BIOINFORMATICS ORIGINAL PAPER
Vol. 25 no. 22 2009, pages 2921–2928
doi:10.1093/bioinformatics/btp541

Structural bioinformatics

Understanding hydrogen-bond patterns in proteins using network motifs

Ofer Rahat[1], Uri Alon[2], Yaakov Levy[3,*] and Gideon Schreiber[1,*]
[1]Department of Biological Chemistry, [2]Department of Molecular Cell Biology and [3]Department of Structural Biology, Weizmann Institute of Science, Rehovot 76100, Israel

Received on January 25, 2009; revised on September 8, 2009; accepted on September 10, 2009
Advance Access publication September 18, 2009
Associate Editor: Anna Tramontano

ABSTRACT

Summary: Protein structures can be viewed as networks of contacts (edges) between amino-acid residues (nodes). Here we dissect proteins into sub-graphs consisting of six nodes and their corresponding edges, with an edge being either a backbone hydrogen bond (H-bond) or a covalent interaction. Six thousand three hundred and twenty-two such sub-graphs were found in a large non-redundant dataset of high-resolution structures, from which 35 occur much more frequently than in a random model. Many of these significant sub-graphs (also called network motifs) correspond to sub-structures of α helices and β-sheets, as expected. However, others correspond to more exotic sub-structures such as 3_{10} helix, Schellman motif and motifs that were not defined previously. This topological characterization of patterns is very useful for producing a detailed differences map to compare protein structures. Here we analyzed in details the differences between NMR, molecular dynamics (MD) simulations and X-ray structures for Lysozyme, SH3 and the lambda repressor. In these cases, the same structures solved by NMR and simulated by MD showed small but consistent differences in their motif composition from the crystal structures, despite a very small root mean square deviation (RMSD) between them. This may be due to differences in the pair-wise energy functions used and the dynamic nature of these proteins.

Availability: A web-based tool to calculate network motifs is available at http://bioinfo.weizmann.ac.il/protmot/.

Contact: gideon.schreiber@weizmann.ac.il; koby.levy@weizmann.ac.il

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Proteins are made out of hundreds of amino acids folded to a well-defined structure that is stabilized by thousands of interactions. Discretization of the 3D structure of proteins can be done in many ways. Bystroff and Baker (1998) constructed a library of sequence-structure motifs, which was the base for the Bayesian separation of the total energy score into components that describe the likelihood of a particular structure. Unger *et al.* (1989) as well as others (de Brevern *et al*., 2000; Kolodny *et al*., 2002; Micheletti *et al.,* 2000; Oliva *et al.,* 1997; Wintjens *et al*., 1996) analyzed short oligopeptides and showed that their structure tends to concentrate in specific clusters rather than to vary continuously. A discrete repertoire of standard structural building blocks taken from these clusters was suggested as representative of all folds, and is often referred to as 'fold motifs'.

High resolution data of a protein can be represented as a contiguous stretch of 3D points, or alternatively as a mathematical graph based on the atomic contact map in which a contact is defined based on a predetermined threshold (typically of 8–10 Å between non-adjacent C_α atoms). Dokholyan *et al.* (2002) used the graph connectivity to predict folding probability. Huan *et al.* (2004, 2006) developed a frequent subgraph mining algorithm, and applied it to the contact map. The authors defined a subgraph as frequent, if it occurs in some predefined fraction of the studied proteins. The algorithm facilitates automated annotation of structures with unknown function. Later, the same methodology proved successful in mining RNA tertiary motifs (Wang *et al.,* 2007).

Contact maps discretize at the level of interatomic contacts, enabling a refinement of the contact map definition by discrimination of interactions according to their chemical nature (e.g. H-bonds, electrostatics). For comparing structures, this gives the overlay of contact maps a clear advantage over the RMSD of distance. The advantage here is the possibility of rationally choosing thresholds, by studying the contact chemical properties. Using such a refined contact map definition, we previously found that contact maps encapsulate the information necessary to detect the secondary structure (Raveh *et al.,* 2007). In another study we used the refined contact map definition to establish a novel notion of modularity in proteins interfaces (Reichmann *et al.*, 2005), and further used this scheme to study the evolution of protein interfaces (Rahat *et al.,* 2008).

A widely used scheme of systems biology suggests that networks are made up of a small set of recurring patterns, called *network motifs*. These are frequent subgraphs, where a probability is assigned to each subgraph individually, based on some background model of a random network. Furthermore, analysis of the significance profile (SP) of these motifs is suggested as a device to identify the networks design principles (Milo *et al.,* 2004). An SP is the vector of occurrences of the network motifs, which can be thought of as a fingerprint of a network. SP is fruitful mostly when it reveals novel, non-trivial design principles of the underlying network.

In this work we studied the architecture of protein folds as represented in the network of backbone H-bonds (hence other types

*To whom correspondence should be addressed.

of interactions were omitted). We compiled a representative dataset of non-redundant proteins with high-resolution crystal structures. For each structure, we calculated all the backbone H-bonds and searched for significant *network motifs*. The motifs found include the known fold motifs (α-helix, β-sheet, 3_{10} helix, etc.) as well as novel ones. To understand motifs dynamics, we performed MD simulations on a number of proteins. We found that the trajectories preserved both the number of H-bonds, and the major organization of the α-helices and the β-sheets. Yet, we observed differences in motifs that form the surroundings of both the α-helix and the β-sheet. Finally, we provide examples of how the motives may help to analyze protein structures.

2 METHODS

2.1 H-Bonds definition

Each structure is enriched with backbone H-bonds by using BndLst version 1.6 (http://kinemage.biochem.duke.edu/software/utilities.php). The critical step is to correctly position the backbone hydrogen atoms. For this purpose Reduce is used (Word *et al.*, 1999). The accuracy of various methods for positioning hydrogen atoms in protein structures was assessed previously by Forrest and Honig (2005), based on ultra-high resolution structures in which hydrogen positions are determined experimentally (over 1000 hydrogen atoms from seven different structures). Nearly 100% of the tetrahedral NH-type and the planar pNH-type hydrogen atoms were placed within 0.2 Å of the respective experimental atomic positions by most methods, including Reduce. Once the hydrogen atoms are positioned, bonds are assigned efficiently using a spherical probe that is rolled around the van der Waals surface of each atom, and leaves a dot when the probe touches another 'not-covalently-bonded' atom (we used default parameters: probe radius = 0, radius scale factor = 1, C = O carbon VDW scaled by 0.943 to a radius of 1.65 Å).

2.2 Graphs of proteins

Each protein structure (solved by X-ray crystallography) was embedded in a mathematical graph $G=(V,E,C)$ in which the amino acid residues are the vertices V, the backbone interactions are the edges E and C is the edges (bonds) colors: 'black' for a covalent bond, 'thin red' for a single H-bond and 'thick red' for a double H-bond (see examples in Fig. 1). Note, that the 'thick red' here is considered a *different color* than the 'thin red'. The analysis was performed on a representative set of 2521 proteins of known structure (852 561 amino-acid residues), 'culled down' from the PDB (Berman *et al.,* 2000) using a list precompiled by PISCES (Wang and Dunbrack, 2003) to represent all the known structures as of Jan. 2007, such that the (pair-wise) sequence identity is <20%, the resolution is <2.0 Å and the R factor is <0.25.

2.3 Network motifs

A graph $H=(W,F,D)$ is a *subgraph* of $G=(V,E,C)$, if $W\subset V, F\subset E$, and $D\subset C$. It is defined as an *induced subgraph*, if in addition it preserves the following property of the structure of G:$F=E\cap(W\times W)$ (i.e. if an edge of G connects nodes of the subgraph H, the edge itself also exists in the subgraph H). For each network, all the edge-colored induced subgraphs of six nodes were enumerated by the FANMOD algorithm (Wernicke and Rasche, 2006) using full enumeration. Two subgraphs with different edge colors are considered different (see for example motifs β-sheet S10 and S27 in Figure 1, three different motifs which differ only in the 'thick' versus 'thin' red color). FANMOD enumerates the subgraphs by iterating the vertices, and at each step extending on to include subgraphs which were not enumerated earlier. To calculate the probability that a subgraph is a recurrent motif, we use a novel random model described below.

2.4 The random model

To capture the uniqueness of protein graphs, we developed the following random network generator algorithm, given a real protein $\mathrm{Pt}_{\mathrm{real}}$. We first create a 3D self-avoiding random walk on grid points, with a shape of an ellipsoid and the minimal size that envelops $\mathrm{Pt}_{\mathrm{real}}$. For each protein, the procedure is repeated until a self-avoiding walk is obtained. Each point of the walk is a node in the random protein $\mathrm{Pt}_{\mathrm{rand}}$, and we furnish $\mathrm{Pt}_{\mathrm{rand}}$ with edges in three steps. First, a 'black' color (which corresponds to a covalent bond in $\mathrm{Pt}_{\mathrm{real}}$) is automatically added for each two neighboring nodes on the random walk (that is, nodes n and $n+1$). Second, for two nodes of $\mathrm{Pt}_{\mathrm{rand}}$ with distance d in the 3D space, a 'thin red' color is added at random using a biased coin with a probability R, where R is the probability that two nodes in $\mathrm{Pt}_{\mathrm{real}}$ with distance d have a 'red' edge (using normal fit for the edge-distance distribution). Third, we pick at random T 'thin red' edges of $\mathrm{Pt}_{\mathrm{rand}}$ and convert their color to 'thick red', where T is the number of 'thick red' edges in $\mathrm{Pt}_{\mathrm{real}}$. We use this procedure to create one random network per real protein, that preserves the number of nodes, edges, degree distribution, radius of gyration and community structure.

For each subgraph M, we check the null hypothesis that the distribution of M occurrence in real proteins is the same as the distribution of M occurrence in random proteins, using the Kolmogorov–Smirnov test for two samples (explained e.g. in DeGroot, 1975). The probability $P(M)$ with which we can reject the null hypothesis is approximated from the statistic by the *kstest2* implementation of *matlab* version 7.3.0.267 (R2006b). The *occurrence* of M is defined as $<M>=$(# residues in which M occurs)/N, where $N=$total number of residues = 852 561. Note, that we ignore motifs which contain leaves, that is, vertices with at most one edge. The probability P of only eight subgraphs, namely subgraphs #36–#43, is such that $6.2\times10^{-9}\leq P<0.05$. These subgraphs were ignored. For subgraphs #44 and on, $P>0.05$. The 35 motifs analyzed here have a P of $<6.2\times10^{-9}$ and thus a statistical fix for multiple comparisons such as *False Discovery Rate*, has been omitted herein.

2.5 MD simulations

The dynamics of motifs were studied by simulating three proteins for 4 ns using molecular dynamics simulations. The selected proteins were: Lysozyme (pdb 1rfp), SH3 domain (pdb 1srl) and the 434 repressor (pdb 1r69). The simulations were performed at room temperature using the CHARMM (Brooks *et al.*, 1983) package using the charmm27 force field and time step of 2 fs. To explore the sensitivity of the motif stabilities to the details of the force field, each protein system was simulated using explicit and implicit solvent models. Initially, each protein was minimized using 200 steepest-descent steps and 400 adopted basis Newton–Raphson. The studied protein was then placed in a TIP3 water box with a water layer of 20 Å surrounding the proteins and were minimized for additional 500 adopted basis Newton–Raphson. The temperature was equilibrated using 50 ps MD simulation to reach a temperature of 300 K. Constant temperature simulation was collected for 5 ns with dielectric constant of 1 and a 14 Å energy cutoff. The implicit solvent simulations were performed using the Generalized-Born (GB) models. For each trajectory we calculated the number of H-bonds, and the RMSD from the native structure.

3 RESULTS

We compiled a list of 2521 protein structures (see 'Methods' section), for which we calculated the contact map, and further furnished the set of contacts (edges) with colors, to distinguish between covalent interactions of the polypeptide chain ('black' edges), and H-bonds ('red', Fig. 1A). We then retrieved all the subgraphs of six nodes (see 'Discussion' section). To evaluate the statistical significance of each subgraph, we developed a novel random model (detailed in the 'Methods' section). Next, we searched for subgraphs in the random graphs, and calculated the

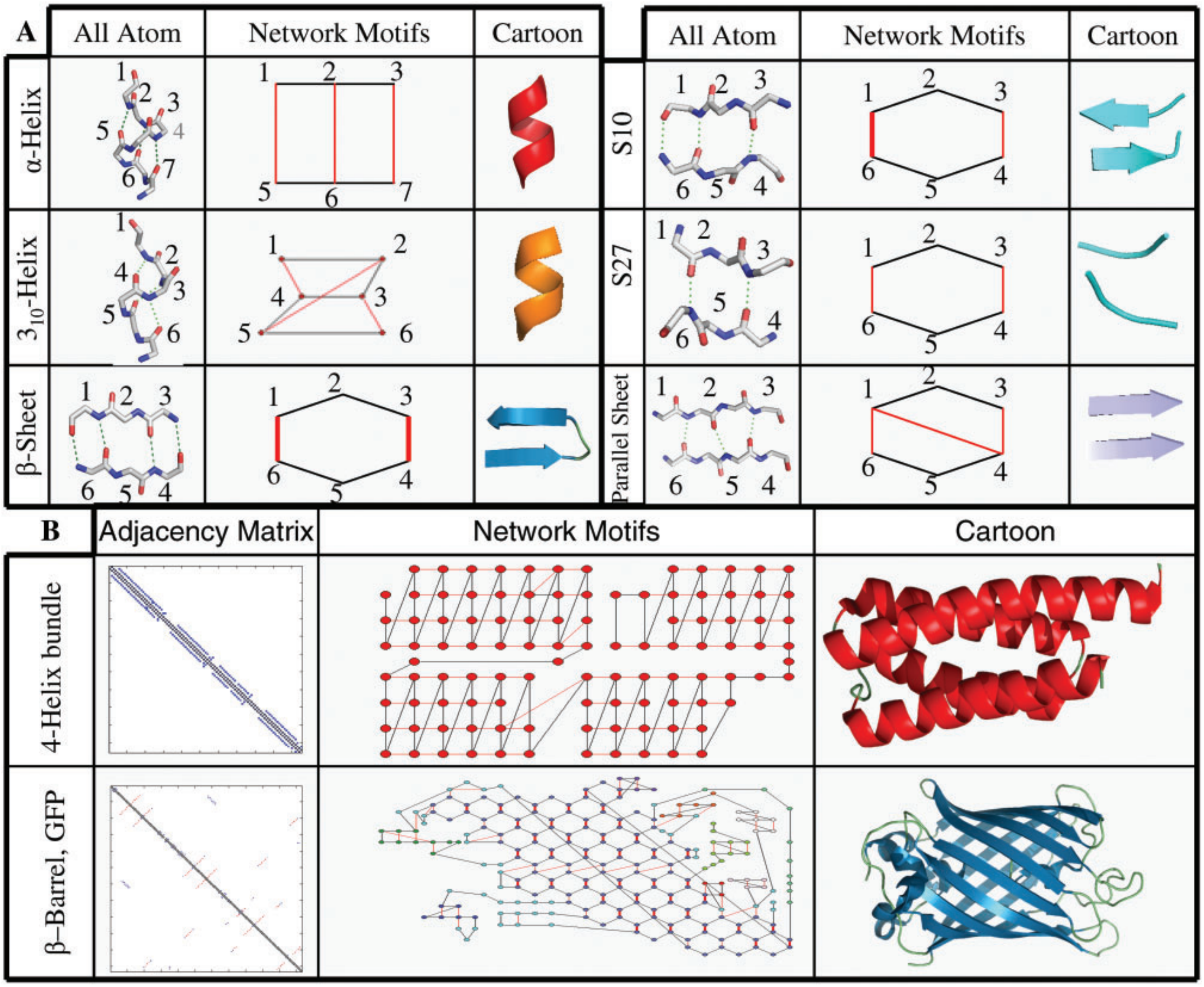

Fig. 1. Proteins described as mathematical graphs. (**A**) Various presentations of the standard motifs: α-helix (H9, top row, caturing a non-countiguous six residues motif, see also Supplementary Fig. S1), 3_{10} Helix (H18) explained by an H-bond of residues n and $n+3$. The '10' stands for the distances in backbone atoms in the chain nitrogen–carbon–carbon (NCC). The standard α-helix is 4_{13}. Below: β-sheet (anti-parallel, S3, S10 and S27) and parallel (S15). A covalent interaction in black, a single H-bond in normal red, and a double H-bond in thick red. See Supplementary Figure S1 for a visualization of all the motifs. (**B**) Examples of a 4-helix bundle (pdb 1tqg, top) and a β-barrel (GFP, pdb 1oxe, bottom) proteins. The left column depicts the contact map as a matrix with covalent bonds and backbone H-bonds. The middle column shows the contact map planar drawing with vertices positions based on the observed motifs: helices are represented by the box-shaped motif H9, while a β–sheet resembles the beehive shape of S3 hexagons. Regions with no motif are displayed in green. The figures were drawn using PyMOL (DeLano, 2002, http://www.pymol.org/).

probability of each subgraph to occur in similar numbers in the random protein network and in real proteins. If this probability is low we consider the subgraph as a *network motif*. Thus, network motifs are backbone H-bond patterns that occur in experimentally determined structures of proteins much more often than in random proteins with similar local connectivity and size.

In the set of 2521 structures, we found 6322 different subgraphs, out of which <43 are *network motifs*. Not surprisingly, the most significant network motifs include the α-helix and the β-sheet (Fig. 1A). For example, α-helix is represented by the box shaped motif number 9 (called H9, 'H' for helix and the number 9 is the position when sorted by probabilities). The same motif is also captured by H6 (see Supplementary Fig. S1). Anti-parallel β-sheet sub-categorizes into S3, S10 and S27 with 4, 3 and 2 H-bonds, respectively (see Fig. 1A, the 'S' stand for Sheet). It is interesting to note the *inverse* correlation between the number of H-bonds in these motifs and the probability to observe them at random, suggesting that unsaturated H-bonds are rare in our dataset, and so the structures used are of high quality. Examples for contact maps of two proteins with mostly helical and sheet structures are given in Figure 1B using both the adjacency matrix and the alternative planar drawing, based on the observed motifs (see 'Discussion' section). A graphical representation of all the motifs is given in Supplementary Figure S1, while the motifs probabilities are depicted

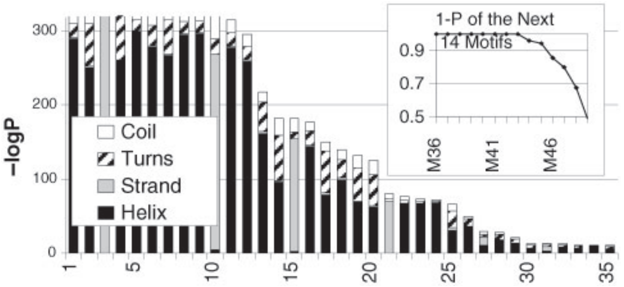

Fig. 2. Motifs probabilities (log-scale) versus DSSP annotations (Kabsch and Sander, 1983). Six-thousand three-hundred and twenty-two different subgraphs exist in the 2521 proteins structures analyzed. For each subgraph, we calculated the probability P that the subgraph is a network motif, and sorted the motifs by P (see 'Methods' section). The first 10 motifs occur with probability $P < 10^{-315}$. (Inset), the probability of the next 14 motifs, using a normal scale. Motifs 36–43 are less significant, with $6.2 \times 10^{-9} \leq P < 0.05$ (see 'Methods' section). For subgraphs #44 and on, $P > 0.05$. For subgraphs #49 and on, $P > 0.5$.

in Figure 2. Sorted by their probability, a clear distinction can be made between the first ten motifs ($P < 10^{-316}$) and the next 25 motifs. The first 10 motifs overlap with the standard α-helix and anti-parallel β-sheet, while the next 25 motifs include other known secondary structure motifs as well as novel ones. For example, motif number 14 (M14) is the Schellman motif (Schellman, 1980). This motif is found in many C-caps of helices (see Fig. 3A for a typical example). Using network motifs analysis, we found that the same motif connects an α-helix and a β-sheet (Fig. 3B), or two β-strands to a sheet (Fig. 3C). S15 and S21 are two alternative representations of the parallel β-sheet. H18 is the 3_{10} helix with occurrence $<\text{H18}>=0.96\%$. Many novel fold motifs were found, including H13, T17 and H22 which are prevalent in helix caps. Figure 3D and 3E visualize H13, called here the B_{10} helix, a bifurcated 3_{10}-like helix. Another set of motifs, namely T2, T7, T20, T26 and T29, appear as a part of a turn, which is found in various surroundings. For example, a turn might connect two beta-strands or two loops. Each one of these four novel motifs represents a different surrounding of a turn, and is also prevalent in helix caps.

To visualize the motifs on protein structures, and compare between different structures in an interactive manner, we created a web tool, protmot (http://bioinfo.weizmann.ac.il/protmot/). In addition to the graphical interface, protmot provides also textual information on the location of all the motifs for further analysis. In Supplementary Figure S2, we provide a case study of how the motif composition of different p21-activated kinases teaches us about the differences between these similar apo and holo enzyme structures. Supplementary Figure S3 shows the apo versus the holo structures of Sir2, which apparently have a RMSD of 12 Å, but show high structural and SP similarity.

SP of homologues protein structures is expected to be similar. Therefore, we were surprised to see that the SP of Deer Hemoglobin (structure solved to 1.98 Å resolution) is highly different from all other mammalian Hemoglobins (Supplementary Fig. S4, compared to Human, Maned wolf and Horse), while the RMSD relative to human is only 0.9 Å (Supplementary Fig. S4A). This suggests that the structure of Deer hemoglobin has local dissimilarities from all the others. Interestingly, a high percentage of this structure is out of the Ramachandran allowed region [Supplementary Fig. S4C, see also Morris *et al.* (1992)]. Apparently, the reason for the discrepancy is a strained backbone of this old structure, resulting in incorrect H-bonds parameters (distance and angle), and hence the motifs are not found although realistically they must be. These examples show that SP reveals local differences between structures even when globally the structures seem to be identical.

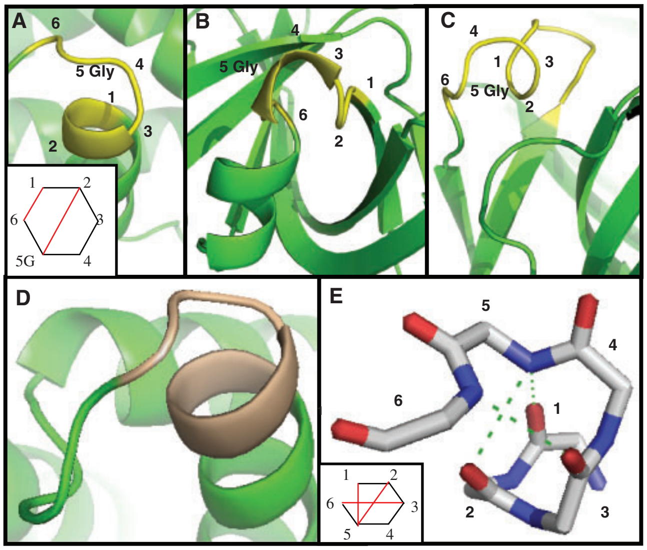

Fig. 3. Various surroundings of the Schellman and the novel B_{10} motifs. (A–C) The Schellman motif, previously considered only in helices caps, is shown here in various surroundings. Note the Glycine preference at position 5. (**A**) Classical helix cap, residues 16–21, pdb 1taf. (**B**) Connecting an α-helix and a β-sheet, residues 50–55, pdb 2d37. (**C**) As a part of a β-sheet, residues 99–109 of pdb 2fd6 (two overlapping instances of the motif, numbering is given only for the second). (**D**) Visualization of the novel B_{10} helix (Motif H13), a bifurcated 3_{10}-like helix, using cartoon; and (**E**) all atoms. The motif is more prevalent than the 3_{10} helix (Fig. 1, occurrence of 0.23 versus 0.2%). Yet, the 3_{10} helix is widely represented in the literature as an alternative helix, due to its 'nice' shape (pdb 1taf residue 62–67).

3.1 SP as a fingerprint of a protein structure

Proteins are dynamic in nature, as is evident by comparing multiple NMR or X-ray solutions of the same structure. SP is a natural tool to analyze these differences, as well as following the time evolution of SP in atomistic MD. The SP is studied globally (count of each motif in the entire protein). Although localization may relate some helical motifs to sequence properties, we are more interested in the global architecture. We followed the pattern of the motifs as a function of time and compared their average population along the trajectories to those found in X-ray and NMR structures for three model proteins: Lysozyme, SH3 and the amino terminal domain of the 434 repressor (Fig. 4). We simulated each protein along 4 ns at room temperature, starting from the crystal structure. During this time frame the global fold did not change. To observe high-resolution variation, we constructed the SP (i.e. the motifs' occurrences vector). To examine the effect of the force-field on the SP, we simulated the proteins using two different water models, Generalized-Born (GB), and explicit solvent. The SP obtained when simulating the

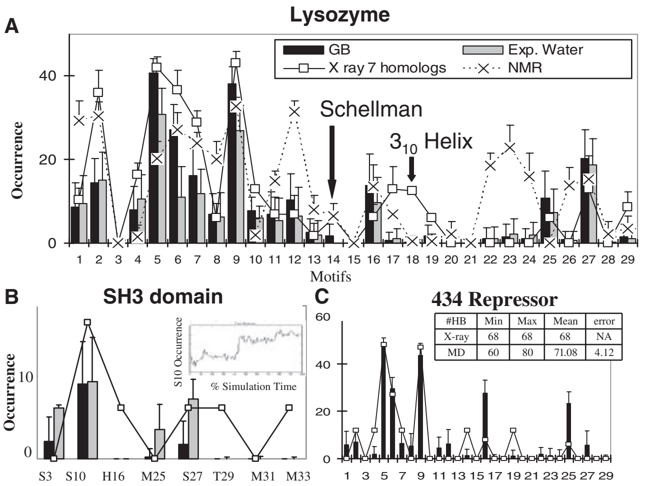

Fig. 4. Significance profiles (SP). Frequencies of motifs in experimental structures (solid line) compared to MD simulation trajectories (bars with error) of 4 ns. High correlation is observed in motifs 1–10, but a significant deviation can be seen for motifs 11–30. (**A**) Lysozyme. X-ray ($n = 7$, average with error) compared to NMR ($n = 50$) and two MD ($n = 100$ snapshots) simulations, using General Born (GB) and explicit water models (see 'Methods' section). (**B**) The NMR structure of SH3 domain: motifs S3, S10 and S27 are hexagons which form a β-sheet (see Fig. 1). These motifs have a rather similar SP in the MD simulation versus the NMR (although somewhat overrepresented in the MDs). However, the less frequent motifs (H16, T29, and M33) do not exist at all in the MDs. The inset shows the time behavior of S10 in the SH3 domain MD simulations. The curve depicts the number of residues in which S10 occurs. After ~45% of the simulation time (equivalent to 1.6 ns), S10 occurrence increases from ~18 to ~24. S3, as shown in Figure 1A, is the most stable β-sheet. The higher occurrence of this motif may compensate for the loss of S16, T29 and M33. (**C**) The 434 repressor.

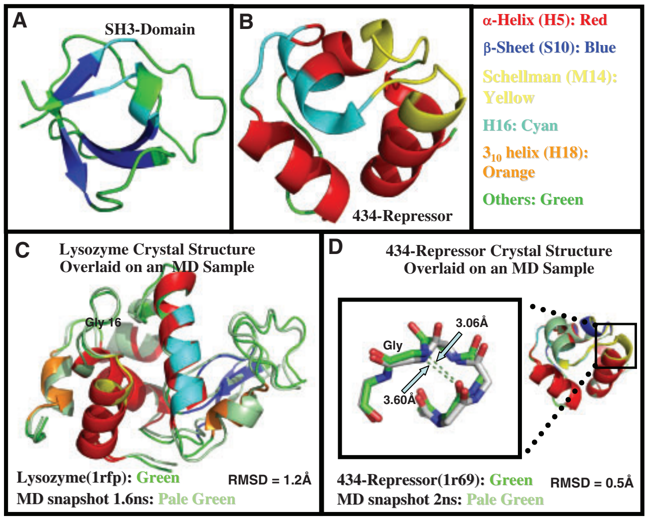
A
SH3-Domain
B
434-Repressor
α-Helix (H5): Red
β-Sheet (S10): Blue
Schellman (M14): Yellow
H16: Cyan
3_{10} helix (H18): Orange
Others: Green
C
Lysozyme Crystal Structure Overlaid on an MD Sample
Gly 16
Lysozyme(1rfp): Green
RMSD = 1.2Å
MD snapshot 1.6ns: Pale Green
D
434-Repressor Crystal Structure Overlaid on an MD Sample
Gly
3.06Å
3.60Å
434-Repressor(1r69): Green
RMSD = 0.5Å
MD snapshot 2ns: Pale Green

Fig. 5. Motifs visualization. Motifs are visualized by color-coding on the three protein structures analyzed in details in this study. (**A**) SH3-domain. (**B**) 434-repressor. (**C**) Lysozyme crystal structure overlaid on an MD sample. (**D**) 434-repressor crystal structure overlaid on an MD sample.

proteins using the GB and the explicit solvent models were within the standard deviation for most motifs (Fig. 4), and therefore we focus hereafter on GB.

The first system studied, Lysozyme, is a helical protein of 129 amino acids (Figs 4A, 5C and 6), in which many motifs are observed beside the α helix, including the 3_{10} helix and the Schellman motif. Moreover, the wealth of available structural data for Lysozyme makes it possible to calculate motif conservation in different crystal forms, as well as to compare them to NMR. Figure 4A shows the SP occurrence in the crystal structures (minimum and maximum of seven crystal structures) versus the MD trajectory (100 conformations sampled along the 4 ns trajectory) and NMR structures (50 minimized models). A high correlation is observed for the first 10 motifs; however, a significant deviation between the three methods is observed for motif number 11 and on. Furthermore, motifs that show a high average correlation do vibrate significantly over time; see for example the α-helix (H5) in Supplementary Figure S5. The second system studied is the SH3 domain, a small β-sheets protein domain that served as a model for numerous structural studies. As can be seen in Figure 4B, H16, T29 and M33 are underrepresented in the MD versus the X-ray structure. These motifs disappeared in the initial minimization step of the simulation. Furthermore, the motifs show a possible cooperativity between them (see 'Discussion' section and Fig. 4B, inset).

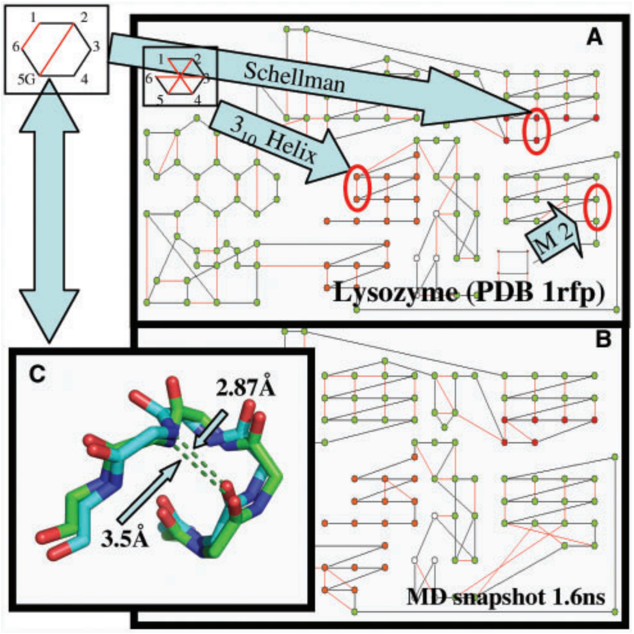

Fig. 6. Lysozyme contact map. (**A**) X-ray crystal structure, compared to (**B**), a snapshot from the MD (using GB) at 1.6 ns. Although the major secondary structure elements are conserved, some H-bonds break (arrows), caused by (**C**) a backbone perturbation of as small as 0.63 Å.

The third system studied is the 434 repressor, a small protein domain of 69 residues, which consists of five short α-helices (H5, red in Fig. 5B). Two of the helices end with the Schellman motif (M14, yellow), and H16 is found in the short 2-turns helix. The average number of H-bonds along the trajectory is similar to that in the X-ray structure (Fig. 4C, inset) and only a small change in the RMSD is observed during the simulation. Furthermore, Figure 4C presents a comparison of the motifs occurrence in the X-ray structure and their average population along the simulation, demonstrating a good correlation for the most common motifs (motifs 1–10), which corresponds to the α-helix and β-sheet. On the other hand, similarly to the previous two systems, a poor correlation is observed between the population of the novel motifs (motif 11 and on) in the MD conformations and the X-ray structure. The lower population of some motifs in the MD simulations is due to their relatively low stabilities (e.g. M14 and M19). Other motifs have short lifetimes (<1 ns) and their population significantly fluctuates in the room temperature simulations. This results in an averaged lower occurrence in comparison to the crystal structure (Supplementary Fig. S6).

4 DISCUSSION

Abstraction of structural data through the use of fold motifs as building blocks is common. These methods use clustering algorithms that are applied to the continuous space of folds. While impressive results were achieved using these approaches, it relies on proper clustering, a process which is not easy to assess (Unger *et al.*, 1989). Moreover, the positions of the protein atoms are in many cases less robust than the interactions they induce. Therefore, inter-residue contact maps (or networks) are likely to be informative by capturing cooperative elements that maintain complex biological architectures.

Networks can be represented either as an adjacency matrix or alternatively as a planar drawing (Fig. 1). Note that the planar drawing is not unique, as the position of each point does not relate to the actual 3D position of the amino acid it represents. Huan *et al.* (2004, 2006) developed and applied an algorithm to mine subgraphs of bio-molecules contact maps (represented as a mathematical graphs). In this manner, the question of defining the boundaries between clusters is reduced to the definition of an interaction, defined based on a distance threshold of 10 Å between non-adjacent C_α atoms. The authors further binned the interactions according to the distance.

Here, we focus on networks of backbone–backbone H-bonds in proteins, and their network motifs (which can be assigned accurately from the structure: see 'Methods' section, Forrest and Honig, 2005). Inter backbone H-bonds are included in the first bin of the previous contact map definitions. However, we suggest that by focusing on validated backbone–backbone H-bonds we can study the general architecture of a protein, and obtain unambiguous raw data (see Fig. 1, and the sharp probability threshold in Fig. 2). The method can be adapted to side chains as well. Here, however, we explore how much only an analysis of backbone hydrogen bonding can elucidate in and between protein structures.

For self-consistency, we limit the motifs to a fixed number of nodes. At least six nodes are needed to capture both α-helix and β-sheet motifs (H9 and S3). More than six nodes may better distinguish between certain turn conformations, such as helical and non-helical turns. However, to raise the number of nodes would significantly increase the complexity of the results.

To check if a certain motif is family-specific, Huan *et al.* (2006) calculated the probability with which they can reject the null hypothesis that the motif is prevalent in two distinct families of structural homology. The randomized entity here is the assignment of structures to families. Here, based on the assumption that important subgraphs occur in high numbers, we draw a different null hypothesis: that a specific subgraph occurs in similar numbers in proteins with experimentally solved structures, and in random, i.e. we check if a motif is overrepresented in proteins structures. To calculate the probability with which we can reject the null hypothesis, we developed a novel random model for proteins. Moreover, we used an algorithm that count exactly all the occurrences of each motif in each network (see 'Methods' section). This is unlike the previous work, which only finds motifs occurring in a high portion of the networks, and hence may overlook motifs which occur in a high number but are limited to a narrow family of proteins. *Network motifs* can simplify the task of planar drawing, as is demonstrated in Figure 1B. Still, one should be aware that network motifs are the fingerprints of a fold, and it is possible for two different network motifs to co-exist in the same fold motif, as is the case for S15 and S21 (parallel β-sheet).

A major strength of the method presented here is the ability to characterize sequence propensity of novel fold motifs, which are otherwise classified as a random coil. In this context, the 35 network motifs found here (Fig. 2 and Supplementary Fig. S1), which include all the known motifs and some novel ones can be studied individually. Surprisingly, analyzing these network motifs using DSSP (Kabsch and Sander, 1983) shows that all the motifs include a high percentage of ordered secondary structure (α-helix or β-sheet or both, see Fig. 2) in addition to some percentage of coil. In other words, every H-bonds network motif has the potential to be embedded in an α-helix or in a β-sheet, and no motif is exclusively related to a random coil. This suggests that knowledge of the local H-bonding pattern is not enough to determine the local fold. Indeed, for certain sequences the secondary structure depends on the global fold and not on its H-bond pattern (Minor and Kim, 1996). It should be noted that some of the random coil (according to DSSP) has no motif attached to it, as their occurrence is not higher than in random.

The helix cap is an extensively studied structure identified from sequence-structure relations as a fold motif (for a review, see Aurora and Rose, 1998). We suggest that network motif analysis provides a way to define helices caps using backbone–backbone hydrogen bonds, which has not been done previously. Harper and Rose (1993) suggested that complete understanding of the fold motifs requires analysis of side-chains. Richardson and Richardson (1998) gave a geometrical definition for helices caps, asserting that a backbone-H-bonds-based definition would be too sensitive to small perturbations. They observed a 33% Glycine propensity at the C-cap of a helix. In fact, C-caps have a few possible forms: 23% are the Schellman motif (Fig. 3A–C) while the rest are 3_{10} helix, the novel B_{10} helix (Fig. 3D–E, see below), and others. While helices ending with the Schellman motif have a Glycine propensity of 66% in position 5 of the motif, the rest of the helices have a Glycine propensity of as low as 10% (see also Nagarajaram *et al.,* 1993). The high Glycine propensity in this motif was shown recently to be due to the ability of Glycine to adopt a positive ϕ/ψ conformation, rather than the enhanced solvation related with the lack of a side chain in Glycine (Bang *et al.,* 2006). We also found that the Schellman *network motif* is prevalent in the surroundings of β-sheets (see Fig. 3A–C).

H18 is the 3_{10} helix (see Fig. 2), which is observed for about 1% of the amino acid residues, and always consists of <2 helical turns. Should this motif be considered as another variant of helix kink, or as a special, though rare sort of a helix? Comparing H18 with other motifs such as B_{10} (Fig. 3D–E, a more prevalent motif that was not documented as a distinct helix type previously, possibly due to its less elegant H-bonding pattern) suggests that α-helices have various fold motifs coexisting at helices caps and kinks. The variation is driven by a bifurcated H-bond between the carbonyl oxygen of residue i and the nitrogens of residues i+3, i+4, giving rise to motifs such as H13, M14, H18 and others. While bifurcated H-bonds have been previously observed in helices (Niimura, 2001; Richardson, 1981, 2004–2006), their high prevalence shown here is unforeseen.

SP is a powerful tool to compare structures of high similarity. RMSD of 0.5 Å is usually considered to be within the experimental fluctuations of X-ray structures. However, a distance change of 0.5 Å causes an H-bond to break. Unlike RMSD, SP analysis makes it possible to distinguish between concerted movements that do not affect bond patterning and specific movements that do. For example, Figure 4C shows that the Schellman motif (M14) is poorly populated in the MD simulation of the 434 repressor. Figure 5D reveals that the short life time of this motif is due to the break of a single H-bond occurring close to the start of the MD simulation, in a place which otherwise seems to be identical to the X-ray structure. In a second example, a snapshot at 1.6 ns of the MD simulation of Lysozyme shows a structure that is almost identical to the X-ray structure (Fig. 5C). However, the deviation in the SP (Fig. 4A, M14 and H18) is explained by the break of a small number of H-bonds in significant positions. Figure 6 compares the crystal structure of Lysozyme (A), to a snapshot from the MD at 1.6 ns (B). Although the major fold is conserved (reflected by a small RMSD of 1.2 Å), the elimination of some H-bonds results in the disappearance of a few motifs. Another inherent problem of comparing proteins using RMSD relates to the global nature of this method, which causes a hinge movement to have a tremendous effect. Although SP is presented here globally, we calculate the motifs occurrence locally for each position, and hence two proteins with similar interaction will have similar SP, despite hinge-like movements. The SP during the simulation is far from being static. Motifs are broken and formed (inset in Fig. 4B and Supplementary Figs S5–S7), and deviate away from the starting crystal structure. Interestingly, the SP of crystal structures shows high self-consistency (among various crystal solutions of the same structure) but a significant deviation from that of the different NMR datasets (Fig. 4A, motifs number 1, 5, 12, 18, 22–24), possibly due to different energy minimization potentials. The deviation is predominantly in motifs 11 and on, where motifs are highly significant, appearing in the proximity of standard α-helices and β-sheets. In a few cases, motifs break in the initial dynamics simulation phase, and do not reappear in the 4 ns simulation, including the Schellman motif (M14) at the 434 repressor (also underrepresented in Lysozyme), 3_{10} helix (H18) in Lysozyme, and H16, T29 and M33 of the SH3 domain. Moreover, motifs anti-correlation was observed between H6 and S27 in Lysozyme (Supplementary Fig. S7), and also in the SH3 domain. Another two examples for using SP to compare proteins are given in Supplementary Figure S2 (PAK) and S4 (Hemoglobin). Here, SP reveals significant local differences between homologues proteins. In the case of PAK, this explains the structural basis for the observed biological differences between the related proteins.

In summary, we applied here a method from graph theory to the vast amount of structural data available to understand the high-order patterns prevalent in bio-molecules. Exploring the repertoire of contact map motifs allows for the unsupervized discovery of new fold patterns that are no longer limited to a continuous stretch. The advantage is that subsequent mining of structural data can base on a simpler unified framework, eliminating the need for separated analysis for helices and for sheets. SP analysis is suggested as a novel scheme to study 3D structures, and their dynamics and folding trajectories, where a large number of snapshots have to be compared. Furthermore, the method can be used to track protein folding through the development of native motifs. For example, one may use this method to investigate which motifs are formed already during the early stages of folding, and how folding is being developed.

ACKNOWLEDGEMENTS

The authors are grateful to Merav Parter and Dr Nadav Kashtan for the development and implementation of the random model; Dr Avi Mayo, Dr Maria Krisch, Dr Yaki Setty and Eli Steinberger for critical reading of the manuscript; Prof. Eytan Domany and Prof. Avigdor Scherz and Dr Orly Dym for helpful discussions.

Funding: Kimmelman Center for Macromolecular Assemblies and the Center for Complexity Science [to Y.L.]; and MINERVA [8525]. Y.L. is the incumbent of the Lilian and George Lyttle Career Development Chair.

Conflict of Interest: none declared.

REFERENCES

Aurora,R. and Rose,G.D. (1998) Helix capping. *Protein Sci.,* **7,** 21–38.

Bang,D. *et al*. (2006) Dissecting the energetics of protein alpha-helix C-cap termination through chemical protein synthesis. *Nat. Chem. Biol.,* **2**, 139–143.

Berman,H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.,* **28**, 235–242.

de Brevern,A.G. *et al.* (2000) Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins,* **41**, 271–287.

Brooks,B.R. *et al.* (1983) CHARMM: a program for macromolecular energy, minimization and dynamic calculation. *J. Comp. Chem.,* **4**, 187–217.

Bystroff,C. and Baker,D. (1998) Prediction of local structure in proteins usinag a library of sequence-structure motifs. *J. Mol. Biol.,* **281**, 565–577.

DeGroot,H. (1975) *Probability and Statistics*, Addison-Wesley, MA, pp. 468–469.

Dokholyan,N.V. *et al. (*2002) Topological determinants of protein folding. *Proc. Natl Acad. Sci. USA,* **99**, 8637–8641**.**

Forrest,R. and Honig,B. (2005) An assessment of the accuracy of methods for predicting hydrogen positions in protein structures. *Proteins,* **61**, 296–309.

Friesner,R.A. (2005) Ab initio quantum chemistry: methodology and applications. *Proc. Natl Acad. Sci. USA,* **102**, 6648–6653.

Gerstein,M. and Chothia,C. (1996) Packing at the protein-water interface. *Proc. Natl Acad. Sci. USA,* **93,** 10167–10172.

Harper,E.T. and Rose,G.D. (1993) Helix stop signals in proteins and peptides: the capping box. *Biochemistry,* **32,** 7605–7609.

Huan,J. *et al.* (2004) Mining protein family specific residue packing patterns from protein structure graphs. *RECOMB'04,* pp. 308–315.

Huan,J. *et al.* (2006) Distance-based identification of structure motifs in proteins using constrained frequent subgraph mining. *Proc. LSS Comp. Sys. Bioinfor. Conf. CSB,* pp. 227–238.

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

Kolodny,R. *et al.* (2002) Small libraries of protein fragments model native protein structures accurately. *J. Mol. Biol.,* **323**, 297–307.

Micheletti,C. *et al.* (2000) Recurrent oligomers in proteins: an optimal scheme reconciling accurate and concise backbone representations in automated folding and design studies. *Proteins,* **40**, 662–674.

Milo,R. *et al.* (2004) Superfamilies of evolved and designed networks. *Science,* **303**, 1538–1542.

Minor,D.L. Jr and Kim,P.S. (1996) Context-dependent secondary structure formation of a designed protein sequence. *Nature,* **380**, 730–734.

Morris,A.L. *et al.* (1992) Stereochemical quality of protein structure coordinates. *Proteins*, **12**, 345–364.

Nagarajaram,H.A. *et al*. (1993) Termination of right handed helices in proteins by residues in left handed helical conformations. *FEBS Lett.,* **321**, 79–83.

Niimura,N. (2001) Neutron protein crystallography in JAERI. *J. Phys. Soc. Jpn.,* **70** (Suppl. 1), 396.

Oliva,B. *et al.* (1997) An automated classification of the structure of protein loops. *J. Mol. Biol.,* **266**, 814–830.

Rahat,O. *et al.* (2008) Cluster conservation as a novel tool for studying protein-protein interactions evolution. *Proteins,* **71**, 621–630.

Raveh,B. *et al.* (2007) Rediscovering secondary structures as network motifs-an unsupervised learning approach. *Bioinformatics,* **23,** e163–e169.

Reichmann,D. *et al.* (2005) The modular architecture of protein-protein binding interfaces. *Proc. Natl Acad. Sci. USA,* **102**, 57–62.

Richardson,J.S. and Richardson,D.C. (1988) Amino acid preferences for specific locations at the ends of alpha helices. *Science,* **240**, 1648–1652.

Richardson,J.S. (1981, 2004–2006) The anatomy and taxonomy of protein structure. In *Advances in Protein Chemistry*, Academic Press, New York, pp. 167–339.

Schellman,C. (1980) The αL conformation at the ends of helices. In Jaenicke R. (ed) *Protein Folding.* Elsevier, North Holland, New York, pp. 53–61.

Unger,R. *et al.* (1989) A 3D building blocks approach to analyzing and predicting structure of proteins. *Proteins,* **5**, 355–373.

Wang,G. and Dunbrack,R.L. Jr (2003) PISCES: a protein sequence culling server. *Bioinformatics,* **19**, 1589–1591.

Wang,X. *et al.* (2007) Mining RNA tertiary motifs with structure graphs. *Nineteenth International Conference on Science and Statistical Database Management (SSDBM 2007)*, p. 31.

Wernicke,S. and Rasche,F. (2006) FANMOD: a tool for fast network motif detection. *Bioinformatics,* **22**, 1152–1153.

Wintjens,R.T. *et al.* (1996) Automatic classification and analysis of alpha alpha-turn motifs in proteins. *J. Mol. Biol.,* **255**, 235–253.

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**, 1711–1745.