Journal of Theoretical Biology 260 (2009) 438–444

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Structural relatedness via flow networks in protein sequence space

Zakharia M. Frenkel [a,*], Zeev M. Frenkel [a], Edward N. Trifonov [a,b], Sagi Snir [a]

[a] Institute of Evolution, University of Haifa, Haifa 31905, Israel
[b] Division of Functional Genomics and Proteomics, Masaryk University, Brno, CZ 62500, Czech Republic

ARTICLE INFO

Article history:
Received 22 March 2009
Received in revised form
9 June 2009
Accepted 2 July 2009
Available online 8 July 2009

Keywords:
Sequence annotation
Network analysis
Protein sequence analysis
Protein structure prediction

ABSTRACT

A novel approach for evaluation of sequence relatedness via a network over the sequence space is presented. This relatedness is quantified by graph theoretical techniques. The graph is perceived as a flow network, and flow algorithms are applied. The number of independent pathways between nodes in the network is shown to reflect structural similarity of corresponding protein fragments. These results provide an appropriate parameter for quantitative estimation of such relatedness, as well as reliability of the prediction. They also demonstrate a new potential for sequence analysis and comparison by means of the flow network in the sequence space.

1. Introduction

One of the tasks of computational biology is in protein structure prediction on the basis of known 3D-structures of related proteins (homology modeling) (Kopp and Schwede, 2004). Recently developed methods for structural annotation of proteins, such as efforts in predicting protein subcellular localization (see, e.g. Huang and Li, 2004; Chou and Shen, 2008), membrane protein types (Shen and Chou, 2008; Yang et al., 2007) and regions (Shen and Chou, 2008), enzyme functional classes (Shen and Chou, 2007a), as well as signal peptides (Chou and Shen, 2007; Shen and Chou, 2007b), and many other approaches provide very useful information for both basic research and drug development (Chou, 2004).

The main difficulty of protein homology modeling is the high diversity of protein sequences. Often, structurally identical proteins have almost no sequence similarity. Most of the approaches for solving this problem rely upon finding a common general pattern or profile for a selected group of structurally and/or functionally related proteins. These efforts are based on the assumption that such common profiles exist, which is often not true.

An example of an alternative way is the *intermediate sequence search* (ISS) for detection and alignment of marginally similar pairs of protein sequences (Park et al., 1997; John and Sali, 2004). The main point of this approach is that when two proteins do not show significant sequence similarity, but both are related to a third protein, this relationship can be used to infer association between the pair under consideration.

Recently, this ISS strategy was substantially expanded by exploring two new ideas (Frenkel and Trifonov, 2007c). The first is the choice of small protein segments (about 20 aa) forming the *natural formatted sequence space* (or just *sequence space*), instead of full protein chains or domains. The second is the discovery that such protein segments can form prolonged '*walks*' in the natural sequence space. The 'walk' is a chain of sequence fragments, a path in a graph formed by these fragments where each element has high similarity to its neighbors in the path. These walks result in formation of connected components of different sizes and configurations (Frenkel and Trifonov, 2007b) in the graph (network). An example of such a walk in a network is demonstrated in Fig. 1a and b.

The idea of viewing the protein space as a network and employing network applications to it is not new. Initially, the 'neutral networks' in the protein sequence space were considered for sequences obtained from the lattice model simulations (Bornberg-Bauer, 1997; Bornberg-Bauer and Chan, 1999). In another study (Dokholyan et al., 2002; Dokholyan, 2005) a protein domain universe graph (PDUG) was defined. The third type of application in the network-based representation is the protein–protein interaction (PPI) networks (see Ciriello and Guerra, 2008 for recent review). In these networks the nodes correspond to proteins, and undirected edges represent physical interactions between them. A concept of 'Functional Flow', somewhat similar to the one used in this work, was introduced in these papers (Nabieva et al., 2005). In another paper with network flow the directed graph of sequences and structures of proteins from the Protein Data Bank was computed (Meyerguz et al., 2007). The flow

* Corresponding author.
E-mail address: zakharf@research.haifa.ac.il (Z.M. Frenkel).

0022-5193/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2009.07.004

in the 'structural' graph correlates with the respective number of matching genomic sequences. The strength of connections between the nodes was found to reflect structural similarity between the proteins.

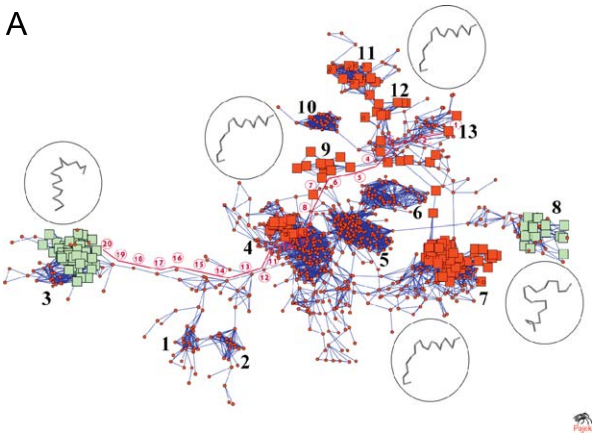
There are many indications that natural proteins have a modular organization (see Trifonov and Frenkel, 2009 for review). ISS (intermediate sequence search) that uses full-length proteins, built from several small modules, may relate absolutely different proteins. This false-positive prediction can be prevented by considering protein fragments less than a single module size, about 15–25 aa as was demonstrated earlier (Frenkel and Trifonov, 2007c). The consecutive small sequence changes along a walk can link fragments from different protein families and with completely different sequences (illustrated in Fig. 1c) but with the same structure and/or function (Frenkel and Trifonov, 2007a, c). The linkage of 'wrong' fragments is relatively a rare case (Frenkel and Trifonov, 2007a) (of course, on condition that all the intermediate fragments are from real natural proteins). The small number of independent connections (often only one) of such fragments to the network usually reveals this 'mistake', as we quantitatively prove in this work.

A series of papers with some relation to our work is 'kernel' papers (Eskin and Snir, 2005; Leslie et al., 2002, 2004). In these works, all *k*-mer words of a protein sequence, with a relatively small *k*, are used and their spectrum in constructed. Proteins are compared on the basis of similarity of the spectra. Subsequently, a support vector machine is used to classify protein families.

The presence of such structurally different fragments in one connected component can be of course interpreted as an accidental connection between two different components. It also can reflect natural structural variety due to interactions with other sequences in the protein, ligand associations, etc. For example, the module Aleph (*P*-loop) responsible for ATP-binding may exist in two structurally different forms depending on the presence of bound ADP (Sobolevsky and Trifonov, 2006). One of the modules of topoisomerases that is found in two alternative structures in the crystallized proteins (Sobolevsky and Trifonov, 2006) is another example.

In this work we demonstrate the relationship between structure similarity of protein fragments and the structural properties of a graph induced by these fragments. Specifically, we show that the connectivity between two vertices in the induced network definitely reflects the structural similarity of protein fragments corresponding to these nodes. Moreover, the level of connectivity also allows predicting such structural relatedness even for the protein fragments with essentially no sequence similarity. Importantly, the network considered here and the objects over which it is defined are principally different from the networks in the papers mentioned above. Here, vertices correspond to short protein fragments, and two vertices with prespecified sequence similarity are connected by an (undirected) edge.

A



B

Point Number	Sequence	Swiss-Prot Code	Position of fragment in protein
1	EDHLFRIYSMTKPLTSLACL	Q9RYU0	60
2	DDTIYRIYSMTKPLTSVAFM	Q89RH4	67
3	DDTCFRIGSITKTFTAVAVM	Q988N4	54
4	ADLRHRVGSITKTFTAAAVL	Q9F3F0	73
5	ADERFPVCSVFKTLAAAAVL	Q82F11	88
6	ENERFPMCSTFKFVLAAAVL	Q8P6D1	67
7	SNYGFPTGSTFKPFLAAAAL	Q8CJW2	401
8	ANSPLQTGSTFKIFGLAAAL	Q8FLP4	366
9	VWSTYEPGSTFKIITLAAAL	Q812W4	285
10	VSSIYEPGSVFKVITAAAAL	Q8R8S8	280
11	TSGLYPPGSIFKVVTAASAL	Q97GR5	205
12	TQGQAAPGSIFKVVSSAGAV	Q9L1G9	349
31	SQGQALEGAIFKVVDSKGKI	Q9CJ55	1663
14	IDGTTLEGAIFKIVDMDGKD	Q814R4	1737
15	VDGTALEGAVFKIVDANDEK	Q814R4	323
16	VDGTALKDAVFKITDLNGND	Q814R4	1362
17	DKKEMLEGAVFKITDLNGND	Q814R4	1550
18	GEKKPLEGAVFKLVDTNNND	Q814R4	419
19	SREKALEIAKTKLVDTNAND	Q8EX25	102
20	SRDKVREIATAKMKDLNAND	Q98N70	101

C

Point Number	Sequence	Swiss-Prot Code	Position of fragment in protein
1	EDHLFRIYSMTKPLTSLACL	Q9RYU0	60
	| | | | |		
9	VWSTYEPGSTFKIITLAAAL	Q812W4	285
1	EDHLFRIYSMTKPLTSLACL	Q9RYU0	60
	| | |		
20	SRDKVREIATAKMKDLNAND	Q98N70	101
9	VWSTYEPGSTFKIITLAAAL	Q812W4	285
	| | | |		
20	SRDKVREIATAKMKDLNAND	Q98N70	101

Fig. 1. An example of a walk within the network, containing several 3D structures. (A) Graphical illustration of the network. Squares appear only next to vertices whose sequences appear in PDB. Those vertices, for which their corresponding structure is similar (RMSD < 3 Å) to the most common structure of the network, are presented as red squares, while other as blue squares. The clusters in this figure form a ring (clusters 4, 9, 13, 7) with few or only a single bond connection between neighboring clusters (as between clusters 4 and 9). The clusters correspond to proteins with different sequences and functions. Nevertheless, they all correspond to the same 3D structure. Such result was expected, since every two clusters in the ring are connected by at least two disjoint paths. The clusters 11 and 12 do not belong to the ring, but as they are connected one to another and by several edges to the ring (13) it is not surprising, that they have the same 3D structure. Well-isolated clusters 3 and 8 are each connected to the central ring by a single path, and, indeed, their structures are different. (B) Description of the sequence fragments composing the 'walk' selected in this figure. Note moderate sequence dissimilarities ($\lesssim$40%) between consecutive nodes. (C) Sequence comparison of the initial, central and final fragments of the walk (1, 9 and 20, respectively). As can be seen the sequence similarity for each pair of the fragments are completely random and therefore cannot serve as a predictor for structure similarity. However, the first two fragments are connected by several disjoint paths and indeed share the same structure, whereas the last is connected by a single path and is structurally different. (For interpretation of the references to the color in this figure legend, the reader is referred to the web version of this article.)

Fig. 1a illustrates this idea of the relationship between structure similarity and the structural properties of a corresponding graph. Clusters of similar structures (red squares: 4, 7, 9, 11–13) are connected to one another through several different paths versus a single path link to clusters of different structure (green squares: 3, 8). (We informally denote them as *clusters* to indicate the group of nodes linked to one another rather to other nodes beyond the group.)

According to the above, although any two fragments in a connected component might be related, it is reasonable to assume that the higher the connectivity (Thomas et al., 2001) (see section Definitions), the higher the relatedness. In this work we quantitatively establish this relationship. We model the network as a flow network and apply flow algorithms to measure relatedness. In addition, once the relationship between connectivity and structure similarity is established, the aim was to refine the method. For a more realistic picture, the length of a path connecting two nodes should be also included into the score of relatedness. For this purpose another representation is proposed by modeling the network as an electrical conductance network. Using this model, more accurate results are obtained.

2. Methods

2.1. Definitions

A word (or a sequence) w is a sequence of letters taken from some alphabet Σ. We denote the length of w by $|w|$. A *k-mer* is a word of length k. We will use the term *match* to signify the event of an identical letter in both sequences at a certain position, and a mismatch is the opposite event. We use the normalized Hamming distance (Hamming 1950) between sequences as a metric over this space of k-mers, i.e., $H(s_1,s_2)$ is the relative (percentage) number of mismatches between the sequences s_1 and s_2.

Given a set W of sequences w_i, the k-spectrum $\Omega_k(W)$ is the set of all k-mers appearing at some sequence in W. For a threshold parameter $0\leq\tau\leq1$, we define the *threshold graph* over $\Omega_k(W)$, $G(W,k,\tau)=(V,E)$ where $V=\Omega_k(W)$ and $(u,v)\in E$ if the $H(u,v)$ is at most τ.

In this paper we focus on the number of independent, i.e., disjoint, paths between two vertices $u,v\in G(W,k,\tau)$. According to Menger's theorem (Ravindra et al., 1993), this equals the minimum number of edges required to remove from G such that u and v remain in two disconnected components. The latter is denoted as the *local edge connectivity* between two nodes u and v.

A directed graph $N=(V,E,c)$ is a *flow network* if every edge $(u,v)\in E$ is associated with a *capacity* function $c(u,v)\geq0$ and there are two distinguished vertices, a *source* s, and a *sink* t. A flow is a real valued function f: $E\rightarrow R$ satisfying the constraints that no edge's flow is greater than its capacity and for every $v\in V-\{s,t\}$, the sum of flow at the edges incoming v equals the sum of flow at the edges outgoing v. The flow value is defined as the total flow outgoing from s. The maximum-flow problem is to find a flow with maximum flow value in a given network.

The local edge connectivity problem in G can be solved by reducing it to a maximum flow problem in N in which $V(N)=V(G)$ and every edge in G is replaced by two opposite directed edges in N, each with a unit capacity (Goodaire and Parmenter, 2005).

A note on notations: Although sequence similarity in the setting of protein sequences takes into consideration the specific pair of aa, since in this work we operate only in the sequence space, we use the term *sequence similarity* only for the normalized Hamming distance.

2.2. Database preparation

Fig. 2 illustrates the process of preparing our network. Our ground set of sequences W was a set of about 320,000 protein sequences taken from 112 complete prokaryotic proteomes. Next, we constructed $\Omega_{20}(W)$, the set of all 20-mers appearing in some sequence in this set. Over this set $\Omega_{20}(W)$ of all 20-mers, we constructed the threshold graph $G(W,k,\tau)$ with $\tau=0.6$. We were interested only in connections between nodes in the same connected component. In this threshold graph, we selected about 61,000 components of magnitude from 11 to 2600 nodes.

All these 20-mers from the set $\Omega_{20}(W)$ were compared with all 20-mers from Astral database (Brenner et al., 2000; Chandonia et al., 2004) of proteins with known 3D-structure from PDB. A fragment with a close relative (having a distance of no more than 0.4, i.e., at least 12 matches) in this database was assigned with this relative's structure. When several such relatives for the same fragment were found, only the closest relative was selected.

2.3. Calculation of maximum flow

As described above, we solved the edge connectivity problem by reducing it to a maximum flow problem where edges have unit capacity. To calculate the flow in the induced network, we chose to use an efficient implementation (Cherkassky and Goldberg, 1997) of the Push-Relabel (Goldberg and Tarjan, 1988) algorithm for maximum flow. This implementation is very fast as it uses a combination of heuristics to speed up running times (we used the code taken from http://www.avglab.com/andrew/soft.html).

The maximum flow is analyzed only between the pairs of nodes such that each has a known structure and their sequences differ substantially from each other ($\leq$50% of similarity). If several nodes from the same connected component were related to the same fragment from PDB, only one representative node was selected for maximum flow calculation, which has the maximal number of connections.

The networks with $\tau>0.6$ are not analyzed because they contain much less non-similar structures. The distribution of structures in the network for which maximum flow was calculated is shown in Fig. 3. One axis shows the amount of pairs inside the interval of 0.1 Å of RMSD, which is shown on another axis.

2.4. Calculation of resistance

The calculation of the resistance between two points of the network was carried out as follows. The source node A and target node B are considered possessing potentials '0' and '1', respectively. For the network of N nodes and K edges (connections) a linear system of $K+N-2$ variables was composed: K-currents through the K edges, and $N-2$ potentials at remaining $N-2$ nodes. $N-2$ equations were introduced according to Kirchhoff's law, which states that the sum of all currents at every point (excluding two mentioned above) equals zero. Other K variables were obtained from Ohm's law $F_j-F_i=I_{ij}*R_{ij}$, where F_j and F_i are potentials at nodes i and j, and I_{ij} and R_{ij} are current and resistance between these nodes respectively. In this work we set uniformly all R_{ij} to '1' for simplicity (see Discussion). Thus, we have a linear heterogeneous system of $K+N-2$ equations with $K+N-2$ variables. From this system one can find a total current I, and total resistance will be $1/I$ (since the general difference between potentials is equal to '1').

The system was solved by Gauss's method. Before this, to save calculation time, the system was reduced to $(N-2)\times(N-2)$ by

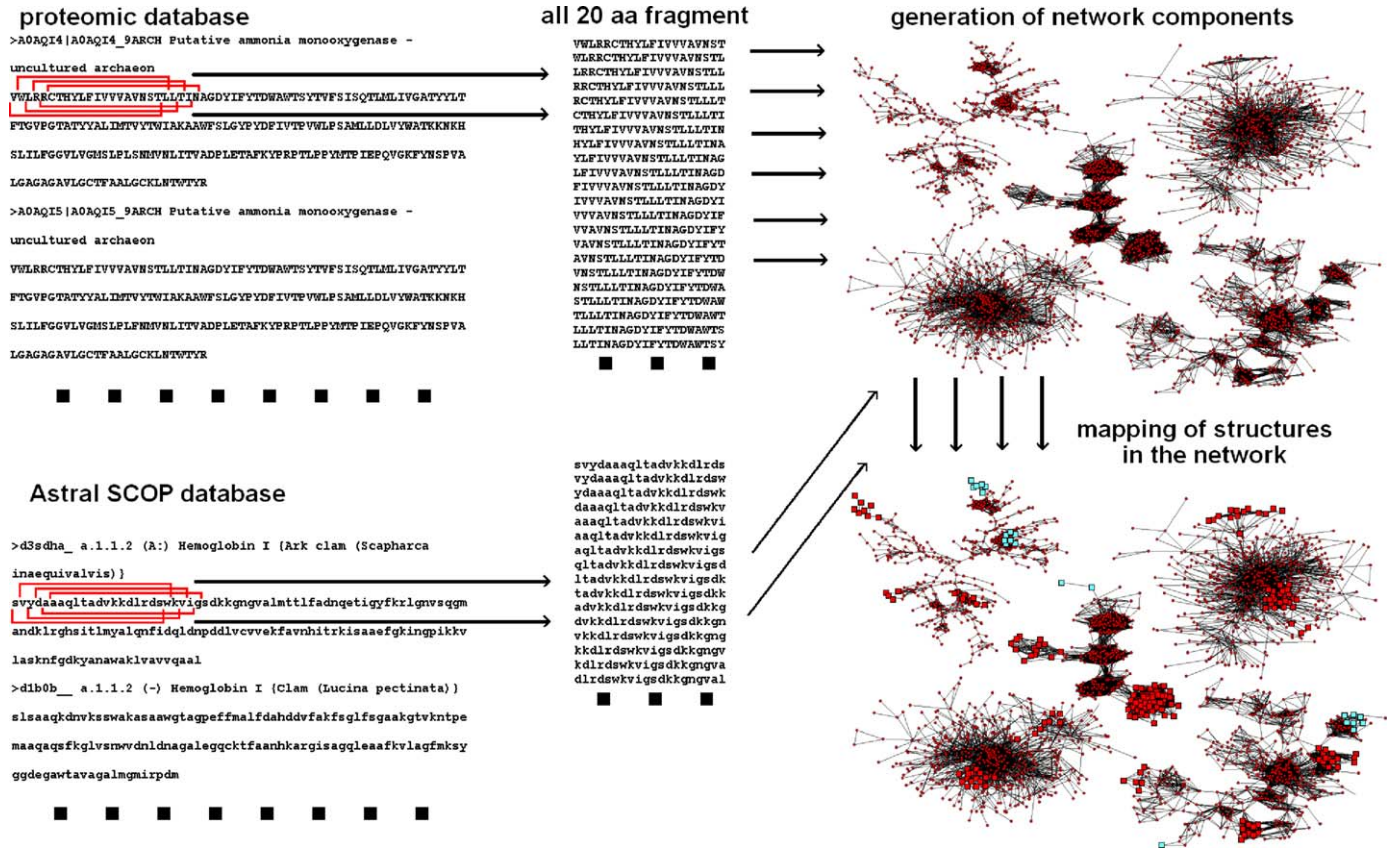


Fig. 2. Scheme of network connected component calculation for the study of flow between different protein fragments. A proteomic database was used for the network nodes production—all 20 aa fragments (with overlaps) were used. At the selected sequence similarity threshold, connections between the nodes were found and, consequently, connected component distribution of the network in the sequence space was produced. The nodes of selected components were compared against all 20 aa fragments from the protein database—Astral SCOP, for which the 3D structure of all fragments is known. Every node that has a similar fragment from Astral (also at selected threshold—60%) is associated with its respective protein structure (selected by the red or blue square, similar to Fig. 1). (For interpretation of the references to the color in this figure legend, the reader is referred to the web version of this article).

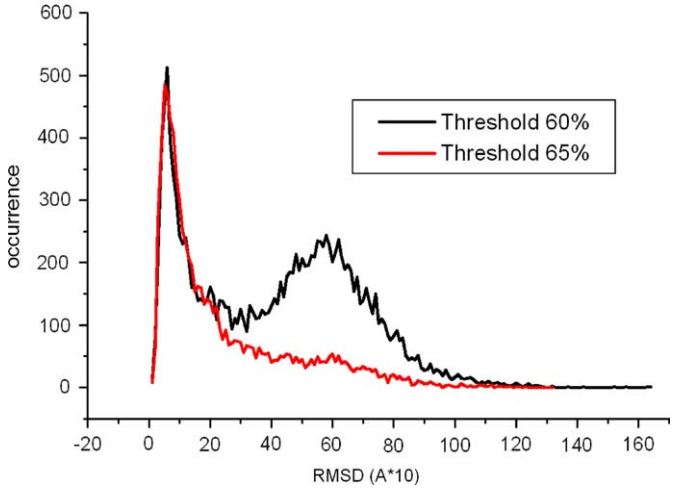


Fig. 3. Distributions of RMSD inside the networks of different thresholds. The 'occurrence' means amount of pairs inside the respective interval of 0.1 Å of RMSD.

assigning Kirchhoff's equations to determine potentials, with the currents taken from Ohm's equations: $I_{ij} = (F_j - F_i)/R_{ij}$.

2.5. Structural comparison of the protein fragments

The structural comparison of the protein fragments was carried out by the standard root-mean-squared-distance (RMSD) calculation between aligned alpha-carbon atoms. The procedure was taken from the available (http://www.cgl.ucsf.edu/Research/minrms/) Minrms program (Jewett et al., 2003).

For network visualization the standard program 'Pajek' (Batagelj and Mrvar, 2002) was used. Other operations: sequence comparison, network construction and analysis (e.g., Gauss's method realization) were carried out by our original programs in Microsoft Visual C++. Complete proteomes of 112 Eubacteria and Archaea were extracted from the SwissProt database, HAMAP project release 2003 (Gattiker et al., 2003). For sequence space construction of all 20 aa fragments (with overlapping) were taken (a total of about 10^8 fragments) from these proteomes.

3. Results

3.1. Maximum flow

For the connected components of the network described above, the flow values were calculated between the nodes, as described in 'Methods'. The relationship between flow values and RMSD is shown in Fig. 4. This curve shows average maximum flow between pairs of nodes with corresponding RMSD (rounded off to multiples of 0.1 Å). The numbers of pairs with a given RMSD value, i.e., on which the average flow was taken, is shown in Fig. 3. The distribution strongly demonstrates that a higher number of independent pathways (e.g., maximum flow) between fragments, indeed, indicates structural relatedness.

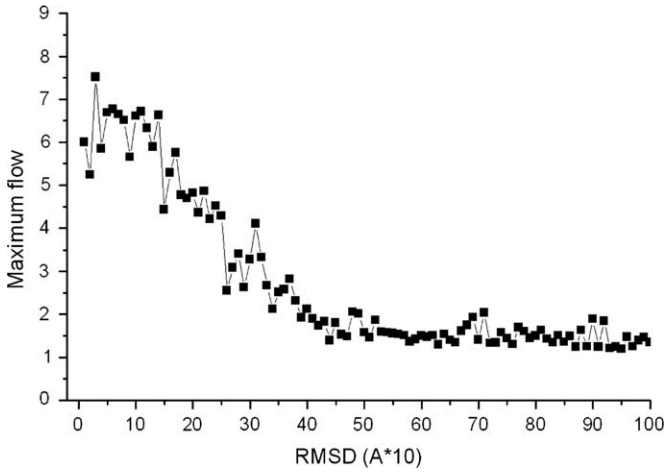


Fig. 4. Correlation of flow values between nodes with their RMSD. An exponential decay of the flow value with respect to the RMSD is observed.

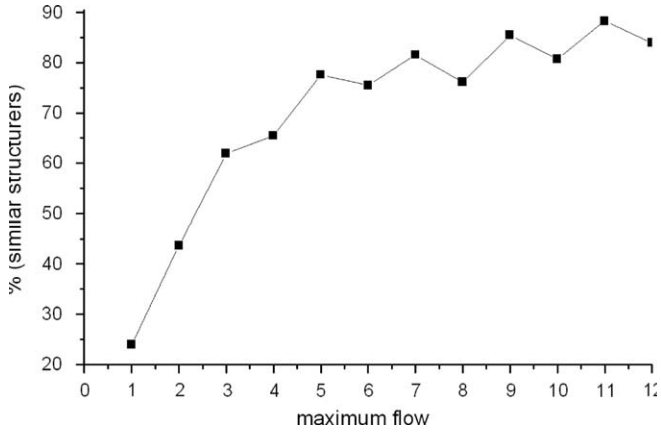


Fig. 5. Confidence of flow values as a function of structure similarity (RMSD < 3 Å).

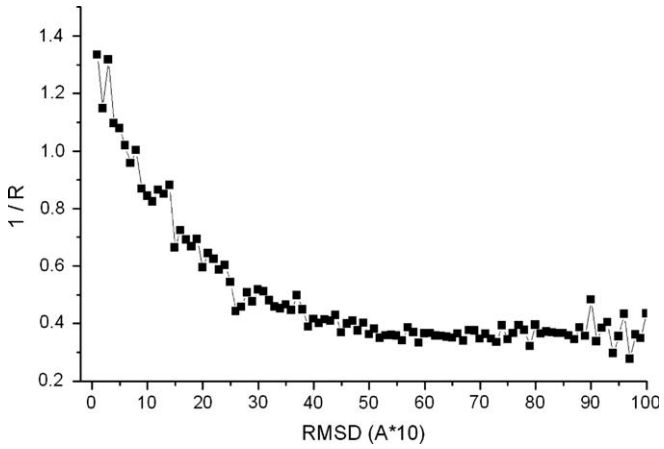


Fig. 6. Correlation of conductance values between nodes with their RMSD.

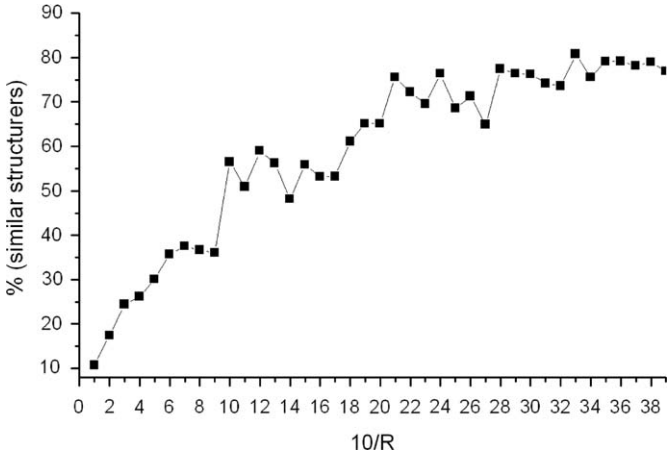


Fig. 7. Confidence of conductance values as a function of structure similarity (RMSD < 3 Å).

The curve shown in Fig. 5 displays the observed frequency of similar 3D-structure (RMSD < 3 Å) between two sequences, as a function of the flow between the corresponding nodes. Ie., for every max flow value, the number of node pairs with similar structure is indicated (percentage of total node pairs with that flow). As can be seen for flow values ≥ 2, i.e., two or more independent paths through the network, the frequency of structure similarity is more than 60%.

3.2. Electrical network

It is clear that the likelihood of nodes to share similar structure depends also on several other factors such as the length of pathways, degree of similarity between neighbors in the pathway, etc. The specific properties of the sequence similarity can be taken into account by introducing different radii of the 'conducting tubes', i.e., edges of the graph. The length of pathways can be considered by estimations made earlier (Frenkel and Trifonov, 2007a), in which the probability of two 20 aa protein fragments from different families but with sequence similarity of at least 60% to have similar structure is about 85%.

Hence, we have applied a more attractive though more complicated model simultaneously taking into account (in one parameter) all factors mentioned above (amount of independent paths, lengths, and local sequence similarity). Let us consider the network in the sequence space as an electrical network. The parameter, reflecting relatedness of fragment will be 'conductivity' (inverse resistance) between corresponding nodes. According to well-known laws of resistance: $R = R_1 + R_2$ for consecutive connection of R_1 and R_2, and $1/R = 1/R_1 + 1/R_2$ for parallel connection. Thus, this parameter takes into account both the amount of independent pathways and their lengths, allowing simple introduction of specific properties of connections.

We repeated the calculations described in the previous section using resistance (R) instead of maximum flow, as described in 'Methods'. The dependence of the conductivity ($1/R$) on RMSD is shown in Fig. 6. This curve is very similar to the one shown in Fig. 4, though of a smoother appearance. The curve of frequency of the structures from two nodes (at selected conductivity between them) to have similar 3D-structure (RMSD < 3 Å) is shown in Fig. 7.

It should be noted here that the value '10' of frequency at abscissa corresponds to $R = 1$, i.e., to pair of neighboring nodes. Since the node correspondence to a 3D-structure from PDB was defined also by a 60% similarity threshold, two similar PDB structures would be connected through three 'transitions': from the first PDB-sequence to the first node, from the first node to the second, and from the second one to the last PDB-sequence. As it has been measured earlier (Frenkel and Trifonov, 2007a), at 60% threshold the probability of 20 aa protein fragments from different families to have similar structure is about 85%. Thus,

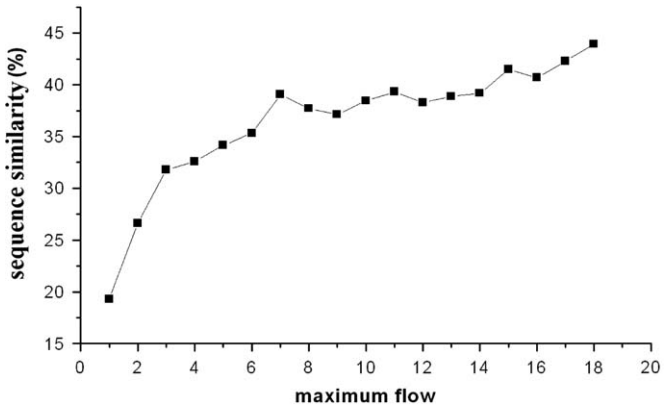


Fig. 8. Average sequence similarity between the nodes with different maximum flow.

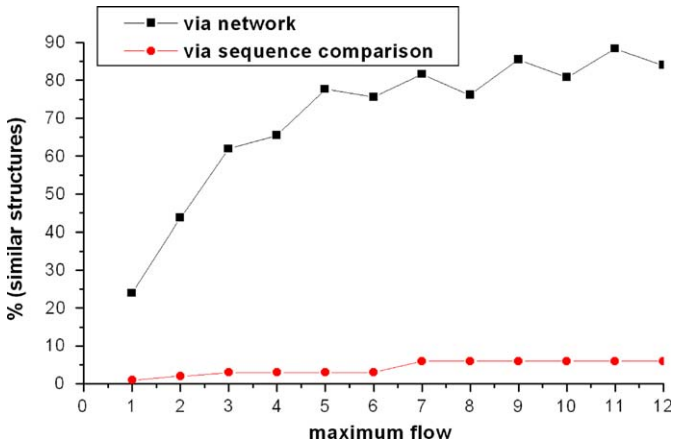


Fig. 9. Comparison of predictions of structure similarity (RMSD < 3 Å) via the network maximum flow and via the sequence similarity (according to Table 1, see text). The black curve is identical to Fig. 5. The red curve shows the structure similarity calculated with the sequence similarity value corresponding to this flow (see Fig. 8). (For interpretation of the references to the color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Structure similarity at different sequence similarity levels.[a]

Sequence identity (%)	Amount of sequence-wise similar fragments	Amount of similar structures (RMSD < 3 Å)	%
25	301003824	5281889	1.75476
30	51063933	1203097	2.35606
35	7064562	245873	3.48037
40	826721	54601	6.60453
45	93244	19506	20.9193
50	17865	11097	62.1159
55	7570	6762	89.3263
60	5281	5096	96.4969
65	3669	3570	97.3017
70	2792	2740	98.1375
75	1991	1973	99.0959
80	1471	1443	98.0965
85	1160	1132	97.5862
90	1041	1010	97.0221
95	829	808	97.4668
100	31250	31201	99.8432

[a] Only fragments with the exact level of similarity (as in column 1) are considered.

for three connections the probability will be about $100 \times (0.85)^3 = 61\%$. As one can see, the curve in Fig. 7 is in a good agreement with this result.

3.3. Sequence similarity

A natural question to ask is how much similarity based on 'flow' or 'current' is better than when only a direct pairwise sequence similarity is used? To check this we calculated the distribution of sequence similarity as a function of flow (i.e., per a given flow value *f*, what is the average sequence similarity between *s* and *t*, taken over all pairs *s*,*t* with flow *f*). As expected, sequence similarity increases with the network flow (Fig. 8). To estimate the level of structure similarity expected at the sequence similarity value obtained, calculations similar to those made previously in Frenkel and Trifonov (2007a) were carried out (summarized in Table 1): We measured the level of sequence similarity between all nodes included in our calculations ('flow' and 'current') and all other 20 aa protein fragments with known structure (our multi-set database obtained from Astral SCOP release rel.71). The distribution of pairs with similar structures (RMSD < 3 Å), as a function of pairwise sequence similarity is shown in Table 1. As can be seen from the table, only sequence similarity ≥50% implies structural similarity between the corresponding fragments also.

Fig. 8 shows distribution of sequence similarity as a function of flow between the fragments used in our calculations. From the figure it follows that all pairs with flow 20 and below have insufficient sequence similarity to allow accurate and straightforward structure similarity prediction. This should be contrasted to Fig. 5 where pairs with flow 3 and above have high-expected structure similarity.

Fig. 9 summarizes the difference in prediction between the two approaches, sequence similarity-based prediction and flow-based prediction: expected prediction for the average sequence similarity for the 'real' values of the network flow (from 1 to 10) is <10%, which strongly contrasts the prediction via sequence flow.

4. Discussion

The demonstration of the fact, that structure similarity, indeed, depends on amount of independent pathways or, more generally, on the network structure, attests to the importance of network approach to the sequence relatedness problem. The proposed measure of the relatedness via 'resistance' provides a new tool for application of the networks. We believe that this approach can provide a new, biologically justified definition of distances between sequences, instead of the common 'statistical' definition (Trifonov and Frenkel, 2009). Usually, specific weights (or 'costs') for mismatches (substitution matrix) and indels are introduced and corresponding 'scores' are calculated. However, every structurally/functionally specific site in the protein should allow only certain correlated types of mutations, which are leveled off when one averaged substitution matrix is used. Most of modern 'individual' approaches consider the existence of a 'sequence pattern' or 'profile'. In the network approach several connected clusters can harbor rather different sequences. Indeed, let us consider two physically interacting protein fragments. When such physical interaction is *the* function of the fragments in their respective proteins, the variety of sequences that would ensure such interaction (by combinations of, for instance, hydrophobic and polar residues) is unlimited. In this case, an introduction of 'average sequences' for the interacting fragments of the same interaction pair taken from different organisms does not make any

sense. The appearance of some 'consensus' would have questionable value, reflecting, for example, domination of one taxonomic group over another. However, combination of our network approach with more traditional pattern/profile analysis, where the conservation of some residues is a must, can be very fruitful.

In addition to the apparent benefit for sequence annotation, these parameters (amount of independent pathways and 'resistance') can be important in studies on protein evolution. It becomes possible not only to reveal new evolutionary relatedness by the networks (Trifonov and Frenkel, 2009), but also to derive a strict measure of this relatedness. Interestingly, the approach allows carrying out this measurement for every pair of sequences by decreasing the sequence similarity threshold, which gathers all sequence fragments into one huge network.

Further development of this proposed approach could lead to improvement in the application of this method. The first improvement that could be investigated is in the introduction of different resistance values between the nodes, depending on sequence similarity or other parameters of connected fragments. Another important step would be to take into account all other known structures of the network. They can be considered as additional 'voltage sources' in the electrical current of the sequence space. In addition, these 'known structures' can reveal key elements of the basic pattern for the structure. This pattern can consist of only 1–2 positions, which would make it impossible for detection by other methods. The closeness of the node fragments to this pattern should also be reflected in the value of resistance. All of these advantages have become possible only by using the networks in the sequence space.

In more general terms, the formalism of flow and resistance provides the sequence space with a frame, organizing it in a biologically meaningful construction, with identifiable and quantitatively measured relations between networks and clusters therein. Noteworthy is that a manifold of seemingly separate, dissimilar fragments can be joined together in a concerted way and partake in outlining functional and structural similarities in the protein sequence space.

5. Conclusions

The work demonstrates that the amount of independent pathways in the network in a sequence space reflects the structural similarity of corresponding protein fragments. The use of a network of short protein fragments of the same length enables the detection of structure similarity between seemingly completely unrelated sequences. Moreover, the flow approach provides an efficient algorithm for the evaluation of structural similarity thus opening new perspectives for annotations of large databases. The maximal flow is an appropriate parameter for reliable quantitative estimation of structural relatedness.

Acknowledgment

The work has been supported by the Center for Complexity Science Grant GR2006-018.

References

Ahuja, R.K., Magnanti, T.L., Orlin, J.B., 1993. Network Flows: Theory Algorithms and Applications. Prentice-Hall, Englewood Cliffs, NJ.
Batagelj, V., Mrvar, A., 2002. Pajek — analysis and visualization of large networks. In Graph Drawing 2265, 477–478.
Bornberg-Bauer, E., 1997. How are model protein structures distributed in sequence space?. Biophysical Journal 73, 2393–2403.
Bornberg-Bauer, E., Chan, H.S., 1999. Modeling evolutionary landscapes: mutational stability, topology, and superfunnels in sequence space. Proceedings of the National Academy of Sciences of the United States of America 96, 10689–10694.
Brenner, S.E., Koehl, P., Levitt, R., 2000. The ASTRAL compendium for protein structure and sequence analysis. Nucleic Acids Research 28, 254–256.
Chandonia, J.M., Hon, G., Walker, N.S., Lo Conte, L., Koehl, P., Levitt, M., Brenner, S.E., 2004. The ASTRAL compendium in 2004. Nucleic Acids Research 32, D189–D192.
Cherkassky, B.V., Goldberg, A.V., 1997. On implementing the push-relabel method for the maximum flow problem. Algorithmica 19, 390–410.
Chou, K.C., 2004. Structural bioinformatics and its impact to biomedical science. Current Medicinal Chemistry 11, 2105–2134.
Chou, K.C., Shen, H.B., 2007. Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides. Biochemical and Biophysical Research Communications 357, 633–640.
Chou, K.C., Shen, H.B., 2008. Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms. Nature Protocols 3, 153–162.
Ciriello, G., Guerra, C., 2008. A review on models and algorithms for motif discovery in protein–protein interaction networks. Briefings in Functional Genomics and Proteomics 7, 147–156.
Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C., 2001. Introduction to Algorithms, second ed. MIT Press and McGraw-Hill, Cambridge, MA and New York.
Dokholyan, N.V., 2005. The architecture of the protein domain universe. Gene 347, 199–206.
Dokholyan, N.V., Shakhnovich, B., Shakhnovich, E.I., 2002. Expanding protein universe and its origin from the biological Big Bang. Proceedings of the National Academy of Sciences of the United States of America 99, 14132–14136.
Eskin, E., Snir, S., 2005. The homology kernel: a biologically motivated sequence embedding into Euclidean space. In: Proceedings of the 2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, pp. 179–186.
Frenkel, Z.M., Trifonov, E.N., 2007a. Walking through the protein sequence space: towards new generation of the homology modeling. Proteins—Structure Function and Bioinformatics 67, 271–284.
Frenkel, Z.M., Trifonov, E.N., 2007b. Evolutionary networks in the formatted protein sequence space. Journal of Computational Biology 14, 1044–1057.
Frenkel, Z.M., Trifonov, E.N., 2007c. Walking through protein sequence space. Journal of Theoretical Biology 244, 77–80.
Gattiker, A., Michoud, K., Rivoire, C., Auchincloss, A.H., Coudert, E., Lima, T., Kersey, P., Pagni, M., Sigrist, C.J.A., Lachaize, C., Veuthey, A.L., Gasteiger, E., Bairoch, A., 2003. Automated annotation of microbial proteomes in SWISS-PROT. Computational Biology and Chemistry 27, 49–58.
Goldberg, A.V., Tarjan, R.E., 1988. A new approach to the maximum-flow problem. Journal of the ACM 35, 921–940.
Goodaire, E., Parmenter, M., 2005. Discrete Mathematics with Graph Theory, third ed. Prentice-Hall, Englewood Cliffs, NJ.
Hamming, R.W., 1950. Error detecting and error correcting codes. Bell System Technical Journal 26, 147–160.
Huang, Y., Li, Y., 2004. Prediction of protein subcellular locations using fuzzy k-NN method. Bioinformatics 20, 21–28.
Jewett, A.I., Huang, C.C., Ferrin, T.E., 2003. MINRMS: an efficient algorithm for determining protein structure similarity using root–mean–squared-distance. Bioinformatics 19, 625–634.
John, B., Sali, A., 2004. Detection of homologous proteins by an intermediate sequence search. Protein Science 13, 54–62.
Kopp, J., Schwede, T., 2004. Automated protein structure homology modeling: a progress report. Pharmacogenomics 5, 405–416.
Leslie, C., Eskin, E., Noble, W.S., 2002. The spectrum kernel: a string kernel for SVM protein classification. Pacific Symposium on Biocomputing 564–575.
Leslie, C.S., Eskin, E., Cohen, A., Weston, J., Noble, W.S., 2004. Mismatch string kernels for discriminative protein classification. Bioinformatics 20, 467–476.
Meyerguz, L., Kleinberg, J., Elber, R., 2007. The network of sequence protein structures flow between. Proceedings of the National Academy of Sciences of the United States of America 104, 11627–11632.
Nabieva, E., Jim, K., Agarwal, A., Chazelle, B., Singh, M., 2005. Whole-proteome prediction of protein function via graph–theoretic analysis of interaction maps. Bioinformatics 21, I302–I310.
Park, J., Teichmann, S.A., Hubbard, T., Chothia, C., 1997. Intermediate sequences increase the detection of homology between sequences. Journal of Molecular Biology 273, 349–354.
Shen, H.B., Chou, K.C., 2007a. EzyPred: a top-down approach for predicting enzyme functional classes and subclasses. Biochemical and Biophysical Research Communications 364, 53–59.
Shen, H.B., Chou, K.C., 2007b. Signal-3L: a 3-layer approach for predicting signal peptides. Biochemical and Biophysical Research Communications 363, 297–303.
Shen, H., Chou, J.J., 2008. MemBrain: improving the accuracy of predicting transmembrane helices. PLoS ONE 3, e2399.
Sobolevsky, Y., Trifonov, E.N., 2006. Protein modules conserved since LUCA. Journal of Molecular Evolution 63, 622–634.
Trifonov, E.N., Frenkel, Z.M., 2009. Evolution of protein modularity. Current Opinion in Structural Biology 19, 1–6.
Yang, X.G., Luo, R.Y., Feng, Z.P., 2007. Using amino acid and peptide composition to predict membrane protein types. Biochemical and Biophysical Research Communications 353, 164–169.