of the fulfilment of the triangle inequality property, we calculate the T-error [5]. Moreover, we calculate the intrinsic dimension and the ball-overlap factor (BOF) [5], which evaluates suitability of the measure for metric indexing. All the quantities are calculated for 500 random proteins from the SCOP database.

Tab. 2 shows that the value of T-error is low but BOF is very high. To change this behavior, we try out several TV-modification [5] (various powers and logarithm). The measure with logarithm modification shows low T-error, BOF and intrinsic dimension. Hence, we investigate this modification more thoroughly. We calculate the qualities again, but for individual structural classes of proteins (see Tab. 3). It shows slight increase of T-error whereas keeping BOF low (with one exception).

Due to the observation of the increase of T-error, we repeat the experiment for some big folds and superfamilies of proteins (including even more structurally similar proteins). On these subsets, T-error can increase significantly (e.g., T-error 19.27% for proteins from superfamily 48726). This increase of T-error can be also observed (not so noticeably) for the measure without the logarithmic modification (for superfamily 48726 the T-error is 0.52%).

class ID	cardinality	dimension	T-error	BOF
46456	825	6.4	0.32%	27.7%
48724	952	6.8	0.60%	15.8%
51349	1115	7.8	0.22%	17.8%
53931	965	8.7	0.21%	26.3%
56572	86	8.3	0.36%	55.3%
56835	100	3.8	0.27%	38.8%
56992	283	8.1	0.76%	38.1%

Table III
PROPERTIES OF THE MEASURE WITH THE LOGARITHMIC MODIFICATION ON DIFFERENT CLASSES OF PROTEINS.

V. CONCLUSIONS

We introduced an effective measure and its symmetric version that holds the semimetric properties. Its degree of the triangle inequality property fulfilment is very good (on random sets of proteins), not so the BOF quality. The logarithmic modification can decrease BOF, but increase number of T-errors (rapidly for sets of structural very similar proteins). However, we believe (possibly with better TV-modification) it is suitable measure for metric indexing.

ACKNOWLEDGMENTS

This research has been supported in part by Czech Science Foundation (GAČR) project Nr. 201/09/0683.

REFERENCES

[1] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *Journal of Molecular Biology*, vol. 147, no. 1, pp. 195–197, March 1981. [Online]. Available: http://dx.doi.org/10.1016/0022-2836(81)90087-5

[2] D. Hoksza, "Ddpin - distance and density based protein indexing," in *CIBCB*. IEEE, 2009.

[3] Y. Zhang and J. Skolnick, "Scoring function for automated assessment of protein structure template quality," *Proteins: Structure, Function, and Bioinformatics*, vol. 57, no. 4, pp. 702–710, 2004. [Online]. Available: http://dx.doi.org/10.1002/prot.20264

[4] 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, no. 4, pp. 536–540, April 1995. [Online]. Available: http://dx.doi.org/10.1006/jmbi.1995.0159

[5] T. Skopal, "Unified framework for fast exact and approximate search in dissimilarity spaces," *ACM Trans. Database Syst.*, vol. 32, no. 4, p. 29, 2007.