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 fulfill 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 DPDDP[2] (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 settings 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 protein database that define the alignment; the length of the query protein LQ.
Output: The TM-score transformation iTM = 1 + T, where L = 27.7% and 28.3%.

Step 3: Distance

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

RMSD = \sqrt{\frac{\sum_{i=1}^{LQ} (Q_i - D_i)^2}{LQ}}

TM-score = 1 - \frac{\sum_{i=1}^{LQ} |Q_i - D_i|}{LQ}

where LQ is the length of the query protein, LQ is the length of the alignment, \delta_Q is the distance between the \(i\)th pair of aligned amino acids and \delta_D = max(1,2|\delta_Q| - 7.5, 1.5) is a scale to normalize the measure.

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 subsequgments consisting of pairs having an identical secondary structure type are considered. In this way we noticeably decrease the runtime while keeping the quality of the heuristic.
- iterative modification of the given alignment. After obtaining the optimal transformation, an alignment can be purified according to it.

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
- superfamiliy - 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 797 query proteins having low sequence similarity.

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:

\[ iTM_e (Q, D) = \max (iTM(Q, D), O, O) \]

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.

\[ T-error(Q, D) = \frac{\sum_{i \neq j \neq k} |Q_i - Q_j| \neq |Q_j - Q_k|}{\sum_{i \neq j \neq k} |Q_i - Q_j| \neq |Q_j - Q_k|} \]

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

Conclusions

We introduced an effective measure and its symmetric version that holds the semimetric properties. Its degree of the triangle inequality property fulfillment 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 TM-modification) it is suitable measure for metric indexing.

References

[5] CIBCB (Center for Information and Knowledge-Based Computing) project Nr. 201/09/0683.