On the Effectiveness of Distances Measuring Protein Structure Similarity

Jakub Galgonek and David Hoksza

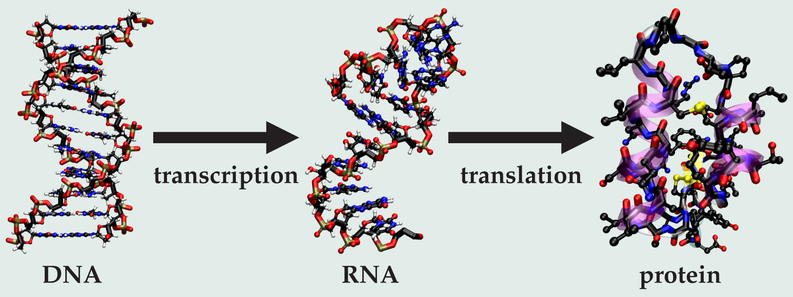
Department of Software Engineering, Charles University in Prague, Czech Republic

{jakub.galgonek, david.hoksza}@mff.cuni.cz

Motivation

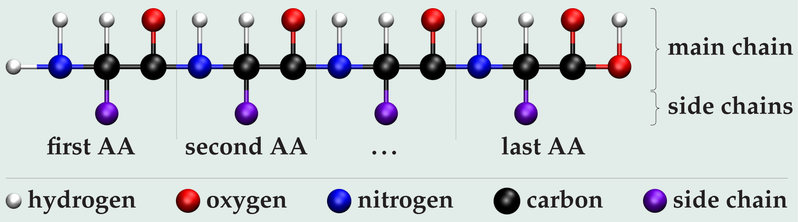
This poster presents one of the most often employed method in protein structure comparison (TM-score), its improvements and possibilities of metric indexing. We show the possible way of its semimetrization together with modifications and show how the various modifications fulfil metric qualities.

Central Dogma of Molecular Biology



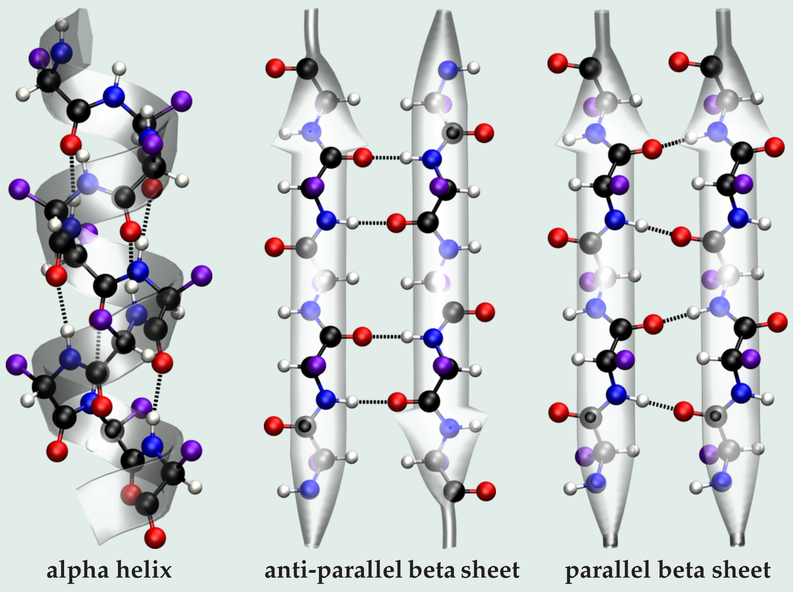
The origin of proteins can be depicted by the so-called central dogma. Genetic information is coded in DNA, transcribed into messenger RNA and finally translated into a protein.

Primary Structure



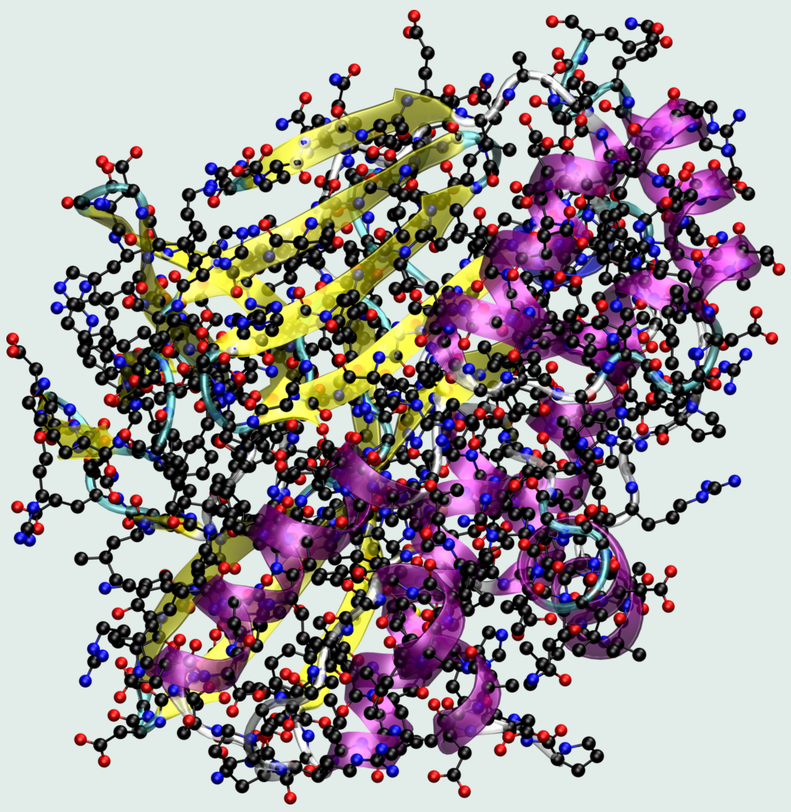
A protein is a sequence of amino acids linked by peptide bonds. This sequence is called the *primary structure*. A protein consists of twenty different amino acids all having identical main-chain parts but differing in their side chains.

Secondary Structure



Secondary structure refers to the three-dimensional form of regular local segments of a protein chain. The most ample variants employ the alpha helix and the parallel or anti-parallel beta sheets.

Tertiary Structure



Tertiary structure, expressed as the coordinates of individual atoms, describes three-dimensional structure of the whole protein.

Measures

Measuring similarity between protein structures consists in general of three subsequent steps:

1. Finding correspondence (alignment) between pairs of amino acids.
2. Determining the transformation (shift and rotation) of one of the proteins to minimize the mutual distance.
3. Computing the distance of the superposed structures based on the mutual positions of the aligned amino acids in the Euclidean space.

Step 1: Alignment

An algorithm determines similarities (based on local, or possibly global, properties) of amino acid pairs and selects a suitable subset.

We employ DDPIn [1] (a method based on pairing amino acids having similar density neighborhood) in combination with the Smith and Waterman dynamic programming algorithm [2] to obtain the alignment.

Step 2: Superposition

Since protein structure is not anchored in the Euclidean space, it is difficult to find such a spatial superposition minimizing the mutual distance of the respective proteins.

We employ following two algorithms:

- **minimizing the sum of the square of distances.** For transformation minimizing the sum of the square of distances (RMSD) we use a fast exact algorithm based on the linear algebra theory.
- **iterative search algorithm.** To minimize a distance for which an exact transformation algorithm is not known (TM-score [3]), a heuristic can be utilized.

  The heuristic algorithm that we use calculates the given distance (TM-score) for various subsets of the original pairing according to the superposition minimizing RMSD (the idea is that the optimal superposition will have some pairs near each other in the Euclidean space, hence their RMSD superposition will be nearly identical).

Input: Coordinate vectors Q and D of query and database proteins that define the input alignment; the length of the query protein L_Q
Output: The TM-score transformation
$d_0 = L_Q > 21$? $1.24 * \sqrt[3]{L_Q - 15} - 1.8 : 0.5$;
for *len = length(Q)* **shift right to** *4* **do**
  for *pos = 1* **to** *length(Q) - len + 1* **do**
    cut = {pos, . . . , pos + len - 1};
    for *i=1* **to** *ITERATION_COUNT* **do**
      T = rmsd_transformations(Q_{cut}, D_{cut});
      Q = transform(Q, T);
      score = 0; cut = ∅;
      for *i=1* **to** *length(Q)* **do**
        d = | Q_i - D_i |;
        if *d < limit(i,d_0)* **then**
          cut = cut ∪ {i};
        score += 1 / (1 + $(d/d_0)^2$);
      if *score > score_max* **then**
        T_max = T; score_max = score;
return *T_max*;

(the core of the TM-score transformation)

Step 3: Distance

As the resulting distance formula we use RMSD (root mean square deviation) and TM-score[a] (and its variant - see Improvements).

$$\text{RMSD} = \sqrt{\frac{1}{L_A}\sum_{i=1}^{L_A} d_i^2}$$

$$\text{TM-score} = 1 - \frac{1}{L_Q}\sum_{i=1}^{L_A}\frac{1}{1+\left(\frac{d_i}{d_0}\right)^2}$$

where L_Q is the length of the query protein, L_A is the length of the alignment, d_i is the distance between the ith pair of aligned amino acids and $d_0 = max(1.24\sqrt[3]{L_Q - 15} - 1.8, 0.5)$ is a scale to normalize the measure.

[a] Originally TM-score is a similarity measure, we modify it to form a distance.

Improvements

Several improvements of the measure based on TM-score have been proposed to obtain a more efficient measure:

- **reducing number of initial states.** Only the original alignment and subalignments consisting of pairs having an identical secondary structure type are considered. In this way we noticeably decrease the runtime whilst keeping the quality of the heuristic.
- **iterative modification of the given alignment.** After obtaining the optimal transformation, an alignment can be purified accordind to it. We apply two types of modifications. First, if there is a segment S_1 missaligned with a segment S_2 by a constant value, then S_2 is moved to match S_1. Second, we extend 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.
- **change the scale.** The original TM-score uses scale d_0 to normalize distances of amino acid pairs. Scale d_0 is parametrized by the length of the query protein. We parametrize d_0 using the lengths of both proteins (only the *min* parametrization is presented in the experimental section - other possibilities do not score as good).

SCOP

SCOP [5] is a manually curated hierarchical evolutionary classification, that was established as the gold standard for organizing protein structures. Proteins are stored in the leaves of the four-level hierarchy:

- **family** - high sequence similarity (>30%) or very similar function or structure
- **superfamily** - common evolutionary origin
- **fold** - same major secondary structures having similar topological distribution
- **class** - similar relative amount of types of secondary structures

Classification Accuracy

We evaluate effectiveness of a measure in the terms of the so-called *classification accuracy* - percentage of correctly classified proteins into superfamilies. The classification is assessed based on the "class" of the nearest protein in the database. The experiments have been carried out against a subset of the SCOP including 4326 database proteins and 979 query proteins having low sequence similarity.

meassure	effectiveness
DDPIn + DP[a]	23.08%
DDPIn + DP + normed RMSD	88.86%
DDPIn + DP + TM-score	93.36%
DDPIn + DP + iTM-score	93.97%

[a] Result of a dynamic programing based on DDPIn representation of protein structures.

Semimetrization

The improved TM-score does not hold the semimetric properties. Due to the identity of indiscernibles property we consider the following semimetric version of the improved TM-score.

$$iTM_{sm}(O_i, O_j) = max(iTM_{L_Q}(O_i, O_j), iTM_{L_Q}(O_j, O_i))$$

where iTM_{L_Q} is the improved TM-score with the standard parametrization of the d_0 scale.

However, effectiveness of this measure decreases to 93.05%. To acquire the former effectiveness, we use the top N most similar proteins for the semimetric *iTM* and re-sort them according to the original non-semimetric *iTM*. In such a way, the effectiveness increases to 94.38%.

BOF and T-error

To find out suitability of a measure for metric indexing we employed T-error and BOF (ball overlap factor) [4].

T-error is defined as the relative number of nontriangular triplets. Higher T-error values indicate higher non-metricity hence possible errors during filtration.

$$T\text{-}error(\mathbb{S}, \delta) = \binom{|\mathbb{S}|}{3}^{-1} \sum_{\{O_i, O_j, O_k\} \subset \mathbb{S}} T(O_i, O_j, O_k)$$

where $T(O_i, O_j, O_k)$ is 1 if the distances of the given objects are not triangular, otherwise 0.

BOF is defined as the relative number of overlapping pairs of the smallest non-empty balls. High BOF values indicate poor filtration ability.

$$BOF(\mathbb{S}, \delta) = \binom{|\mathbb{S}|}{2}^{-1} \sum_{\{O_i, O_j\} \subset \mathbb{S}} I(R_{NN}(O_i), R_{NN}(O_j))$$

where $R_{NN}(O)$ is a ball with center O and radius $\delta(O, 1NN(O))$. $I(R_i, R_j)$ equals 1 if regions R_i and R_j have a non empty intersection, otherwise 0.

Experimental Evaluation

T-error and BOF have been calculated for the semimetric version of the improved TM-score and its TV-modifications. As a dataset, we used 500 random proteins from the database used for computation of classification accuracy.

The logarithm modified measure shows low T-error and BOF. We calculate its qualities again for distinct structural classes. It shows slight increase of T-error whereas keeping BOF low (with one exception). On the other hand for internally very similar subspaces (e.g. superfamilies) the T-error deteriorates substantially.

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

class ID	card	T-error	BOF
46456	825	0.32%	27.7%
48724	952	0.60%	15.8%
51349	1115	0.22%	17.8%
53931	965	0.21%	26.3%
56572	86	0.36%	55.3%
56835	100	0.27%	38.8%
56992	283	0.76%	38.1%

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 increases number of T-errors (rapidly for sets of structural very similar proteins). However, we believe (possibly with a 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.

The figures have been generated by VMD [6].

References

[1] D. Hoksza, "DDPIn - Distance and Density Based Protein Indexing," in *CIBCB*. IEEE, 2009.

[2] 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.

[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.

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

[5] 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.

[6] W. Humphrey, A. Dalke, and K. Schulten, "VMD: Visual molecular dynamics," *Journal of Molecular Graphics*, vol. 14, no. 1, pp. 33–38, February 1996.