Understanding Hydrogen-Bond Patterns in Proteins using a Novel Statistical Model

Ofer Rahat[1], Uri Alon[2], Yaakov Levy[3*] and Gideon Schreiber[1*]

[1] Department of Biological Chemistry
[2] Department of Molecular Cell Biology
[3] Department of Structural Biology, Weizmann Institute of Science, Rehovot 76100, Israel

Correspondence should be addressed to Gideon.Schreiber@weizmann.ac.il; Koby.Levy@weizmann.ac.il.

Abbreviations: MD – Molecular Dynamics. SP – Significance Profile. H-Bonds – Hydrogen Bonds. RMSD – Root Mean Square Deviation. M<u>X</u>* – Motif Number *<u>X</u>*. ns-nanoseconds.*

Proteins are built from basic structural elements and their systematic characterization is of interest. Searching for recurring patterns in protein contact maps, we found several network motifs, patterns that occur more frequently in experimentally determined protein contact maps than in randomized contact maps with the same properties. Some of these network motifs correspond to sub-structures of alpha helices, including topologies not previously recognized in this context. Other motifs characterize beta-sheets, again some of which appear to be novel. This topological characterization of patterns serves as a tool to characterize proteins, and to reveal a high detailed differences map for comparing protein structures solved by X-ray crystallography, NMR and molecular dynamics (MD) simulations. Both NMR and MD show small but consistent differences from the crystal structures of the same proteins, possibly due to the pair-wise energy potentials used. Network motifs analysis can serve as a base for many-body energy statistical energy potential, and suggests a dictionary of basic elements of which protein secondary structure is made.

Introduction

Embedding a continuous entity such as a protein structure in a discrete model can be done in many ways. Bystroff and Baker [1] 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.* [2] as well as others [3-7] 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.

Secondary structures are key fold motifs, both for the process of folding and in stabilizing the structure [8] suggesting a possible resolution to the Levinthal paradox [9] by reducing the sample space. Currently, secondary structure predictions from sequence have a success rate of about 80% (being stable at these rates for over a decade now) using algorithms like ASSAM [10] and SPASM[11]. Interestingly, also the assignment of secondary structure for a given solved structure is not absolute, evidenced by an agreement rate of about 80% between various algorithms such as DSSP [12].

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. Recently, we showed that contact maps of proteins are modular [13], [14] and encapsulate the information necessary to detect the secondary structure [15]. Contact map based abstraction has one clear advantage: discretizing takes place at an earlier stage, that of atomic contacts, for which physics is better understood. A widely used scheme of systems biology suggests that networks are made up of a small set of recurring patterns, called Network Motifs. Further, analysis of the significance profile of these motifs is suggested as a device to identify the networks design principles [16]. A significance profile (SP) is the vector of occurrences of the network motifs, which can be thought of as a fingerprint of a network. However, SP is fruitful only to the extent to which it reveals novel, non-trivial design principles of the underlying network.

In this work we compiled a representative dataset of non redundant proteins with high resolution crystal structure. Each protein was embedded in a mathematical graph in which the amino acid residues are the vertices and the backbone interactions are the edges (a contact map), in which we searched for network motifs. The motifs found include the known fold motifs (α-helix, β-sheet, 3_{10} helix, etc) as well as novel ones, suggesting a novel framework to study sequence-structure correlations. To understand the dynamics of the motifs, 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. These findings suggest that unique cooperativity patterns exist in proteins, patterns that are weakly captured by the force-field potentials used.

Results

We compiled a list of 2521 protein structures (see Methods), 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. To evaluate the statistical significance of each subgraph, we developed a novel random model for proteins. This model generates self-avoiding chain with the same length and radius of gyration of the real protein, with a contact map that preserves the number of links and the degree frequency of the original proteins contact map. Next, we searched for subgraphs in these random networks, and calculated the 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 (see Methods). Thus, network motifs are patterns that occur in real proteins much more often then in random proteins with similar local connectivity and size. The present results are a first glimpse at network motifs in proteins and more stringent random model may further refine the results.

Not surprisingly, the ten most significant network motifs include the α helix and the β-sheet (Fig. 1A**)**. Examples for contact maps of two proteins with mostly helical and sheet structures are given in Fig. 1B using both the adjacency matrix and the alternative planar drawing, based on the observed motifs (see Discussion). A graphical representation of all the motifs is given in SI Fig. 6, while the motifs probabilities are depicted in Fig. 2. In this figure the 35 significant motifs are shown 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, motifs are sorted by probability) is the Schellman motif ([25], Fig. 2), which appears in many C-caps of helices, but we found it also as a network motif in the surroundings of β-sheets (see Discussion). M15 and M21 are two alternative representations of the parallel β-sheet. M18 is the 3_{10} helix with occurrence <M18>=0.96%. Many novel fold motifs were found, including M13, M17 and M22 which are prevalent in helix caps and M2, M12, M16 and M29 which represent various surroundings of a turn, in addition of being also prevalent in helix caps. A sub-categorization of the anti-parallel β-sheet includes M3, M10 and M27 with 4, 3 and 2 H-bonds, respectively. It is interesting to note the inverse correlation between the number of H-bonds in these motifs and the probability to observe them in random.

Motifs conservation along an MD simulation. To understand the dynamics of the motifs and their cooperation in maintaining the structure, we studied the time evolution of SP in atomistic MD simulations. 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: the amino terminal domain of the 434 Repressor, Lysozyme, and an SH3 domain (Fig. 3). 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)(Fig. 4). For each protein, SP is compared to the average SP along the simulation.

The 434 repressor protein is a small 69 residues domain, which consists of five short α helices (M5, red in Fig. 3B). Two of the helices end with the Schellman motif (M14, yellow), and M16 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. 4A inset) and only a small change in RMSD is observed during the simulation. Furthermore, Fig. 4A presents a comparison of the SP occurrence of the X-ray structure and their average population along the simulation. High similarity is observed for the most common motifs (M1 to M10) corresponding to the α helix and β-sheet. On the other hand, a poor correlation is observed between the population of the novel motifs (M11 to M30) 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. Accordingly, a few motifs have short life time (< 1ns) and their population significantly fluctuates at the room temperatures simulations. This results in an averaged lower occurrence in comparison to the crystal structure (SI Fig. 7).

The second system studied, Lysozyme, is a larger helical protein (129 amino acids) (Fig. 3C, 4B), in which many more motifs are observed, including the 3_{10} helix and the Schellman motif. For Lysozyme, a wealth of available crystallographic data made it possible to calculate motif conservation in different crystal forms, as well as to compare their occurrence in the NMR models. Fig. 4B shows the SP occurrence in the Crystal structures (minimum and maximum of 7 crystal structures) vs. the MD trajectory (100 conformations sampled along the 4ns trajectory) and NMR (50 minimized models). Again, a high correlation is observed for the first 10 motifs, however, a significant deviation is observed for M11 to M30 between the three methods. Furthermore, motifs that show high average correlation do vibrate significantly over time; see for example M5 in SI Fig. 8. The third system studied is the SH3 domain, a small β-sheets protein domain which served as a model for numerous structural studies. As can be seen in Figure 4C, M16, M29 and M33 are underrepresented in the MD vs. the x-ray structure. These motifs disappear already in the initial minimization step of the simulation. The motifs show a possible cooperativity (see discussion, Fig. 4C inset, and SI Fig. 9).

Discussion

Abstraction of structural data is essential, since a researcher who is not an expert in the intricacies of structural biology may be overwhelmed by the thousands of details of the 'All-Atom' visualization scheme. Moreover, the positions of the protein atoms are in many cases less robust than the interactions they induce. Therefore, inter-residues 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 planar drawing (Fig. 1). The planar draw is not unique, as the position of each point does not relate to the actual 3D position of the amino acid it represents. Network motifs can simplify the task of planar drawing, as is demonstrated in Fig. 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 M15 and M21 (parallel β-sheet).

Secondary structure prediction algorithms show a very high success rate for core regions, predicted to be either α-helix or β-sheet. On the other hand, the prediction is poor for about 15% of the residues. Some attempts were made previously to characterize sequence propensity of novel fold motifs, which might be classified currently as a random coil. In this context, the 35 network motifs found here (Fig. 2 and SI Fig. 6) which include all the known motifs and some novel ones, can be studied individually. Surprisingly, analyzing these network motifs using DSSP [12] shows that all the motifs include a high percentage of ordered secondary structure (α-helix or β-sheet or both, see the bar colors of Fig. 2) in addition to some percentage of coil. In other words, every recurrent pattern of H-bonds 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-bonds 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 [21].

One example for sequence-structure relations derived from fold motifs is the helix-cap, which was extensively studied previously (for a review see [22]). It was suggested that a complete understanding of the fold motifs requires analysis of the side-chains [23] but this aspect is out of the scope of our current work. Richardson and Richardson [24] adopted the geometrical definition for helices caps, asserting that backbone-H-bonds-based definition is too sensitive for small perturbations. This sensitivity is related to the fact that most protein structures are solved at a resolution of ≥1.2 Å and hydrogen atoms have to be inferred, which introduces ambiguity. Here we suggest that network motifs analysis provides a framework to overcome this ambiguity, in the following way. A certain fold motif may have a different pattern of H-bonds, which depend on the H-bond definition. However, the pattern should be the same in all the occurrences of the motif. If this is the case, different assignment methods will give essentially the same motifs with the same sequence propensity for each position of the motif, albeit a different pattern of H-bonds. The analysis of Richardson [24] resulted in sequence propensities for helix caps (most notably a **33%** Glycine propensity at the C-cap of a helix). Using motif analysis, a more detailed understanding of this phenomenon was obtained by dividing C-caps into the following two different forms. About 23% of the helices end with the Schellman motif, while the rest end with motifs such as the 3_{10} helix, M13, and others. For helices ending with the Schellman motif there is a high 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 [26]). 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 [27]. Furthermore, visual inspection of the Schellman network motif revealed that it is prevalent in the surroundings of β-sheets as well.

M18 is the 3_{10} helix (see Fig. 2), which is observed for about 1% of the amino acid residues, and always consists of less than 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 M18 with other motifs such as M13 (a more prevalent motif that was not documented as a distinct helix type previously, possibly due to its less elegant H-bonds pattern) suggests that α helices have various fold motifs coexisting at helices caps and kinks. The variation is driven by bipolarity between the carbonyl oxygen of residue i and the nitrogens of residues i+3, i+4, giving rise to such motifs as M13, M14, M18 and others.

SP is a powerful tool to compare structures of high similarity. RMSD of 0.5Å is usually considered to be within the experimental fluctuations of X-ray structures. However, a distance change of 0.5Å causes an H-bond to break. SP analysis makes it possible to distinguish between concerted movements that do not affect bond patterning to specific movements that do, independently on the resulting RMSD between the structures. For example, Fig. 4A shows that the Schellman motif (M14) is poorly populated in the MD simulation of the 434 repressor. Fig. 3D reveals that the short life time of this motif is due to a 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. 3C**)**. However, the deviation in the SP (Fig. 4B, M14 and M18) is explained by a break of a small number of H-bonds in significant positions. Fig. 5 compares the contact map of Lysozyme crystal structure (**A**), to a snapshot from the MD at 1.6 ns (**B**). Although the major fold is conserved (reflected by small RMSD of 0.63Å), breaking of some H-Bonds results in eliminating a few motifs.

The SP during the simulation is far from being static. Motifs are broken and formed (SI Figures 7 and 8), and deviate away from the starting X-ray structure. Interestingly, the SP of X-ray structures show also a deviation from that of the different NMR solutions (Figure 4B, motifs 1, 5, 12, 18, 22-24), possibly due to different energy minimization potentials. The deviation is predominantly in M11-M30, a region that include motifs which are highly significant in the dataset of representative proteins of known structures, and appear frequently 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 show again in the 4 ns simulation, including the Schellman Motif (M14) at the 434 repressor (also underrepresented in Lysozyme), 3_{10} helix (M18) in Lysozyme, and M16, M29 and M33 of the SH3 domain. An anti-correlation between motifs was observed between M6 and M27 in Lysozyme (SI Fig. 9), and also in the SH3 domain. In the latter, the disappearance of certain motifs (M16, M29, M33) may give rise to a later compensation by a higher occurrence of M10, after a delay of 1.6 ns (equivalent to 45% simulation time, see Fig. 4C Inset).

The relatively poor stability of some motifs in the MD simulations, particularly above M10, may suggest that some motifs are inherently highly dynamic but their weak population is sufficient to retain the protein fold. It is likely, that using different force fields may result with different motifs stability and change in the SP. One might conjecture that the underrepresentation of high-order structural cooperativity patterns in the simulations originates from the pair-wise energy functions used in the current available force-fields. Many-body potentials such as [28] ,[29],[30] come with the cost of high computational complexity, while the exact form of the electronic Schrödinger equation is a problem with an exponential computational complexity, and hence, a brute force solution is intractable [31]. We suggest that a statistical-based energy potential that takes into account many-body cooperativity patterns, but does not exhaustively traverse all the possible ones may solve the problem. For a given conformation, one can enumerate the motifs and introduce a new cooperativity term $\sum_{M \in Motifs} Y(M)$ (in addition to the usual terms such as van der Waals and electrostatic contribution) in order to bias the simulation towards the desired SP. For example one may use for Y a screened Coulombic (Yukawa) potential ([32],[33]) defined below. Given a motif M that has m_M H-bonds we set $r_1 ... r_{m_M}$ to be the distances between the oxygen and the hydrogen of each H-bond. Theoretically m_M can be as high as 30, but for the 35 motifs found here we always have $m_M \leq 5$ (A similar consideration applies also to the H-bond angles, but is not covered here, see also [34]). To satisfy the motif, one has to minimize $E(M) = \sum_{i=1}^{m_M} (r_i - r)^2$, where r is the ideal H-bond distance (usually around 1.9Å). The suggested smooth term is a sum over all the known motifs of

$$Y(M) = P_M \frac{e^{-\sqrt{E(M)}}}{\sqrt{E(M)}} \quad (1)$$

where P_M is the probability of motif M (see SI Fig. 10). The rationale here is to preserve observed motifs but at the same time not to freeze a specific motif-related conformation. To estimate the time complexity (that is, the running time as a function of the input size) of the algorithm, one has first to observe that the bottleneck is enumerating the motifs for the contact map of a specific iteration. This depends upon the motifs algorithm in use, but in the worse case the exhaustive enumeration algorithm is polynomial in the number of residues, as appose to the exponential time of evaluating all the possible dependencies of the protein atoms. Further, it is our experience that motifs enumeration is much faster than energy minimization. We suggest that by using motifs as structural constrains one may obtain dynamic simulations that better represent crystal structures.

Summary

Estimation of the free energy gain of protein folding is a difficult task because of its small net value of 5-10 kcal/mol (about the energy change related to the formation of one or two H-bonds). Hence, detailed understanding of the H-bond cooperativity cannot be achieved directly. Here, we applied a method from graph theory to the vast amount of structural data available to understand the high-order patterns prevalent in bio-molecules. The problem of protein structure prediction might be reduced to a problem of tessellation of the network motifs, the known α helix and β-sheet as well as the other motifs. In this sense, exploring the repertoire of contact map motifs makes it possible to understand secondary structure as a key folding step. Further, it allows for the unsupervised discovery of new fold patterns which are no longer limited to a continuous stretch, and may unify the two major known motifs, helices and sheets, under one framework (each sheet is made out of a few non-continuous strands).

Network motif analysis may also be applied in the future to address questions of function, e.g., prediction of enzymatic-cleft location, metal-binding sites, and protein-protein interfaces. In this context, it might be useful to add edges of different colors for other sorts of non-covalent interactions ($\pi-\pi$ interactions, salt bridges, etc.) and a different vertex color for the various amino acid types (hydrophobic, bulky, etc.). Yet another interesting application is structure motifs of RNA and DNA. Our results may suggest that two ingredients are important for these analyses: first, different edge colors are needed to build up a large repertoire of motifs, and secondly, an appropriate random model to separate the important motifs from the noise.

Methods

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, and the backbone interactions are the edges *E*, similarly to [15]. Backbone interactions can be either peptide covalent bond or H-bond. We extract H-bonds by using BndLst (v.1.6) with default parameters, based on the tool Reduce [17]. Here (unlike [15]) we introduce different edge colors, based on the type of interaction. For each interaction represented by an edge $(v,u) \in E$ where $v,u \in V$ are amino acid residues, we define the color of the edge to 'black' if it is a covalent bond, 'thin red' if it is a single H-bonds and 'thick red' color if it is a double H-bond (see Fig. 1 for an example). The analysis was performed on a representative set of 2,521 proteins of known structure (852,561 amino-acid residues), 'culled down' from the PDB [18] using a list precompiled by PISCES [19] to represent all the known structures as of Jan 2007, such that the (pair-wise) sequence identity is <20%, the resolution is <2.0Å, and the R factor is <0.25.

Network Motifs. For each network, all the edge-colored subgraphs of six nodes were enumerated by the FANMOD [20] algorithm, using full enumeration. FANMOD enumerates the subgraphs by iterating the vertices, and at each step extending on to include subgraph which were not enumerated earlier. To calculate the probability of each subgraph to be a recurrent motif, we use a novel random model, describe below.

The Random Model. Networks of proteins, as defined above, have geometrical properties. Notably, the network can be mapped onto a 3D space such that the distribution of adjacent-node distances is normal (the distance between the center of mass of two H-bonded residues peaks around 5Å). To capture this feature we developed the following random network generator algorithm, given a real protein Pt_{real}. We first create a 3D self-avoiding random walk on grid points, restricted by the minimal ellipsoid which envelops Pt_{real}. Each point of the walk is a node in the random protein Pt_{rand}, and we furnish Pt_{rand} with edges in three steps. First, a 'black' color (which corresponds to a covalent bond in Pt_{real}) is automatically added for each two neighboring nodes on the random walk. Second, for two nodes of Pt_{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 Pt_{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 Pt_{rand} and convert their color to 'thick red', where *T* is the number of 'thick red' edges in Pt_{real}. The random network preserves the number of nodes, edges, degree distribution and radius of gyration.

For each subgraph M, we first calculate the distribution of the number of occurrences of M in proteins in the real and in the random datasets. We then apply the Kolmogorov-Smirnov test to calculate the probability that these two distributions are the same. P(M), the *probability of finding M at random* , is defined as the results of this test. The *occurrence* of M is defined as <M> = (#residues in which M occurs)/N, where N = total number of residues = 852,561. Note: we ignore motifs which contain leafs, that is vertices with at most one edge. The probability of only 5 subgraphs fall in the twilight zone of $8.7*10^{-7} < P \leq 0.05$ (it is a twilight zone since we need to correct for multiple tests). Only another four subgraph have probability of $6.2*10^{-9} < P < 8.7*10^{-7}$. We ignore this 9 subgraphs, and define *Motif* as a subgraph with probability $P < 6.2*10^{-9}$ (a total of 35 motifs exists).

MD Simulations. The dynamics of motifs was studied by simulating three proteins for 4 ns using molecular dynamics simulations. The selected proteins are: SH3 domain (pdb 1srl), lysozyme (pdb 1rfp), and the 434 repressor (pdb 1r69). The simulations were performed at room temperature using the CHARMM [35] 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 two different implicit solvent models: distance dependent dielectric constant and the Generalized-Born (GB)[36] models. In the distance dependent dielectric constant models we tested the motifs dynamics using dielectric constants of 0.5, 5, 50, 500, 5000, and 10,000. One expects that at low value of dielectric constant the H-bonds will be very dominant and therefore the motifs will be highly stable. At high value of dielectric constant, on the other hand, the motifs are expected to be very weak. For each trajectory we calculated the number of H-bonds, and the RMSD from the native structure. For the distance dependent dielectric constant models for the solvent, a dielectric constant of 0.5 (though not physiological) captured best the protein environment. For example, the mean number of H-bonds is 22.4±4.2 for the SH3 domain MD, compared to 24 H-bonds in the NMR structure, and RMSD of

2±0.05 from it. However, the GB model captured better the protein environment, and without a need to adjust for the dielectric constant. The figures presented are based on the GB potential.

References

[1] Bystroff C, Baker D (1998) *J. Mol. Biol.* **281** 565-77.

[2] Unger R, Harel D, Wherland S, Sussman JL (1989). *Proteins: Struct. Funct. Genet.* **5** 355-373.

[3] Kolodny R, Koehl P, Guibas L, Levitt M (2002) *JMB* **323** 297-307.

[4] Micheletti C, Seno F, Maritan A (2000) *Proteins: Struct. Funct. Genet.* **40** 662–674.

[5] de Braven AG, Etchebest C, Hazout S (2000) *Proteins: Struct. Funct. Genet.* **41** 271–287.

[6] Wintjens RT, Rooman MJ, Wodak SJ (1996) *J. Mol. Biol.* **255** 235–253.

[7] Oliva B, Bates PA (1997) *J. Mol. Biol.* **266** 814–830.

[8] Rose GD, Fleming JF, Banavar JR and Maritan A (2006) *Proc. Natl. Acad. Sci. USA* **103.45** 16623-33.

[9] Levinthal C (1969) Mössbauer Spectroscopy in Biological Systems, eds Debrunner P, Tsibris JCM, Münck E (Univ of Illinois Press, Urbana), pp 22-24.

[10] Artymiuk PJ, Poirrette AR, Grindley HM, Rice DW, Willett P (1994) *J. Mol. Biol.* **243:2** 327-344.

[11] Kleywegt GJ (1999) *J. Mol. Biol.* **285:4** 1887-1897.

[12] Kabsch W, Sander C (1983) *Biopolymers* **22** 2577-2637.

[13] Reichmann D, Rahat O, Meged R, Dym O, Schreiber G (2005) *Proc Natl. Acad. Sci. USA* **102(1)**:57-62.

[14] Rahat O, Yitzhaki A, Schreiber G (2007) *Proteins* 2007 Oct 30.

[15] Raveh B, Rahat O, Basri R, Schreiber G (2007) *Bioinformatics* **23(2):** 163-169.

[16] Milo R, Itzkovitz S, Kashtan N, Levitt R, Shen-Orr S, Ayzenshtat I, Sheffer M and Alon U (2004) *Science* , **303** :1538-42.

[17] Word JM et al. (1999) *J. Mol. Biol.* **285:** 1711-1745.

[18] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) *Nucleic Acids Research* , **28** :235-242.

[19] Wang G, Dunbrack RL Jr. (2003) *Bioinformatics* **19**:1589-1591.

[20] Wernicke S, Rasche F (2006) *Bioinformatics* **22**:1152-3.

[21] Minor DL, Kim PS (1996) Nature **380**:730-734.

[22] Aurora R, Rose GD (1998) *Prorein Science* **7**:21-38.

[23] Harper ET, Rose GD (1993) *Biochemistry* **32**:7605-7609.

[24] Richardson JS, Richardson DC (1988) *Science* **240** :1648-1652.

[25] Schellman C (1980) Jaenicke R, ed. Protein folding. New York: Elsevier/North Holland. pp 53-61.

[26] Nagarajaram HA, Sowdhamini R, Ramakrishnan C, Balaram P (1993) *FEBS* **321(1)** :79-83.

[27] Bang D, Gribenko AV, Tereshko V, Kossiakoff AA, Kent SB, Makhatadze GI (2006) *Nature Chemical Biology* **2** 139-143.

[28] Stillinger F, Weber TA, *Phys. Rev. B* **31**, 5262 (1985).

[29] Tersoff J, *Phys. Rev. B* **37**, 6991 (1988).

[30] Brenner DW *Phys. Rev. B* **42**, 9458 (1990).

[31] Friesner RA (2005) *Proc Natl. Acad. Sci. USA* **102(19)** :6648-6653.

[32] Gerald Edward Brown and A. D. Jackson, (1976) *The Nucleon-Nucleon Interaction.* North-Holland Publishing, Amsterdam,

[33] N. Barkai, M. Rose and N. Wingreen (1998) *Nature* **396** 422-423

[34] Kortemme T, Morozov AV, and Baker D (2003) *J Mol Biol* **326** 1239-59.

[35] Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: a program for macromolecular energy, minimization and dynamic calcualtion. 1983; **4**: 187-217.

[36] Feig M. & Brooks C.L. (2004) Recent advances in the development and application of implicit solvent models in biomolecule simulations. Curr. Opin. Struct. Biol. **14:2**, 217-224 (2004).

Acknowledgement

We thank Merav Parter and Nadav Kashtan for the development and implementation of the random model, as well as for helpful discussions and fruitful ideas; Dr. Avi Mayo for suggesting the Yukawa potential; Dr. Maria Krisch for critical reading of the manuscript; Prof. Eytan Domany and Prof. Avigdor Scherz for helpful discussions. This research was funded by MINERVA grant 8525 and the Ministry of Science and Technology (MOST) grant 0263, and the Kimmelman Center for Complexity Science. Y. L. is the incumbent of the Lilian and George Lyttle Career Development Chair.

Figures Legends

Fig. 1. Describing proteins as mathematical graphs. **(***A)* Examples for two of the most probable motifs. The α-helix motif M9 (top row) and the anti-parallel β-sheet motif M3 (bottom row), in various presentations. A Covalent interaction in black, a single H-bond in normal red, and a double H-bond in thick red. The occurrence (per residue) are <M3, M9 > = <6.95%,9.7%>. (*B*) Examples for a 4-helix bundle (pdb 1tqg, top) and a β-barrel (GFP, pdb 1oxe, bottom) proteins. The left column depicts the contact map using a distance threshold of R_C=8Å between the C_α, with contacts in blue and H-bonds in red. The middle column shows the planar drawing of the contact map, with vertices positions based on the observed motifs: helices are represented by the box-shaped motif M9, while a β–sheet resembles the beehive shape of M3 hexagons.

Fig. 2. Motifs probabilities using a logarithmic scale (see methods), compared to DSSP annotations [14]. See SI Fig. 6 for a visualization of all the motifs. The first ten motifs occur with probability $<10^{-315}$. M14 is the Schellman Motif. Note the Glycine preference at position 5. M18, the 3_{10} helix, is explained by an H-bond of residues n and residue 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}. Motif 13 is more prevalent than the 3_{10} helix (M18) (occurrence of 0.23% vs. 0.2%). Yet, 3_{10} helix is widely represented in the literature as an alternative helix, due to its 'nice' shape. (*Inset*), The probability of the next 14 motifs, using a normal scale. Only M36 to M43 seems to be in the 'twilight zone' of significance, for which statistical fix for multiple comparisons may be applied. Subgraphs #44 and on cannot be considered motifs at all.

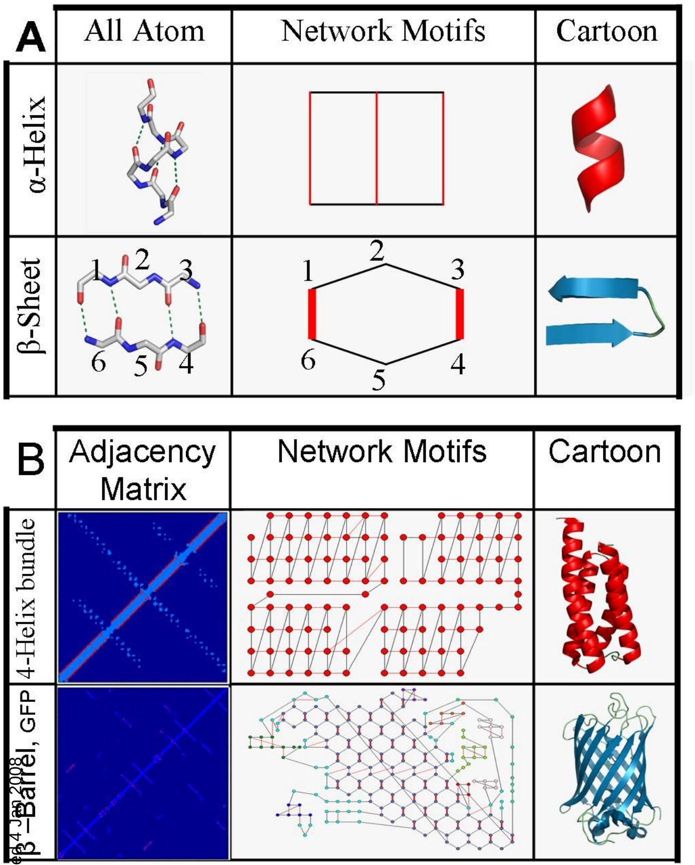
Fig. 3. Motifs are visualized by color-coding on the protein structures analyzed in details in this study. The figure was drawn using PyMol.

Fig. 4. Significance Profiles of the proteins drawn in Fig. 3. Frequencies of motifs in experimental structures (red bars) compared to MD simulation trajectories (blue lines with error bars) of 4 ns. High similarity is observed in M1-M10, but a bias can be seen

in the region M11-M30. (*A*) The 434 repressor. (*B*) Lysozyme. X-ray (n=7) compared to both NMR (n=50) and MD snapshots structures (n=100). (*C*) The SH3 domain NMR Motifs 3, 10, and 27 are hexagons which form a β-sheet (see Fig. 1). These motifs have a rather similar SP in the MD simulation vs. the NMR (although somewhat overrepresented in the MD). However, the less frequent motifs (M16, M29, and M33) do not exist at all in the MD. The *inset* shows the time behavior of M10 along MD simulation. The curve depicts the number of residues in which M10 occurs. After about 45% of the simulation time (equivalent to 1.6 ns), the number of M10 occurrences increases from about 18 to about 24. This could be explained as a compensation for the loss of M16.

Fig. 5. Contact Map of Lysozyme. (*A*) X-ray crystal structure, compared to (*B*), a snapshot from the MD at 1.6 ns. Although the major secondary structure elements are conserved, some H-Bonds break (arrows), caused by a backbone perturbation of 0.63Å (**C**).

Figure 1


A
All Atom
Network Motifs
Cartoon
α-Helix
β-Sheet
1
2
3
6
5
4
1
2
3
6
5
4
B
Adjacency Matrix
Network Motifs
Cartoon
4-Helix bundle
β -Barrel, GFP

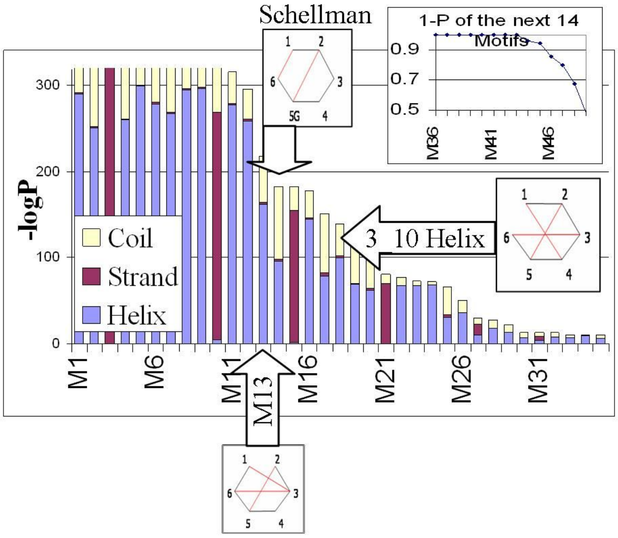

Schellman
1 2 6 3 5G 4
1-P of the next 14 Motifs
0.9
0.7
0.5
M36
M41
M46
300
200
100
0
-logP
Coil
Strand
Helix
3_10 Helix
1 2 6 3 5 4
M1
M6
M11
M16
M21
M26
M31
M13
1 2 6 3 5 4

Figure 2

Figure 3. Motifs Visualization.

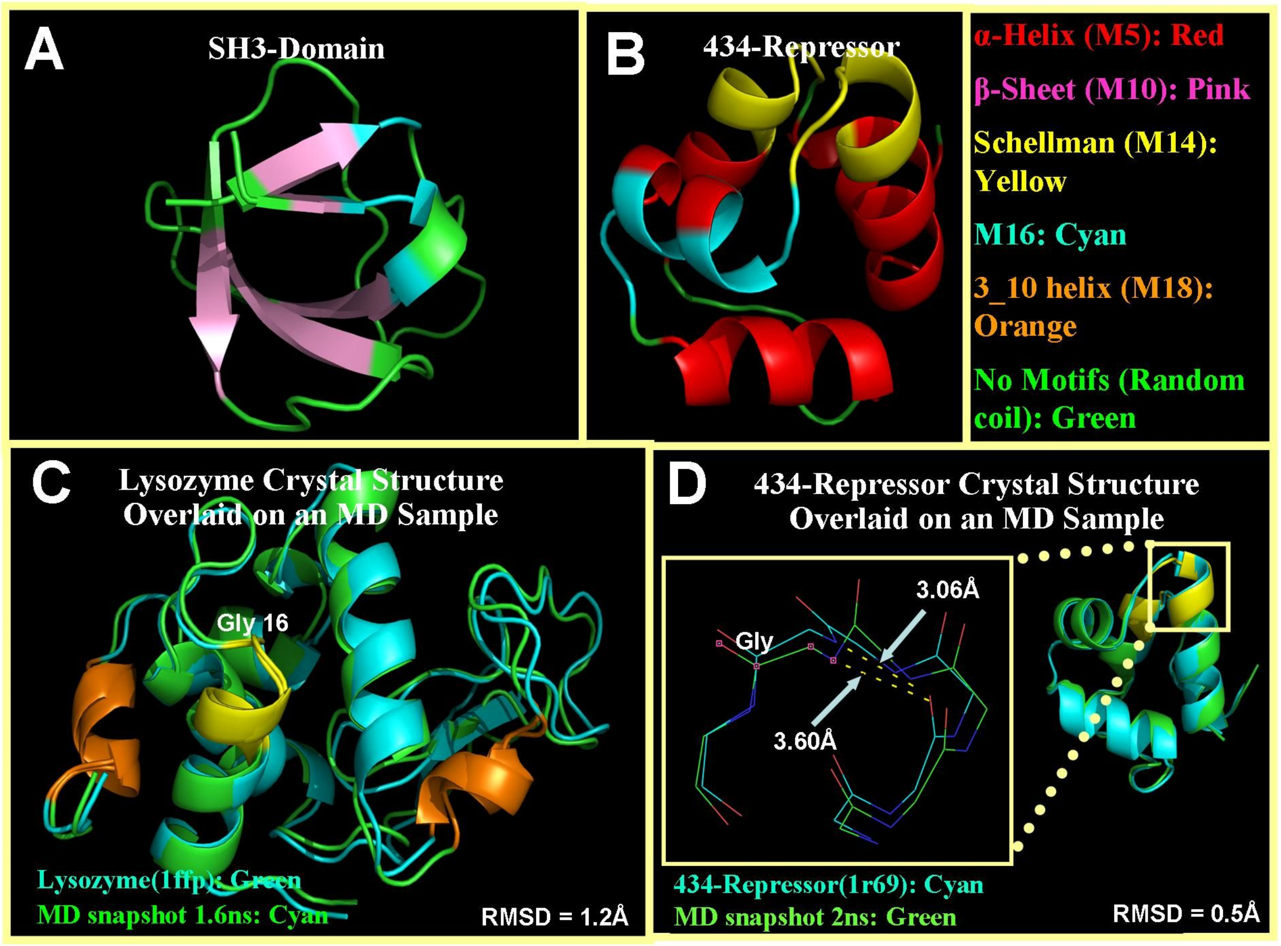
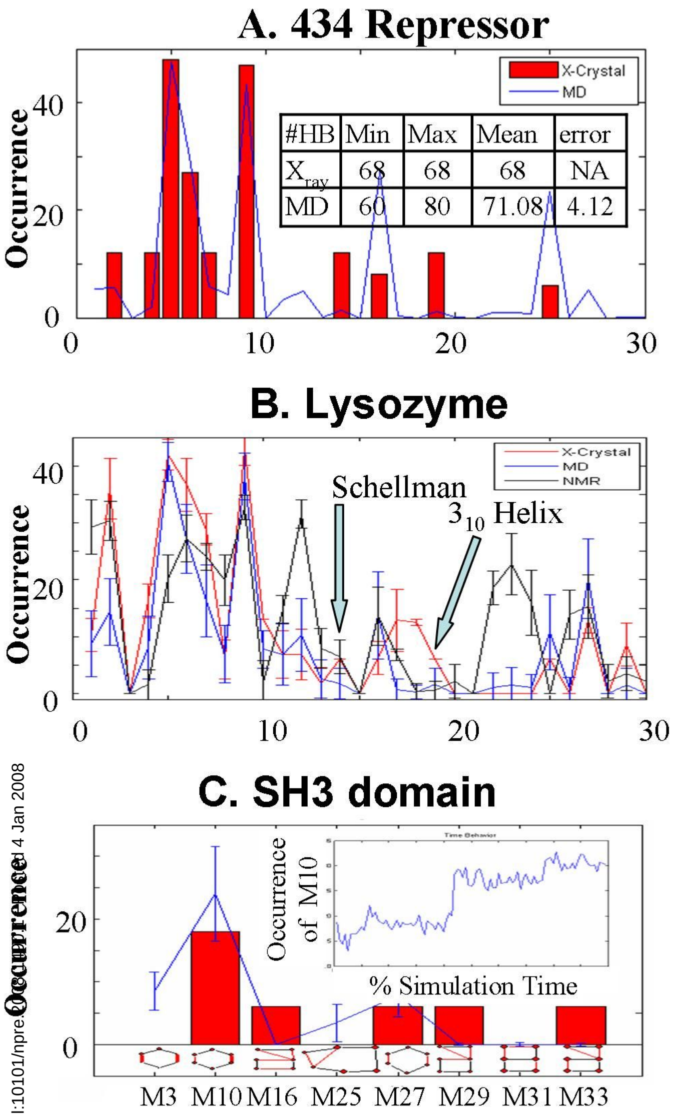
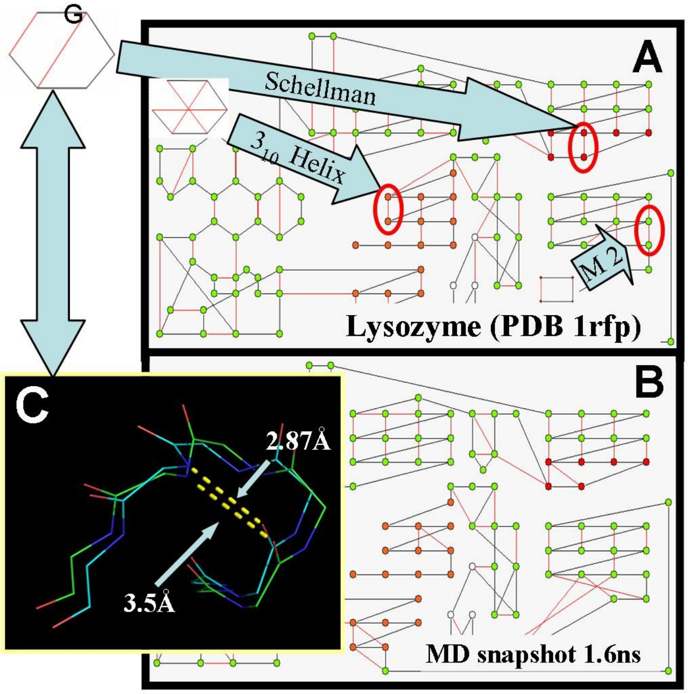


Figure 4



#HB	Min	Max	Mean	error
X$_{ray}$	68	68	68	NA
MD	60	80	71.08	4.12

Figure 5


G
A
Schellman
3₁₀ Helix
M 2
Lysozyme (PDB 1rfp)
B
MD snapshot 1.6ns
C
2.87Å
3.5Å