Protein Structure Models for Determining Protein Structure Similarity

Stoicho Stoichev, Dobrinka Petrova

Abstract: *The sequence and structure of a great number of proteins are becoming increasingly available. It is desirable to explore mathematical tools for efficient extraction of information from such sources. The principles of graph theory are now being adopted to investigate protein structure and folding. Two protein structure models are presented, according to different level of interest, preferring accuracy or simplicity. Graphs are used to describe protein 3D structure and algorithms from graph theory can be used to handle them for determining protein structure similarity and classification.*

Key words: *protein structure models, protein structure similarity, classification, graph model.*

INTRODUCTION

Proteins are long chains of Amino Acids with one interesting characteristic – the unique folding, which is the fold with minimum level of energy. This conformation depends on specific conditions and their violation will change the fold and protein will lose its properties. This is the reason to suppose that protein's functions are strongly connected with their unique fold, which is a sequence of Secondary Structure Elements (SSE).

Many algorithms are based on alignment of secondary structure elements - helices and sheets to compare and classify all known protein structures. One of the primary goals of these structural alignment programs is to measure the level of structural similarity between all pairs of known protein structures. This data can provide several meaningful insights into the nature of protein structures and their functional mechanisms. Comparison of all structures against each other can show relationships, both functional and structural, between proteins that were previously not known to be related. In addition, structure based distance measures are critical to classifying structures into families that share similar folds or motifs. Identifying these shared structural motifs using structural alignment techniques can provide significant insight into the functional mechanisms of the protein family.

One of the important aspects in such a method is the appropriate protein structure representation. Some of these methods use algorithms from graph theory – graph is preferred in VAST [1], MWBM [3], Stoichev- Milusheva method [2].

VAST represents all pairs of secondary structure elements (one from each structure) that have the same type as nodes of a graph. Two nodes are connected by an edge if the distance and angle between the corresponding pairs of secondary structure elements from the two proteins are within some threshold. The graph therefore represents correspondences between pairs of secondary structure elements that have the same type, relative orientation, and connectivity. This correspondence graph is then searched to find the maximal subgraph such that every node in the subgraph is connected to every other node in the subgraph and is not contained in any larger subgraph with this property. This is referred to as clique detection in graph theory.

MWBM use bipartite graph matching technique, where secondary structure elements of protein structures A and B are represented as nodes in weighted bipartite graph. An edge is defined as degree of similarity between two nodes (SSE) each from different protein. Then the similarity is found by choosing such a set of edges, which has maximum weight.

An interesting representation is used in Stoichev-Milusheva method where graph of protein is defined as G = (V, E), where V are nodes representing the SSEs of the protein and E are edges representing the connections between the various SSEs in the molecule. The secondary structure is represented in the space of the graph by a point corresponding to the coordinates of its centroid. The method calculates the centroids of all secondary structures (on the basis of the Ca atoms that compose them) and compares their

coordinates. These two protein graphs are compared to see if they share common features by largest common subgraph detection algorithm.

NEW PROPOSALS FOR PROTEIN STRUCTURE MODELS

Previously discussed alignment algorithms evaluate distance between SSEs to be compared. In this situation one of the most important problems to solve is to define correctly the end points of that distance and the way of computing it as a value. Decisions for end points can be centroids[2] or other equivalent points from two elements, which satisfy some requirements. No matter which of those are used for presenting the SSEs all they have a little disadvantage– they present sequence of points with one point. When the sequence is a chain of Amino Acids, the problem is harder to solve, because the chain is randomly folded, due to the chemical characteristics of the protein as a molecule.

Two different protein structure models are proposed here as ways of protein structure representation. Both of them use graph and aim to overcome distance-computing problem. Approaches to solve this problem can be:

- to avoid distance-computing when model is constructed;
- to evaluate distance between Cα atoms of all Amino Acids which construct the chain, but not between SSEs as a whole.

The choice for one of those approaches is made according our level of interest, which can be defined as:

- **Low level** is the one at which constructed models aim to reduce the complexity in protein structure representation and avoid distance computing. These models can be used as fast filters to determine if two structures could be similar and for classification of proteins in families.
- **High level** – here the accuracy of the model is preferred to its simplicity. Models at high level of interest use second approach and could be applied to obtain more precise results in protein structure comparison.

Model at low level of interest

The rule, which governs here, states, that the sequence of secondary structure elements is sufficient to define fold, when at least one SSE is a sheet. Number of visits for each SSE is the one parameter in this model. Number of visits, greater than 1, can be reached for nodes, representing SSE sheet. The explanation for this situation is that sheets consist of strands. The strands in one sheet could be arranged non sequential.

When Amino Acids sequence 'starts' a sheet (i.e. makes a strand/strands), but 'runs' to another sheet or helix and then 'turns' back to this unfinished sheet, number of visits of the node, representing this sheet is greater than 1.

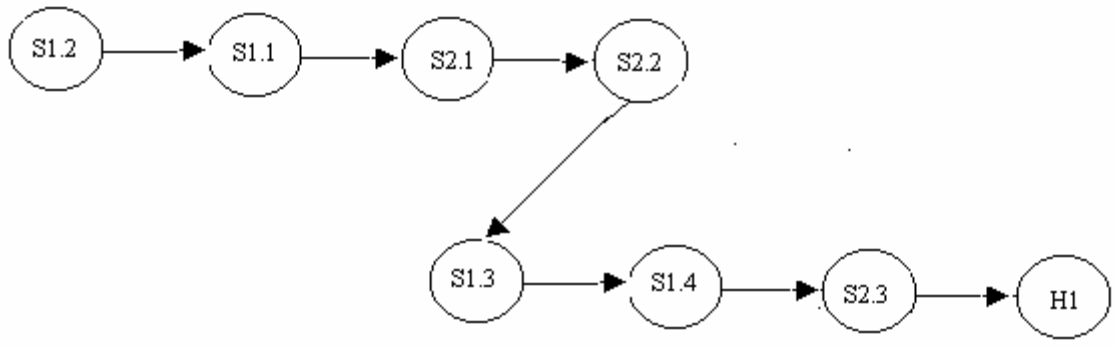


Fig. 1 Sequence of strands in sheets and helix in HIV Type 1

Fig.1 shows the sequence of sheets with their strands and helix in structure of Human Immunodeficiency Virus Type 1. Sequence of that protein starts with second and then first strand of the first sheet, followed by first and second strand of second sheet, next are third and forth strand of first sheet, third strand of second sheet and at the end is one helix.

The number of nodes in graph can be further reduced with additional rules:

- All strands in sheet are represented by a single node, holding number of the sheet;
- Node can represent a single helix or uninterrupted sequence of helices.

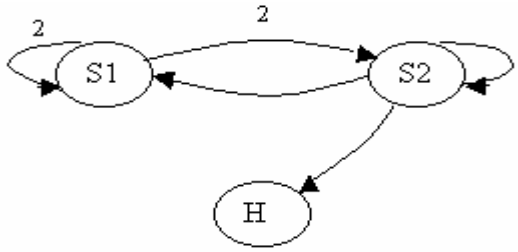


Fig. 2 Graph Model of HIV Type 1

Resulting graph model of Human Immunodeficiency Virus Type 1 is shown in Fig.2.
Number of helices: 1- H1;
Number of sheets: 2 – S1 and S2;
Numbers above each edge defines number of visits, if there is not a number – this edge is used only once.

Results from the model at low level of interest

Targets in this research are different virus proteins, toxins, oxygen-transport proteins, electron transport proteins and blood coagulation proteins. Different virus proteins and their graph models are presented here.

Starting point for the model is the information in PDB files, end- resulting graph with all rules for its construction. Resulting graphs are small and easy to compare, which is important, when the aim is fast comparison. This graph model has one significant advantage- allows protein from PDB to be checked for similarity with already constructed graph models without constructing the graph model of the new protein. In such case, when the purpose is only to conclude if a protein from PDB is similar to a protein with constructed graph model the sequence of SSEs from PDB file can test if it could be generated by walking through the graph model.

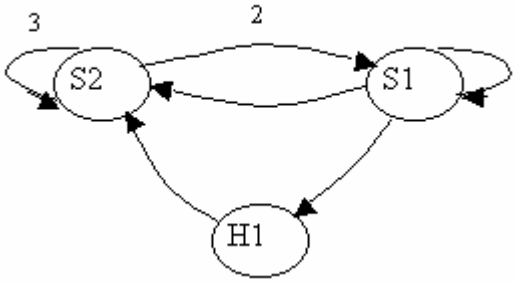


Fig.3 Graph model of Bence-Jones Immunoglobulin

BENCE-JONES IMMUNOGLOBULIN REI VARIABLE PORTION, T39K MUTANT
Type: VIRUS/VIRAL PROTEIN
Number of helices: 1- H1;
Number of sheets: 2 – S1 and S2;

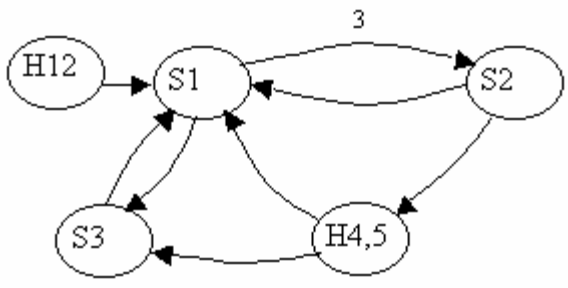


Fig.4 Graph model of Foot-And-Mouth Disease Virus

FOOT-AND-MOUTH DISEASE VIRUS/ OLIGOSACCHARIDE RECEPTOR
Type: VIRUS/VIRAL PROTEIN
Number of helices: 5- H123 and H45;
Number of sheet: 3 – S1, S2 and S3;
Rule 2 is applied here for low level of interest algorithm – helices H1, H2 and H3 compose uninterrupted chain of helices, so they are combined in one node.

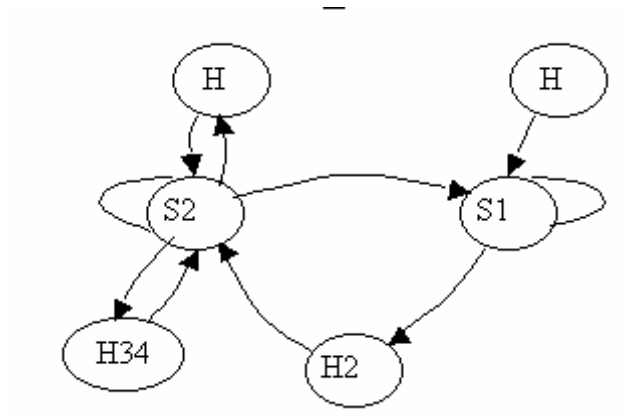


Fig. 5 Graph Model of HUMAN ADENOVIRUS SEROTYPE 2

HUMAN ADENOVIRUS SEROTYPE 2 FIBRE HEAD
Type: VIRUS/VIRAL PROTEIN
Number of helices: 5- H1, H2, H34 and H5;
Number of sheets: 2 – S1 and S2;
Rule 2 is applied here for low level of interest algorithm too – helices H3 and H4 compose uninterrupted chain of helices, so they are combined in one node.

Graph models, constructed for virus proteins (Fig. 2, Fig. 3, Fig. 4 and Fig. 5) can be compared, using algorithms from graph theory. Results from comparison of all known virus proteins could be suitable basis to find out some rules for their secondary structure composition and for structure-based functional analysis.

Model at high level of interest

Protein structure model at high level is used, when precise model is necessary. When the accuracy is important characteristic of the representation, evaluation of distances could be hardly ignored. To avoid inaccurate evaluation of distances between SSEs second approach is applied - SSEs are presented with all Amino Acids, which compose them. As a result in the model at high level of interest nodes are Amino Acids, presented with their Cα atoms. There are some different points of view to define the edges. The decision to make here is whether to connect sequence neighbors (Fig.6) Amino Acids or not and whether to take care of spatial neighbors, which belong to one SSE. After building and analyzing different models to limit the complexity sequence neighbors are excluded from rules, which define edges. Edges in the model are between spatial neighbors.

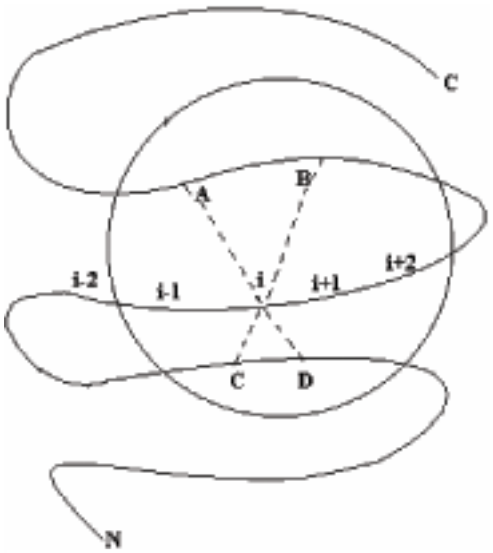


Fig.6 A schematic representation of the sequential and spatial neighbors of a Amino Acid i

A distance threshold is taken and Amino Acids with Cα, which fall within the sphere with threshold as a radius are spatial neighbors of Amino Acid i.
Edges in the model are defined according the rules:
There is an edge between two nodes if

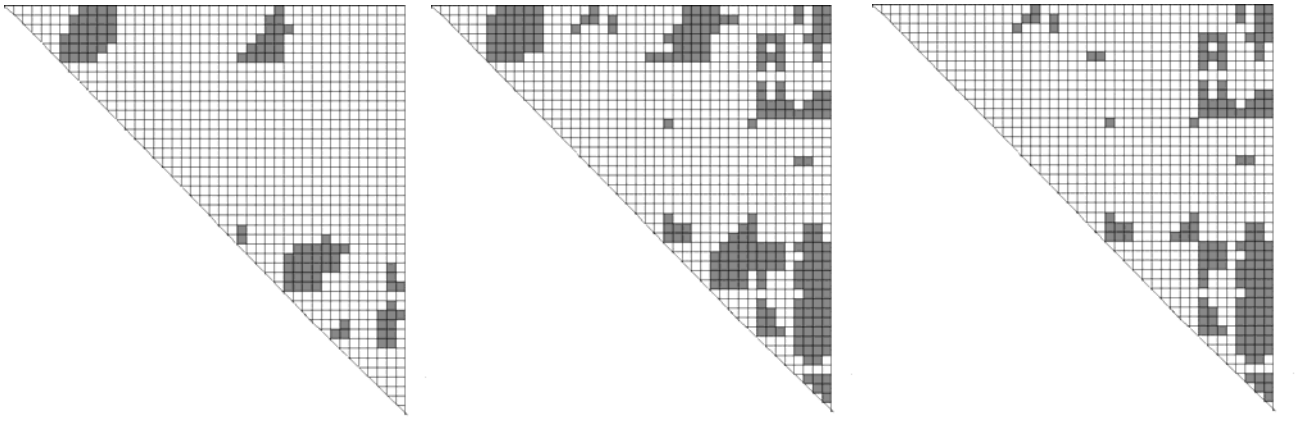
1. The nodes represent two Amino Acids from different structure elements; and
2. The nodes represent two Amino Acids, which are spatial neighbors, i.e. the distance between them is below given threshold.

Equation (1) is used to evaluate the distance between Cα atoms of Amino Acids x and y from different SSEs.

$$D(x,y)=\sqrt{\sum_{i=1}^{N}(x_i-y_i)^2} \qquad (1)$$

Fig. 7 shows graph models for HIV Type 1 resulting from different distance threshold and including or excluding rule 1 for edge definition. Spatial neighbors from different SSEs can't be captured with distance threshold 10A. Fig. 7(a) shows graph model with spatial neighbors within one SSE. When threshold is 15A graph model construction could include rule 1 from edge definition – Fig.7(c), or exclude it – Fig. 7(b). Resulting graphs for two

proteins can be compared to see if they share common features by graph isomorphism detection methods.



(a) distance– 10 A, rule 2 (b) distance– 15 A, rule 2 (c) distance– 15 A

Fig. 7 Adjacency matrices of graph models of HIV Type 1; (a) all spatial neighbors and distance 10 A; (b) all spatial neighbors and distance 15 A; (c) spatial neighbors from different SSEs and distance 15 A

Table1 shows results for graph models of HIV Type 1 constructed with MWBM, STOICHEV-MILUSHEVA method and current method- low level of interest and high level of interest. Numbers of nodes and numbers of edges are compared as elements in the resulting models. Number of nodes and number of edges in graph model, constructed by MWBM technique can't be defined without knowing the value of *n*, where *n* is the number of SSEs of the protein, which will be compared with HIV Type 1. This technique doesn't allow constructing a model for a single protein, but only for two proteins to be compared. If parameter $n = 0$, number of nodes and edges will define part of the model, that corresponds to representation of HIV Type 1. Stoichev- Milusheva method and current method at low level of interest produce similar results as number of nodes and edges, but current method has an advantage- it decreases the complexity in graph model construction by avoiding distance computing. Nodes and edges in model at high level of interest are many in number in accordance with high precision and accuracy of the model.

Table 1 - number of nodes and edges of graph model of HIV Type 1

	MWBM	STOICHEV-MILUSHEVA METHOD	CURRENT METHOD- LOW LEVEL OF INTEREST	HIGH LEVEL OF INTEREST- FIG 7(A)
Number of nodes	$3+n$	3	3	43
Number of edges	$3n$	Depends on precision	5	182

CONCLUSIONS AND FUTURE WORK

Two models of protein structure are presented, depending on level of interest. These representations are suitable for further work on constructing algorithms for protein structure similarity detection and classifying proteins in families, superfamilies and classes. Usage of light representation (model at low level) reduces the number of nodes and rules,

associated with edges, so the resulting graphs can be easy compared. The effect of applying the rules for reducing nodes of the model will be further analyzed. Model at low level of interest has as an advantage its simplicity, while model at high level of interest accentuate to accuracy of the result. Protein structure graph models at high level of interest will be analyzed for determining optimal distance threshold and appropriate rules for edge definition.

As for comparison with other methods, mentioned above, here is some statistics for number or nodes and edges for different methods. When parameter n = 0, values for number of nodes and edges are for part of the model, corresponding to a single protein.

Table 2 - number of nodes of graph models

	MWBG	STOICHEV/MILUSHEVA METHOD	CURRENT METHOD-LOW LEVEL
G1 (Fig.2)	$3+n$	3	3
G2 (Fig.3)	$3+n$	3	3
G3 (Fig.4)	$8+n$	8	5
G4 (Fig.5)	$7+n$	7	6

Table 3 - number of edges of graph models

	MBGM	STOICHEV/MILUSHEVA METHOD	CURRENT METHOD – LOW LEVEL
G1 (Fig.2)	$3n$	Depends on precision	5
G2 (Fig.3)	$3n$	Depends on precision	6
G3 (Fig.4)	$8n$	Depends on precision	8
G4 (Fig.5)	$7n$	Depends on precision	10

REFERENCES

[1] Gibrat, J.F., Madel, T., and Bryant, S.H. (1996) Surprising similarities in structure comparison. Curr. Opin. Struct. Biol., 6, 377-385.

[2] Stoichev.St., Milusheva,Il.- private communication

[3] Wang, Y., Makedon, F., Ford, J. A Bipartite Graph Matching Framework for Finding Correspondences between Structural Elements in Two Proteins. In Proceedings of the 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pages 2972-2975, San Francisco, California, 2004

ABOUT THE AUTHORS

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

PhD student Dobrinka Petrova, Department of Computer Systems, Technical University at Sofia, branch Plovdiv, Phone: +359 32 659 704, E-mail: d_petrova2000@yahoo.com