An Application of an Algorithm for Common Subgraph Detection for Comparison of Protein Molecules

Stoicho Stoichev, Dobrinka Petrova

Abstract: *The application of an algorithm for largest common subgraph detection for comparison of protein molecules is investigated. Proteins are presented with graph models, which are searched for their largest common subgraph. Heuristic algorithm is applied to speed up the comparison. Parameters of the model and the algorithm are examined and analysed to find their optimal values. Different definitions for similarity measures are tested to determine their consistency for comparison of protein molecules.*

Key words: *Protein Structure Comparison, Largest Common Subgraph Algorithm, Primary Structure Model, Heuristic Algorithm.*

INTRODUCTION

Comparison of protein molecules is employed in almost all branches of bioinformatics. Different methods for protein structure comparison are used for classification of protein structures in databases, protein structure modelling for protein structure prediction or structure-based functional analysis. The application of the method defines specific requirements for the three components, which compose it – model of the protein, comparison algorithm and similarity measure.

The model of the protein presents the protein molecule at the atomic level, when the goal is precise and detailed representation. Coordinates of the basic atoms, which compose the protein backbone - $C\alpha$ atoms, are used to construct the model. The requirement for precise and detailed model brings a lot of structural information, which complicates the representation and the task for comparison belongs to NP-complete class. The algorithms, which compare such models, often use different heuristics to reduce the complexity and to complete the job in reasonable time.

The models, which present the proteins at the level of secondary structure elements, are simple for construction and service. The number of entities in such model is smaller than the atomic level model, the complexity is lower and the comparison process is fast. However, these models are not suitable for precise comparison. They can be part of methods, which are applied as fast filters before the real comparison detection.

There is a third approach, which combine the previous discussed models. It uses fast comparison of representations at the level of secondary structure elements, which is followed by precise comparison at the atomic level. The low level comparison uses and refines the results, which are produced by the fast comparison at the higher level of representation.

A method, which follows the first approach for model construction is examined and analyzed in this paper. The protein structure is presented with a graph model. An algorithm for finding largest common subgraph is applied to compare the models of two proteins. Heuristic version of the algorithm is examined to reduce the complexity and to speed up the comparison. The number of vertices and edges in the resulting common graph defines the similarity measure.

PROTEIN STRUCTURE COMPARISON METHOD - OVERVIEW

Model of the protein

The model of the protein is proposed in [1]. Protein structure is examined at the low atomic level, where the coordinates of $C\alpha$ atoms are used to construct the model.
Basic properties of the model can be summarized as follows:

- The model of the protein is undirected graph.

- Every vertex in the graph model presents single amino acid with the coordinates of its Cα atoms.
- An edge between two vertices is defined if the amino acids, which are presented with these vertices, are spatial neighbors and the distance between them is within some threshold.

The distance threshold δ is defined as a parameter of the model. The distances between all pairs of amino acids are determined by calculating the Euclidean distances between their Cα atoms. Given a threshold value of δ, all of the amino acids, which have distances below this value, are spatial neighbours.
The distance threshold δ is one of the parameters, which are examined and analyzed. Different values for δ are tested to find the optimal solution.

Comparison algorithm

Graph models of the primary structure of compared proteins are constructed. An algorithm for finding largest common subgraph is applied to compare the models. Specific property of such model is the huge number of its vertices, which equals the number of amino acids in modeled protein. This number varies in the range [30..300], while the most common values are in the range [100..120].

The problem for largest common subgraph detection is NP-Complete task. The complexity of the task defines the requirement for finding and applying different heuristic algorithms for solving the problem. The need for reducing the complexity is evident, when the number of vertices in compared graphs is greater than 30, which is the lower limit for amino acids in a single protein molecule.

The algorithm for largest common subgraph detection, which is applied for comparing graph models of protein structures, is proposed in [2]. Approximate solutions are searched to speed up the comparison. The precision of the result is a parameter, which is examined and analyzed.

Similarity measure

Comparison algorithm produces the largest common subgraph of the two graph models. The number of vertices and the number of edges in the common subgraph are used to calculate the degree of similarity between compared proteins.

Let A and B are two protein molecules to be compared. Their graph models G_A and G_B are constructed, where $G_A = (V_A, E_A)$ is the graph model of the first protein and $G_B = (V_B, E_B)$ is the graph model of the second protein. The largest common subgraph $G = G_A \cap G_B$ is defined with its vertices V and edges E. A global similarity measure S_1 is proposed to evaluate the result – eq. (1).

$$S_1 = \frac{2V}{V_A + V_B} + \frac{2E}{E_A + E_B} \quad (1)$$

The number of vertices and the number of edges of the compared graphs and their common subgraph are taken into account, when the similarity is evaluated. The term 'global' means that the protein molecules are considered as a whole.

Three different definitions for local similarity scores are investigated to evaluate specific properties of the solution. Let $V_{\min} = \min(V_A, V_B)$ is the smaller of the numbers of vertices of compared graphs and $E_{\min} = \min(E_A, E_B)$ is the smaller of the numbers of edges of compared graphs. The ratio of the number of vertices in the common subgraph and $V_{\min}$ is defined with eq. (2).

$$r_V = \frac{V}{V_{\min}} \quad (2)$$

Eq. (3) gives the corresponding ratio of the number of edges:

$$r_E = \frac{E}{E_{\min}} \quad (3)$$

The sum of the scores, defined with eq. (2) and eq. (3) is used as a general local score S_2 - eq. (4).

$$S_2 = r_V + r_E \quad (4)$$

The forth definitions for similarity measures are examined in the tests to determine their consistency for comparison of biological macromolecules, such as proteins.

TESTING AND ANALYSIS OF THE METHOD

Test sets and parameter adjustment

Two methods – DALI [3] and CE [4], which present the structure of compared proteins at the atomic level, are used to prepare the test sets. Different proteins are selected for test set generators - mainly cytokines and signalling proteins. DALI is applied to determine the structural neighbours of a single protein, which is the test set generator. Groups of proteins are chosen among the structural neighbours of this test set generator according the degree of similarity and generate the test sets. DALI and CE evaluate the neighbours at the same time.

DALI is online available as a program for pairwise protein structure comparison called DaliLite on http://www.ebi.ac.uk/DaliLite/. http://cl.sdsc.edu/ can be used to find structural alignment between two protein molecules by CE method.

The structural information, necessary for comparison is extracted from PDB files [5], which contain atomic coordinates, primary and secondary structure information, stored in labeled records.

The model of the method is tested in advance for optimisation of its parameter – distance threshold δ. Different cases are examined and the results from the experiments can be summarized in following conclusions:

- δ < 5A. Distance threshold, smaller than 5 Angstroms is not suitable for constructing the graph model. The smaller the value for threshold is, the possibility for finding a spatial neighbor within this distance decreases. The graph model turns into a chain – every vertex is only connected with the vertices of the previous and the next amino acids in the primary structure of the protein.
- 5A < δ < 10A - the most appropriate value for the threshold is in this range.
- δ > 10A. Distance threshold, greater than 10 Angstroms is also not suitable for constructing the graph model. Increasing the value for threshold allows more vertices to be determined as spatial neighbors and the model looses its precision.

Further tests are made for distance threshold 5A < δ < 10A, which is shown to be the most appropriate range.

The huge number of vertices of the graph model defines the requirement of optimisation of the comparison algorithm, which belongs to NP-complete class. Essential part of the tests is dedicated to this problem. The algorithm for finding largest common subgraph has been developed to produce result with various precisions. Solutions with medium precision – 40-60% are tested for consistency.

The similarity measure between compared protein models is evaluated with all of the proposals – eq. (1) – eq. (4). Results are compared with the results, produced by DALI and CE.

Results

Results are produced for different signaling proteins and cytokines. The cytokine eotaxin is used for a test generator for the first test set. Part of this test set with its scores

for similarity, detected with DALI and CE is given in table 1. The last columns of the table show the results from the tests, which are done with the proposed method. The values of global similarity measure – S_1 between compared proteins are given for different distance threshold (8A and 10A) and precision of the solution (40%, 50% and 60% - only for 10A).

Table 1 Structural neighbours of eotaxin with their similarity scores

PDB name	DALI Z score	CE Z score	S1 (8A, 40%)	S1 (8A, 50%)	S1 (10A, 40%)	S1 (10A, 50%)	S1 (10A, 60%)
1g2t	13.2	5.3	1.228	1.330	0.993	1.215	1.452
1b3a_a	9.2	4.9	1.045	1.410	1.078	1.188	1.395
1eqt_a	9.2	4.9	0.945	1.282	1.038	1.279	1.507
1eqt_b	9.2	4.9	1.029	1.337	1.134	1.167	1.398
1zxt_b	9.3	4.9	0.997	1.267	0.965	1.204	1.380
1zxt_d	9.2	4.9	1.198	1.358	1.007	1.283	1.491
1zxt_c	9.1	4.7	0.982	1.278	1.068	1.418	1.532
1zxt_a	9.1	4.9	1	1.298	1.042	1.420	1.487
2ra4_a	9.3	4.9	0.956	1.248	0.964	1.180	1.423
2ra4_b	9.3	4.9	1.012	1.325	0.952	1.241	1.463
2q8r_e	9.2	4.9	0.931	1.179	0.923	1.197	1.432
1jqy_g	2	3.5	0.792	1.031	0.830	1.083	1.265
1jqy_n	2	3.5	0.808	1.030	0.809	1.029	1.293
1jqy_w	2	3.5	0.773	1.067	0.825	1.039	1.286

Similarity measures with different order of magnitude are applied for DALI, CE and proposed method. In order to compare these results they are presented with graphs in fig.1, fig. 2 and fig. 3. The comparison of the curves in these figures leads to the conclusion that the tested method determines correctly the similarity between compared protein molecules. The proteins, which are determined by DALI and CE to be closely related, are determined as similar by proposed method too. The opposite is also true – proteins, which are not so closely related, are distinguished by the three methods at the same way. The curve of the scores for global similarity measure, which are evaluated for distance threshold 8A and precision 40%, is the most similar with the curves of DALI and CE.

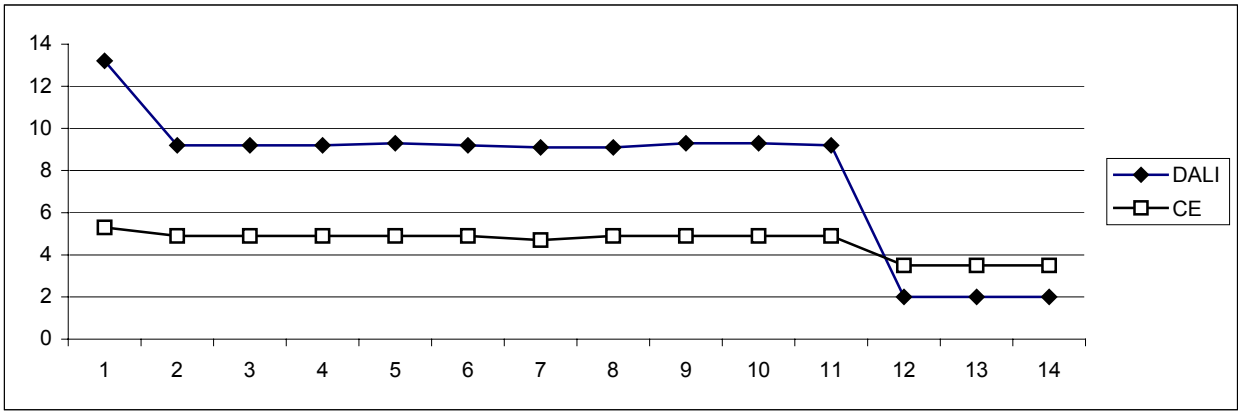

Fig. 1 Structural neighbours of eotaxin, evaluated by DALI and CE

The value of the global similarity measure increases with the growth of the precision of the solution. However the accuracy of the measure not always increases in these cases. There is an optimal value for the precision, which is connected with the value of the distance threshold. The optimal precision for 8A threshold is 40%, while the optimal

precision for 10A threshold is 60%.

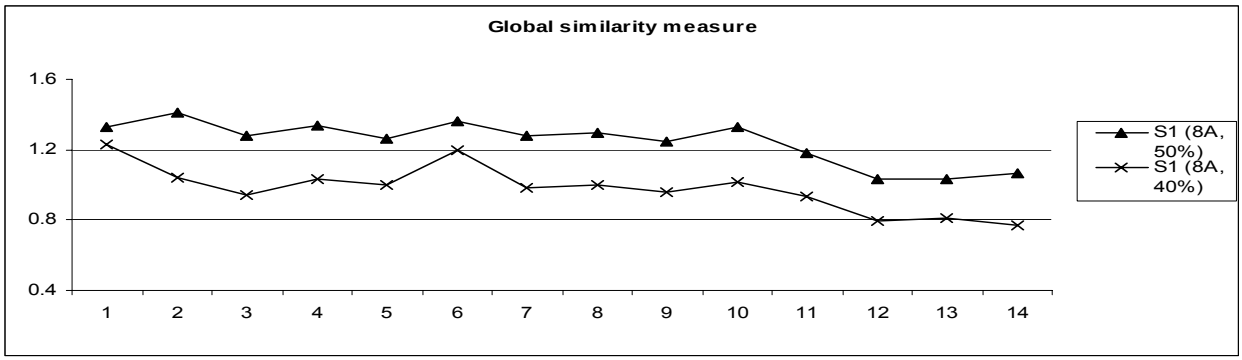

Fig. 2 Structural neighbours of eotaxin, evaluated by the global similarity measure of the tested method, 8A threshold

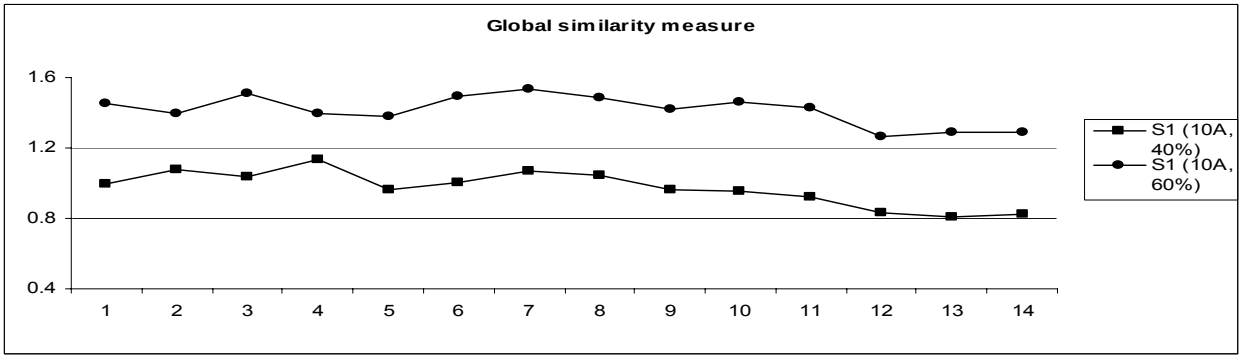

Fig. 3 Structural neighbours of eotaxin, evaluated by the global similarity measure of the tested method, 10A threshold

The values for the three local similarity measures - eq. (2) – eq. (4) for the optimal threshold and precision are given in table 2. Conclusions for the values in the table stand for all of the tested combinations of parameters threshold-precision.

Table 2 Structural neighbours of eotaxin, evaluated by local similarity measures

PDB name	**(8A, 40%)**			**(10A, 60%)**		
	r_V	r_E	S_2	r_V	r_E	S_2
1g2t	0.831	0.401	1.232	0.845	0.610	1.455
1b3a_a	0.672	0.407	1.078	0.836	0.603	1.439
1eqt_a	0.567	0.410	0.977	1	0.556	1.556
1eqt_b	0.657	0.408	1.064	0.836	0.605	1.441
1zxt_b	0.609	0.401	1.009	0.783	0.611	1.394
1zxt_d	0.812	0.402	1.214	0.986	0.523	1.508
1zxt_c	0.594	0.403	0.997	1	0.553	1.553
1zxt_a	0.609	0.403	1.011	0.956	0.553	1.510
2ra4_a	0.6	0.410	1.010	0.892	0.602	1.494
2ra4_b	0.662	0.404	1.065	0.938	0.600	1.539
2q8r_e	0.561	0.410	0.970	0.894	0.601	1.495
1jqy_g	0.408	0.404	0.812	0.680	0.590	1.270
1jqy_n	0.417	0.411	0.829	0.689	0.611	1.300
1jqy_w	0.388	0.408	0.796	0.689	0.599	1.288

The edge ratio r_E has almost constant value, which equals the value of the precision. In contrast the value of the ratio of vertices r_V varies and can be used to distinguish the

proteins according their local similarity score with the test set generator. r_V has a value of 1 for two proteins in table 2 – 1eqt_a and 1zxt_c, when the distance threshold is 10A and the precision of the comparison algorithm is 60%. This value means that the number of vertices of the resulting common subgraph equals the number of vertices of the smaller of compared graphs and it cannot be greater – i.e. the optimal solution is reached for precision of 60%.

This situation shows the application of the local similarity scores for determining and evaluating specific properties of the solution in contrast with the global similarity score, where the molecules are considered as a whole.

r_V is an essential part of the sum of local scores - S_2, because of the constant value of r_E. The values of S_2 are similar to the values of the global similarity measure - S_1 and this is the reason why S_2 and r_V can also be qualified as consistent for similarity measures.

The results from other test sets support the conclusions, which are made for the test set of eotaxin.

CONCLUSIONS AND FUTURE WORK

A method for protein structure comparison is examined and analysed. It presents the protein molecule at the atomic level with graph model, which defines the accuracy and the precision of the representation. The heuristic algorithm for finding largest common subgraph of compared proteins, which is applied, speeds up the comparison of the models. Different similarity measures for global and local similarities are investigated and compared with the results from other protein structure comparison methods to determine their consistency for evaluating the similarity between biological macromolecules such as proteins. The results show that the tested method, which uses heuristic algorithm for largest common subgraph detection can be applied and works properly for comparison of protein structures.

REFERENCES

[1] Stoichev, S., D. Petrova. Protein Structure Models for Determining Protein Structure Similarity. In Proc. Of the International Conference CompSysTech'06, Veliko Tarnovo, Bulgaria., 2006, p. IIIB.10.1-IIIB.10.6.

[2] Stoichev, St., Z. Zahariev. A New Algorithm for Determining Maximum Common Subgraph, 17-th International Conference on Systems for Automation of Engineering and Research (SAER-2003), Varna, 2003.

[3]. Holm, L., C. Sander, Protein structure comparison by alignment of distance matrices, J Mol Biol., 1993 Sep 5;233 (1):123-38 8377180.

[4]. Shindyalov, I.N. P.E. Bourne. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Engineering, 1995, 11(9) p. 739-747.

[5]. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I., P.E. Bourne. The Protein Data Bank. Nucleic Acids Research, 2000, Vol. 28, p. 235-242.

ABOUT THE AUTHOR

Prof. Doctor of Technical Sciences Stoicho D. Stoichev, Department of Computer Systems, Technical University at Sofia, Phone: +359 965 33 85 , E-mail: stoi@tu-sofia.bg.

Assistant of Prof. Dobrinka Petrova, Department of Computer Systems, Technical University at Sofia, branch Plovdiv, Phone: +359 32 659 727, E-mail: d_petrova2000@yahoo.com.