BIOINFORMATICS

Vol. 23 ECCB 2006, pages e163–e169
doi:10.1093/bioinformatics/btl290

Rediscovering secondary structures as network motifs—an unsupervised learning approach

Barak Raveh[1,2,*,†], Ofer Rahat[2,†], Ronen Basri[1] and Gideon Schreiber[2]

[1]Department of Computer Science & Applied Mathematics, and
[2]Department of Biological Chemistry, Weizmann Institute of Science, Rehovot, 76100, Israel

ABSTRACT

Motivation: Secondary structures are key descriptors of a protein fold and its topology. In recent years, they facilitated intensive computational tasks for finding structural homologues, fold prediction and protein design. Their popularity stems from an appealing regularity in patterns of geometry and chemistry. However, the definition of secondary structures is of subjective nature. An unsupervised de-novo discovery of these structures would shed light on their nature, and improve the way we use these structures in algorithms of structural bioinformatics.

Methods: We developed a new method for unsupervised partitioning of undirected graphs, based on patterns of small recurring network motifs. Our input was the network of all H-bonds and covalent interactions of protein backbones. This method can be also used for other biological and non-biological networks.

Results: In a fully unsupervised manner, and without assuming any explicit prior knowledge, we were able to rediscover the existence of conventional α-helices, parallel β-sheets, anti-parallel sheets and loops, as well as various non-conventional hybrid structures. The relation between connectivity and crystallographic temperature factors establishes the existence of novel secondary structures.

Contact: barak.raveh@weizmann.ac.il; gideon.schreiber@weizmann.ac.il

1 INTRODUCTION

1.1 Secondary structures in the eye of the beholder

In the early 1950s, two exciting breakthroughs lay the foundations to the field of structural biology. Both showed elegant symmetry of recurring patterns in the structure of biological molecules. The second, and more famous to the public, was the discovery of the double-stranded structure of DNA (Watson and Crick, 1953). Two years earlier, in a series of seven seminal papers, Linus Pauling and Robert Corey established the existence of periodical helical and pleated sheet patterns in proteins (Pauling *et al*., 1951; Eisenberg, 2003). These include α-helix, β-sheets, β-turns and less frequent structures like π-helix and 3_{10}-helix.

Secondary structures form the first layer of simplification for describing tertiary and super-secondary protein topologies. Cartoon drawings of proteins use secondary structures as a useful tool for schematic visualization of protein backbones. Secondary structures are widely used in databases of structure classification, algorithms of structural alignments, fold prediction, protein design or docking and other applications in the field of computational structural biology (Murzin *et al*., 1995; Orengo *et al*., 1997, Yang and Honig, 2000). The theoretical framework for defining these elementary structures relied on laborious enumeration of dozens of theoretical structures, choosing the ones with favorable geometric patterns that allowed an optimal set of hydrogen bonding patterns while preventing sterical clashes between neighboring atoms. In the late 1950s, a slow drip of solved X-ray structures, starting with the first crystal structure of Myoglobin (Kendrew *et al*., 1958), confirmed the abundance of Pauling's helices and sheets. However, the overall tertiary topology of protein structures does not show any simple coherent symmetry, at least not one apparent to the human eye. Today we know this inherent complexity allows the enormous range of roles played by proteins in a living cell. Moreover, native secondary structures do not conform to the exact ideal parameters predicted by Pauling and Corey, who did not account for the twists and curvatures exhibited by sheets and helices (Barlow and Thornton, 1988; Martin *et al*., 2005), as well as other irregularities like β-bulges and helical kinks. Consequently, definitions of secondary structures are inherently subjective. Crystallographers often disagree in their assignments of secondary structure (Andersen and Rost, 2003). Therefore, great caution should be used when relying on both man-made and automatic assignments of secondary structures.

Numerous computational tools have been designed for automatic assignment and prediction of secondary structures, aiming to capture man-made intuitions about what a secondary structure is. DSSP (dictionary of secondary structures in proteins) (Kabsch and Sander, 1983) is considered a golden standard for automatic assignment of secondary structures to protein folds. It searches for predefined periodicity patterns of hydrogen bonds in protein backbones. Other common assignment methods are DEFINE and STRIDE that also use Φ-Ψ dihedral angles (see Andersen *et al*., 2003). Two recent works were able to characterize secondary structures using descriptors of Voronoï and Delaunay tessellations (Dupuis *et al*., 2004, Taylor *et al*., 2005). All of the above methods are known to exhibit significant disagreements, especially in telling loop regions from beta sheets or short helices, and in deciding helix boundaries (Andersen *et al*., 2003). These differences are inherent to the subjective nature of the problem, and each assignment can be considered legitimate in some sense. For instance, an algorithm of rigid protein docking might favor a geometrically driven definition, whereas another algorithm might rely on patterns of hydrogen bonds.

1.2 Unsupervised learning

Unsupervised learning usually refers to the detection of patterns without the use of explicit prior knowledge about these patterns.

[*]To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

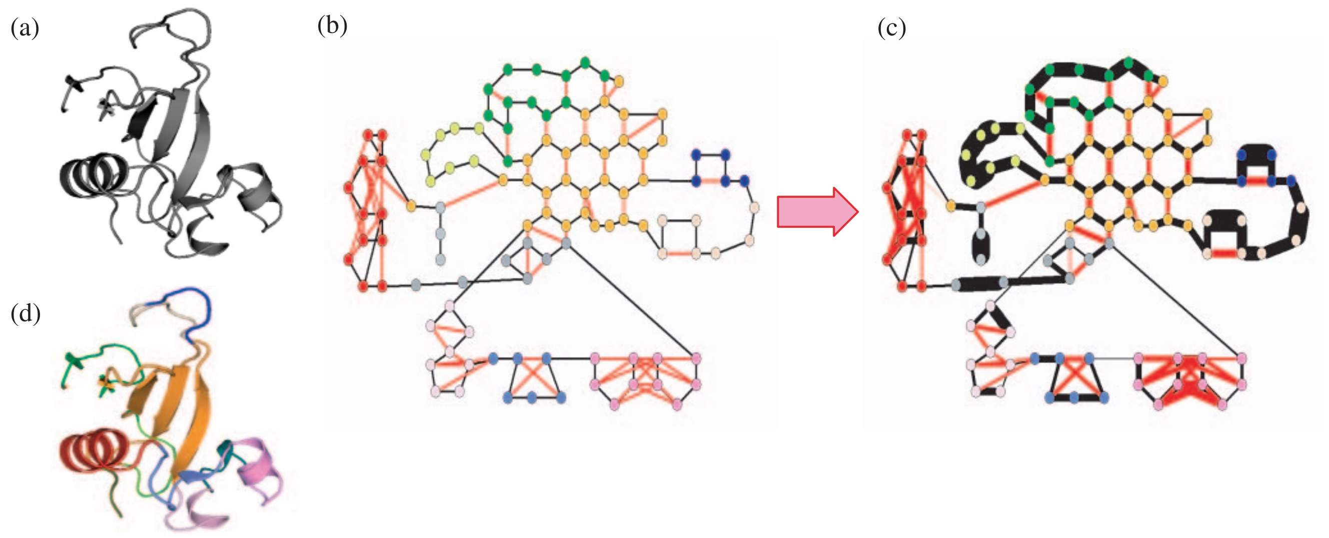

Fig. 1. Unsupervised clustering of the ribonuclease inhibitor Barnase (1BRS chain d). **(a)** Input PDB structure 1BRS_d (image created in PyMOL). In **(b)** and **(c)** we see the enrichment process of the graph which precedes a normalized-cuts clustering step. Node colors are the final result of this clustering process, they are added for illustration purposes only. (b) Graph representation of Barnase. Each node maps to a single residue. Edges stand for inter-residue backbone interactions: black for covalent bonds, red for hydrogen bonds. Note the different patterns in the graph: the anti-parallel sheet (orange nodes) generally looks like a beehive of hexagons. Helical regions (red and violet) and some of the loop regions (blue, pink, dark green) are densely bonded, but with varying network patterns, and contrary to other coil regions (lime, off-white). (c) An enrichment of (b) reinforces significant patterns. Edges were reweighed using the recurring patterns of network motif profile affinity matrix (see text). A normalized cuts algorithm partitions the graph into clusters of high intra-connectivity. The reweighting of the edges assures that significant patterns would be clustered together. **(d)** A coloring of 1BRS with respect to the cluster colors in (b) and (c). Since the clustering procedure searched for uniform patterns of network motifs, the anti-parallel sheet is successfully separated from the neighboring green loop region.

This is in contrast to supervised methods like neural networks, decision trees and support vector machines, in which explicit labeling of data is given in advance. While unsupervised learning implies no prior assumptions about the specifics of a pattern, we still incorporate our very basic notions of common sense to define what kinds of patterns we consider interesting (but not anything beyond that). For instance, in clustering algorithms the aim is to group similar objects together, without a predefined set of examples for the correct partitioning. However, we usually define what this ‘similarity’ is. Different definitions are bound to yield completely different lines of partition.

Methods of unsupervised learning were successfully applied to various biological problems. Clustering algorithms are often used for analysis of biological micro-array chips (Getz *et al*., 2000) and for detection of metabolic and regulatory pathways in biological networks (Segal *et al*., 2003). In the field of systems biology, Milo *et al*. (2002) used an exhaustive search procedure for small network motifs in biological regulatory networks, which yielded important theoretical insights into mechanisms of transcriptional regulation. Many of these mechanisms were experimentally validated.

In the case of protein structures, DSSP, STRIDE and DEFINE should all be considered supervised methods in the sense that they rely on predefined expert assignments. Numerous studies have sought novel patterns of recurrence within protein structures in an unsupervised manner and are used for structure prediction (Unger and Sussman, 1993; Bystroff and Baker, 1999). They cluster local segments of proteins based on geometrical similarities. However, the scope of such methods does not include global structures like whole β-sheets.

In this study, we take a global approach for *de novo* detection of secondary structures. We embed protein structures in undirected graphs that capture the essence of the biochemical network of interactions within a protein (Fig. 1b). We formulate a general unsupervised methodology for detecting patterns of regularities in graphs and networks, which is employed for the *de novo* discovery of secondary structures.

2 METHODS

Our dataset consists of 220 high resolution non-redundant X-ray structures from the July 2005 version of the PDBSelect dataset (resolution <1.20 Å, R-factor <0.22, <25% identity, all smaller than 1000 amino acids, see Hobohm and Sander, 1994).

2.1 Embedding protein structures in graphs

We map a protein structure to an undirected graph $G = (V, E)$ where V are graph nodes (vertices) and E is a set of undirected edges $\{v_i,v_j\}$. Each node v represents a residue and each edge e is either a covalent bond or a hydrogen bond between backbone atoms of two residues (Fig. 1a and b). Hydrogen bonds were extracted using the program BndLst (v.1.6), based on the tools Probe and Reduce (Word *et al*., 1999). While we considered other graph representations that included side-chain interactions such as hydrophobic packing and salt-bridges, we eventually limited ourselves to covalent and backbone hydrogen interactions, since in this study we are mainly interested in a backbone description of folds.

2.2 Calculating network motif vector

Having built graph representation of proteins, we observed an interesting pattern of small network motifs that appear frequently in graphs. For instance, hexagons are typical to graph regions of anti-parallel β-sheets (Fig. 1b), squares characterize parallel sheets (results not shown), etc.

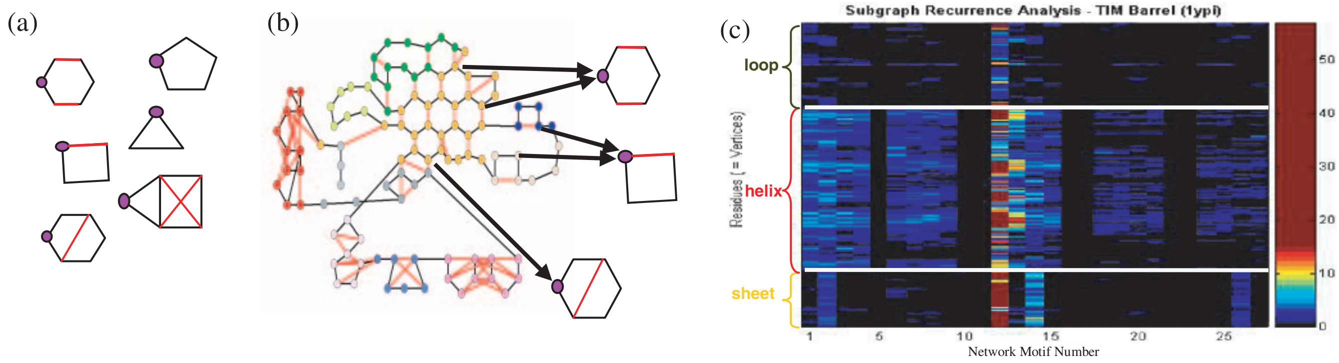

Fig. 2. Profiles of network motifs. (**a**) Consider 1500 different small network motifs, encompassing nearly all graphs over up to six nodes, with edges labeled as 'covalent' (black), 'H-bond' (red) or both. In purple is the 'origin' of the motif, a node that is labeled differently from all other nodes (**b**) For each residue, we count how many times it appear in each network motif. Only ~180 of theoretical motifs appear in graphs of native proteins (see Methods). Different residues appear in different motifs (**c**) Visualization of the resulting matrix for a TIM-barrel fold (1YPI), showing the vector of network motifs for each residue. Each row is the network motif vector of a single residue. Each column represents another network motif (first 30 columns shown). Rows are sorted according to secondary structure as labeled by DSSP (top to bottom: coils, helices and sheets). The color of each cell indicates the number of times a given residue appears in a certain subgraph motif (see color-bar). It is readily observed that profiles of network motifs can distinguish helices, sheets and loops.

We exploited these motifs for capturing the local environment of each node in the graph and detect recurring patterns in the graphs. Therefore, we used an exhaustive enumeration of 1500 small network motifs in graphs of proteins structures (Fig. 2). A motif $Gm = (Vm, Em)$ is a graph over up to six nodes, where one of the nodes is labeled as the 'origin' (Fig. 2a). An edge is labeled either as 'covalent' or 'H-Bond'. Such motif represents a local micro-environment of backbone interactions between up to six residues.

Suppose a graph G contains a subgroup of nodes $V' = (v_1, v_2, v_3, \ldots, v_k)$, and that there is an isomorphic mapping from the induced subgraph over V' to some motif Gm. If node v_1 maps to the 'origin' node of Gm, we say that v_1 appears in the motif Gm (Fig. 2b). We counted the number of times each node (= residue) appears in each of the 1500 possible motifs. In practice, only 180 motifs appeared in structure networks. Hence, for each residue we fill in a vector p_i of length 180 such that

$P_{i,\ m}$ = [No. of occurrences of the residue i in motif Gm].

We refer to this vector as the network motif vector of residue *i*.

Note that if we were to choose a supervised learning scheme, we could readily use motif vectors to distinguish helices, sheets and coils at this stage (Fig. 2c). To validate this assumption, we used support vector machines to learn the assignments of DSSP based on the network motif vectors and were able to successfully predict DSSP assignments with ~90% success rate on all categories using 10-fold cross-validation tests. Although our main purpose in this work is the rediscovery of secondary structures without assuming prior knowledge, this validation reassures us that network motif vectors preserve the essential information needed for secondary structure classification.

2.3 Affinity matrix of network motif profiles

Using all residues from the 220-proteins dataset, we retrieve 40 000 network motif vectors, one per residues. We applied *k*-means clustering (Matlab 7.0.4, $k = 25$, Euclidean metrics, 5000 residues used to find centers of mass) to partition the residues into 25 bins, denoted from here on as network motif profiles (NMPs). Each such bin represents a profile of similar motif vectors. We note that the number 25 is arbitrary. However, different number of bins, from 20 to 30, did not have significant impact on our results. To conclude, we label each node in each graph with a number between 1 and 25, denoting its network motif profile. This number captures a profile of local backbone interactions in the environment of a given residue.

In order to extract patterns in the graphs, we count the number of times each of the 25 NMPs is connected by an edge to another NMP. Formally, we fill in a 25 × 25 matrix S:

S_{ij} = [number of edges (u, v) such that NMP of $v = i$ and NMP of u is j].

We normalize the columns of S to sum to 1. This means a column *i* is a distribution of neighbors for NMP #*i*, and a row *j* represents the share of NMP #*j* among the neighbors of the other NMPs. Denote this asymmetric normalized matrix by N. We calculate a final 25 × 25 affinity matrix A by summing the innerproducts of rows and of columns in N:

$$A = N^{\tau}N + NN^{\tau}$$

Intuitively, A_{ij} measures how much two network motif profiles *i* and *j* tend to share a similar environment of profiles in all graphs. We use both rows and columns in our calculation, in order to account for both how much an NMP is important to its neighbors, and how much its neighbors are important for it.

2.4 Enrichment of the graphs

A high value of A_{ij} implies that NMPs *i* and *j* belong to a similar pattern, in the sense that they prefer the same profiles of neighbors. We would like to manipulate our original graphs, and increase weights of edges connecting similar patterns (Fig. 1c). Let (*u*,*v*) be an edge in a graph of some protein, and let *i* and *j* be the NMPs of *u* and *v*, respectively. We set the weight of (*u*,*v*) to be A_{ij}. We refer to this process of reweighting as an 'enrichment process'. We postulate this process reinforces edges that form significant patterns, while reducing the weights of less significant edges.

2.5 Clustering the enhanced graphs to 'sites'

Following the enrichment process, we reduced the problem of finding secondary structures in an unsupervised manner into a problem of clustering graph nodes. We would like to find clusters of nodes with dense intra-connectivity. This is a natural framework for the normalized-cuts family of clustering algorithms. This family of algorithms cut graphs into groups of nodes, maximizing a score:

$$\mathrm{SCORE}_{\mathrm{NC}} = \frac{\sum [\text{edge weights within clusters}]}{\sum [\text{edge weights between clusters}]}.$$

Graph-cut algorithms were successfully employed in the past to detect protein domains (Xu *et al.*, 2001).

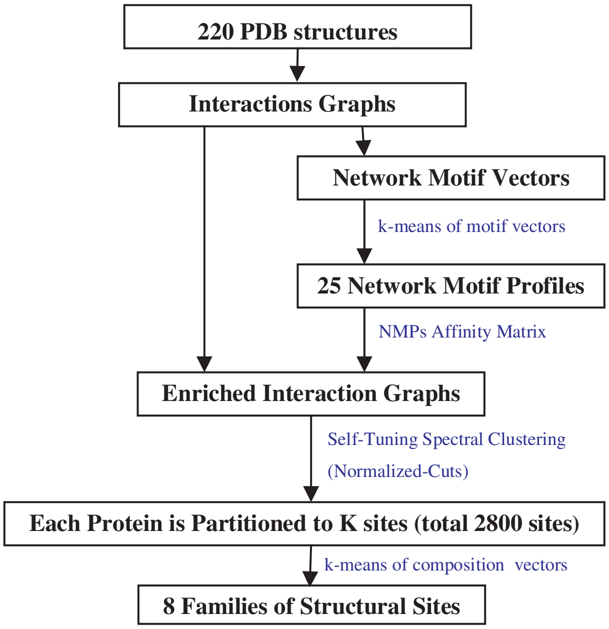
220 PDB structures
Interactions Graphs
Network Motif Vectors
k-means of motif vectors
25 Network Motif Profiles
NMPs Affinity Matrix
Enriched Interaction Graphs
Self-Tuning Spectral Clustering
(Normalized-Cuts)
Each Protein is Partitioned to K sites (total 2800 sites)
k-means of composition vectors
8 Families of Structural Sites

Fig. 3. An overview of the unsupervised learning scheme.

We used the self-tuning spectral clustering algorithm (Zelnik-Manor *et al.*, 2004), which approximates both the normalized cuts score and the total number of clusters for each protein. After applying the clustering algorithm, we return to the original protein, and get the unsupervised partitioning over its residues (Fig. 1d). We refer to each cluster of residues as a 'site'. Overall, we get ~2800 sites for the full dataset.

2.6 Meta-clustering into families of sites

We would now like to cluster the sites themselves into different families. These families are analogous to the conventional families of secondary structures, i.e. helix, sheets, etc. For this aim, we would like to characterize each of these sites quantitatively.

Let a site S be a set of protein residues. Recall that each residue is assigned one of 25 network motif profiles. We characterize each site by NMPs composition vector C_s:

$C_{s,\ i}$ = {percentage of residues in site S that belong to NMP i}

Example:Let S be a site composed of 10 residues:

S = [100, 101, 102, 103, 117, 118, 119, 120, 121, 122]

Suppose these residues correspond to the following NMP profiles:

NMP(S) = [1, 1, 1, 1, 2, 3, 3, 3, 3, 3]

In this case, we characterize S be by its composition of NMPs:

C_s = [40%, 10%, 50%, 0% ...]

Using the k-means algorithm, we cluster the composition vectors of all 2800 sites into eight groups. We get eight families of site types, encompassing our novel 'secondary structures'. These families are retrieved without postulating the types of structures we look for in advance. Figure 3 summarizes our procedure for learning secondary structures.

3 RESULTS

In order to understand the meaning of the sites and site families we received, we explore some of their properties.

3.1 Comparison to DSSP

Here and in the sequel, we compare our results with that of DSSP (Kabsch and Sander, 1983). DSSP was one of the first algorithms for secondary structure assignments, and is still one of the most widespread used. We applied it [DSSPCMBI April 2000] to our set of proteins. We grouped the eight DSSP labels into three categories: helix = 'HGI' (α-helix, 3-helix and π-helix), sheet = 'EB' and loop = 'TS' (turns and bends).

First, we manually inspected the sites we got for each protein in our dataset by looking at the colored structures (like in Fig. 1d), comparing them with DSSP. This initial inspection showed conventional helices and sheets conform well to sites we received after the normalized-cuts step. However, some sheets come out together with loop residues from the vicinity of the main sheet. We will show below that the identity of such loops is open to discussion.

Site Purity (with respect to DSSP) For each of our sites, we give an overall site classification of helix/strand/coil, according to the DSSP labels of the majority of its residues. We now calculate site 'purity', defined as number of residues in the site labeled the same as DSSP. For instance, if a site has 23 residues, of which 19 are labeled as Helix by DSSP, the site will be labeled as Helix site with purity of 19/23 = 83%. Average purity of helices calculated in this way is **78%**, sheet sites purity is **68%** and loop sites purity is **86%**. For sheet sites, loop residues are most of the 32% non-sheet residues. We claim this is an inherent ambiguity regarding the definitions of sheets, with similar disagreements shown between other assignment methods, see comparison to B-factors below.

Site Families versus DSSP In the last 'meta-clustering' step, we assigned each site to one of eight site families. A comparison to DSSP reveals the relation between these eight families and conventional secondary structures. In Figure 4a and b we see the distribution of DSSP assignments for each of the eight site families. Types 7 and 8 are analogous to helices, but type 7 (magenta in Fig. 4c) helices are shorter on average than type 8. Type 3, and to a lesser degree type 2, corresponds to DSSP sheets. Manual inspection revealed that type 3 is mostly assigned to anti-parallel sheets, whereas type 2 is assigned to parallel sheets (Fig. 4c). This means our unsupervised clustering was able to discern these two fundamental types of sheets. Types 1, 4, 5 and 6 all contain mostly loop areas, but type 1 has a high content of sheets, and type 5 captured some helices, mostly short ones. One such helix appears in grey in Figure 4c (encircled), with less than one turn. Its classification by DSSP as a helix could be easily challenged.

3.2 Independent comparison to crystallographic B-factors

Each atom in X-ray structures in the PDB database is assigned a numerical value termed B-factor, also known as temperature factors, which describes the positional uncertainty for this atom. Higher B-factors imply greater positional uncertainty, reflecting the degree of thermal motion and static disorder of an atom in a protein crystal structure (Drenth *et al.*, 1994). A good definition for families of secondary structures should account for different flexibility patterns in different regions of the protein, which has important functional impacts. Loop regions are generally considered areas of higher backbone flexibility. In Figure 4d–f we show how our site families capture these differences. We compare B-factor values of Cα atoms

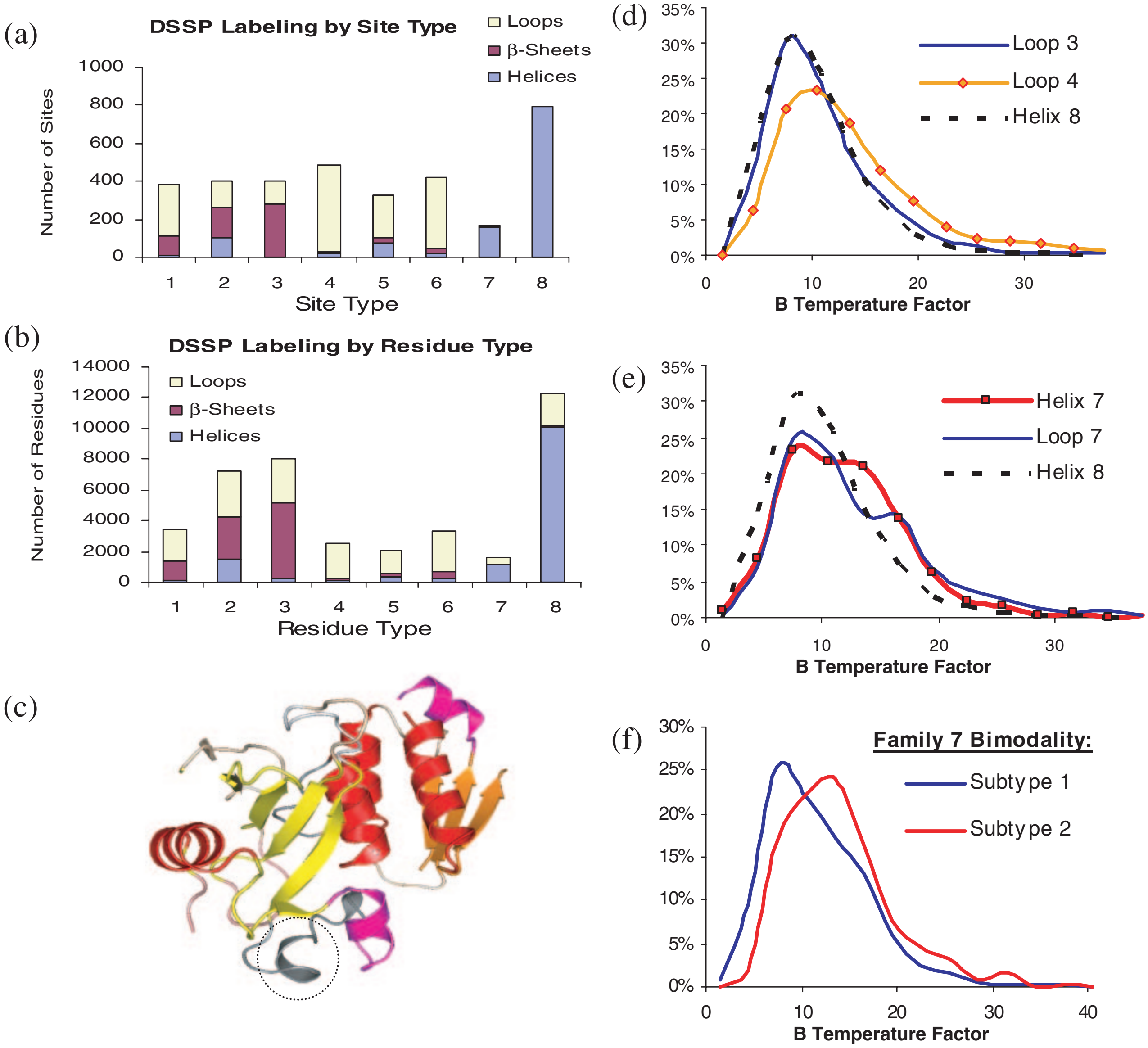

Fig. 4. (**a**) The composition of the eight novel secondary family types according to DSSP. We considered a site to be a DSSP helix if the majority of its residues are helical according to DSSP, and the same for sheets and loops. Family types 7 and 8 are analogous to helices. Type 3 and to a lesser degree type 2 to sheets, and types 1, 4, 5 and 6 to DSSP loops. (**b**) The same analysis done at a residue level. (**c**) The Barnase–Barster complex (1brs chains A,D) colored according to the eight novel families of secondary structures. Long helices mapped to family 8 (red), shorter helices to 7 (magenta), anti-parallel sheets to 3 (yellow, in front) and parallel ones to type 2 (orange, in back). Encircled, a short DSSP helix of less than a single turn was mapped to family type 5. It could be easily argued this is not really a helix (Color Coding: 1, wheat; 2, orange; 3, yellow; 4, pink; 5, grey; 6, white; 7, magenta; 8, red). (**d–f**) DSSP classifications of some of the loop areas might not tell us the full story. Crystallographic B-factors (temperature factors) are known to correlate with positional flexibility. These figures show the distribution of B temperature factors of Cα atoms according to various secondary families. (**d**) Demonstrates the difference between loops we classify as types 3 and 4. Although 36% of type 3 residues are classified as loops by DSSP, the B-factors of these residues (solid blue curve) are significantly lower than those of type 4 loops (solid orange). Moreover, distribution of type 3 loops is practically identical to well-ordered helix residues of type 8 (black dotted line) and very similar to type 3 sheet residues (not shown). (**e**) Despite the fact that both family types 7 and 8 are mostly helical (see a and b), helix residues classified as type 7 show much higher flexibility than those in type 8 (solid red versus black dotted line), and are almost identical to loops of type 7. This correlates well with type 7 helices being shorter and hence more flexible. (**f**) In (e), a 'bump' is observed in both loops and helices of type 7. By further clustering of family type 7, we revealed a bimodal distribution responsible for this bump.

for different site families. In Figure 4d we focus on sheet family 3. This family is analogous to anti-parallel sheets, but DSSP classifies some of its residues as loops. We reveal an inherent difference between DSSP loop regions of family types 3 and 4; type 3 loop residues are much 'colder' than type 4 loops (p-value < 10^{-9} for the two distributions to be the same, using Wilcoxon Rank Sum Test). In fact, the flexibility pattern of type 3 loops is identical to that of long structured helices of type 8, and very similar to type 3 sheet residues. In Figure 4e we compare helix families 7 and 8. We show shorter helices of type 7 are much more flexible than helices of type 8. Moreover, those residues of type 7 classified as loops actually had similar or even lower flexibility than DSSP helix residues of type 7. We also noted a ‘bump’ in the B-factor distribution of family 7, which might indicate a bimodal distribution. We hence clustered family 7 into two subgroups of populations, which indeed had significantly different patterns of flexibility (Fig. 4f).

In addition, we were able to discern loop types 1, 4, 5 and 6: type 4 showed the greatest flexibility, accounting for regions of great backbone flexibility, and is much shorter on average than types 1 and 6. These are all examples of finer divisions that are not revealed by conventional definitions of secondary structures.

4 DISCUSSION

In this work we address the question of secondary structures from a physical perspective. We look for uniform patterns in interaction networks of proteins, without defining in advance what a secondary structure is, and without knowing what to expect. We believe this gives our methodology the power to discern patterns that are otherwise non-evident to the bias of the human eye.

4.1 Comparison to conventional definitions of secondary structures

DSSP and other conventional assignment methods try to match traditional expert definitions of secondary structures. In contrast, we try to redefine secondary structures from scratch, and end up with significant similarities to conventional definitions. We show high correlations between the novel eight families of sites and conventional definitions of secondary structures. To our knowledge, this is the first time that sparse global structures like β-sheets are rediscovered in an unsupervised manner, without having to rely on conventional definitions. In particular, we were able to rediscover the subdivision to parallel and anti-parallel sheets. We also show that some of the disagreements with traditional definitions stem from real structural and functional differences, not captured by current assignment methods. Some examples shown in Figure 4 are the differences in flexibility and in length between helices of family types 7 and 8 and between loops of type 3 and 4, as well as the ‘‘pseudo-helix’’ shown in Figure 4c. Since our novel definitions capture flexibility patterns, they might improve the topological descriptions of proteins fold, and help in tasks such as structural alignment and homology modeling.

4.2 A correct number of site families

Deciding a correct number of clusters is one of the hardest obstacles for unsupervised methodologies. Although our choice to split structures between eight families was quite arbitrary, it allowed us to retrieve finer grained details than conventional assignments. We showed one of our family types can indeed be further split into subgroups (Fig. 4f). In future work, we wish to devise better ways to asses the correct number of family types.

4.3 Embedding structures in sparse networks

This study relied on an extremely sparse representation of a protein structure, which completely ignored structural geometry, side-chain interactions and atomic level description. Since we limited ourselves to backbone interactions, the degree of each node in the network was bounded from above by 4 (two covalent bonds and two possible hydrogen bonds). Computationally, such a sparse graph can be manipulated very efficiently. This is in contrast to common graph representations like the full graph of distances between all protein atoms. This brings up questions about the conservation of information in such sparse networks. We were able to extract useful information about the structure, using this sparse representation alone. Hence, we suggest this kind of sparse representations could facilitate complex computational tasks like homology modeling. We believe a good representation would keep relevant interactions, while ignoring others that might add both noise and complexity into the system. For a recent survey of graph representations for protein structures, see Brinda *et al*., 2005.

4.4 Biochemical network motif profiles

Small network motifs played a major role in systems biology in recent years (Milo *et al*., 2002). We extended their use to quantify the local environment of a node in a graph. We suggest this compact quantitative description can be useful for detecting patterns in other networks as well.

4.5 Summary

Secondary structures are fundamental to the field of structural biology. We believe current advances in computational biology calls for novel definitions for the hierarchy of protein folds. These would improve the way we describe protein topologies, and consequently facilitate complex biological and computational tasks.

In this work we were able to rediscover conventional secondary structures de-novo, without assuming their existence in first place. We retrieve new insights about both novel structures and conventional ones. Since our framework is very general, we hope it will be used to discover patterns in other biological and non-biological networks.

ACKNOWLEDGEMENTS

This work was partially funded by the Israel Ministry of Science and Technology. O.R. thanks Yossi Shaul for fruitful discussions about the problem of assigning secondary structures.

REFERENCES

Andersen,C.A.F. and Rost,B. (2002) Secondary structure assignment. In Bourne,P.E. and Weissig,H. (eds), *Structural Bioinformatics*. Wiley-Liss, Hoboken, NJ, USA .

Barlow,D.J. and Thornton,J.M. (1988) Helix geometry in proteins. *J. Mol. Biol*., **201**, 601–619.

Brinda,K.V. *et al*. (2005) Insights into the quaternary association of proteins through structure graphs. *Biochem. J*., **391**, 1–15.

Bystroff,C. and Baker,D. (1998) Local structure prediction using a library of sequence-structure motifs. *J. Mol. Biol*., **281**, 565–77.

Dupuis,F. *et al*. (2004) Protein secondary structure assignment through Voronoi tessellation. *Proteins*, **55**, 519–528.

Drenth,J. (1994) *Principles of Protein Crystallography*. Springer-Verlag, NY.

Eisenberg,D. (2003) The discovery of the α-helix and β-sheet, the principal structural features of proteins. *Proc. Natl Acad. Sci. USA*, **100**, 11207–210.

Getz,G. *et al*. (2000) Coupled two-way clustering analysis of gene microarray data. *Proc. Natl Acad. Sci. USA*, **97**, 12079.

Hobohm,U. and Sander,C. (1994) Enlarged representative set of protein structures. *Protein Sci*., **3**, 522.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Kendrew,J.C. *et al.* (1958) A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, **181**, 662–666.

Martin,J. *et al.* (2005) Protein secondary structure assignment revisited: a detailed analysis of different assignment methods. *BMC Struct. Biol.*, **5**, 17.

Milo,R. *et al.* (2002) Network motifs: simple building blocks of complex networks. *Science*, **298**, 824–827.

Murzin,A.G. *et al.* (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

Orengo,C.A. *et al.* (1997) CATH—a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.

Pauling,L. *et al.* (1951) The structure of proteins—two hydrogen-bonded helical configurations of the polypeptide chain. *Proc. Natl Acad. Sci. USA*, **37**, 205–211.

Pauling,L. and Corey,R.B. (1951) The pleated sheet, a new layer configuration of polypeptide chains. *Proc. Natl Acad. Sci. USA*, **37**, 251–256.

Segal,E. *et al.* (2003) Module networks: identifying regulatory modules and their condition specific regulators from gene expression data. *Nat. Genet.*, **34**, 166–176.

Taylor,T. *et al.* (2005) New method for protein secondary structure assignment based on a simple topological descriptor. *Proteins*, **60**, 513–524.

Unger,R. and Sussman (1993) The importance of short structural motifs in protein structure analysis. *J.Comp Aided. Mol. Design*, **7**, 457–472.

Watson,J.D. and Crick,F.H.C. (1953) The structure of DNA. *Cold Spring Harbor Symposia on Quant. Bio.*, **18**, 123–131.

Word,J.M. *et al.* (1999) Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. *J. Mol. Biol.*, **285**, 1711–1733.

Word,J.M. *et al.* (1999) Asparagine and glutamine: using hydrogen atom contacts in the choice of sidechain amide orientation. *J. Mol. Biol.*, **285**, 1733–1745.

Xu,Y. *et al.* (2000) Protein domain decomposition using a graph-theoretic approach. *Bioinformatics*, **16:12**, 1091–1104.

Yang,A.S. and Honig,B. (2000) An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance. *J. Mol. Biol.*, **301**, 665–678.

Zelnik-Manor,L. and Perona,P. (2004) Self-tuning spectral clustering. In *18th Annual Conference on Neural Information Processing Systems.*