Motif Abundances in Protein Structure Networks

Blake Stacey*
MIT Department of Physics
(Dated: May 12, 2005)

I test a claim of R. Milo *et al.* [2] concerning the preponderance of particular 4-node subgraphs within networks derived from protein structures. Applying the same statistical tests to a sample 170 times larger, it is seen that the same subgraph "motifs" are most strongly represented. However, the effect of duplicated protein domains (and of exceptionally large single peptides in general) tends to favor the complete 4-graph, regardless of biological considerations.

1. INTRODUCTION

In recent years, several authors have proposed methods for analyzing complex systems as networks, sets of nodes connected by edges. These networks may be derived from biological systems, such as the web of protein-protein interactions within a cell; from social systems, such as a network of human friendships; or from computational sources, such as the URL strucutre of the World Wide Web. In all these cases, the first step to any analysis is to decide what constitutes a node and what requirement must be met for two nodes to be linked. Analyzing human society through network theory, for example, presupposes that friendship is well-defined. Here, I shall treat a model of protein geometry in which nodes are secondary-structure elements (α-helices and β-sheets), and nodes are linked if they are closer than some threshhold distance to each other.

One method to study a network is to hunt through it for repeated *motifs,* subgraphs containing a few nodes apiece, in the hope that a subgraph which occurs many times is physically or biologically significant [1]. (This need not be true *a priori:* the heme unit within a hemoglobin molecule occurs only four times, so one could plausibly invent a network representation of hemoglobin which downplays the heme component.) Milo *et al.* advance the claim that a particular set of four-node motifs are "overrepresented" in protein structure; that is, these particular motifs occur in protein geometry more often than expected by chance. However, this claim was only supported by three proteins [2]. In this paper, I shall examine how well it applies to 512 more.

2. SEQUENCE PROFILES

Given a network (produced by whatever means), we can count the number of times a particular subgraph occurs. Take a set of motifs labeled by i, and let the number of occurences for each motif be n_i. (In this paper, $i = 1, \ldots, 6$.) The quantities n_i are not the most convenient way to measure the abundances of network motifs, since they scale with the overall network size. Besides, the quantity of real interest is not how many times a particular grouping of nodes occurs, but how many occurences we find *compared to what we would observe due to random chance.* Imagine a randomized ensemble of networks, each with the same degree profile as the original, but with the edges rearranged. Presumably, if a particular motif is prevalent (or lacking) for biological reasons, it will occur more (or less) frequently than in this randomized ensemble. We define a measure of the statistical significance of a particular n_i, relative to chance expectation:

$$Z_i \equiv \frac{n_i - \langle r_i \rangle}{\sqrt{\langle r_i^2 \rangle_c}}. \tag{1}$$

Here, $\langle r_i \rangle$ denotes the mean number of times motif i occurs per network, averaged over all networks in the ensemble. The cumulant in the denominator, $\langle r_i^2 \rangle_c$, is the variance of r_i, likewise computed over all the randomized networks.

Proteins are turned into networks in the following manner: represent each α-helix or β-sheet by a node, and connect nodes if their corresponding structure elements come within 10 Å of each other. α-helices and β-sheets are three-dimensional objects, of course, so the distance between them is defined (with appropriate arbitrariness) to be the minimum separation between C_α atoms of their respective amino acids. This information can be computed from the protein's PDB file with relative ease.

The networks in the randomized ensemble were generated by swapping edges in the network generated from the protein geometry. This method has the advantage that it preserves connectivity properties of the network: each node will have the same number of ingoing and outgoing connections as it did in the original.

To compare Z-scores calculated for different proteins, we introduce a new variable, S_i, which is normalized so that the sum over all i is unity:

$$S_i \equiv \frac{Z_i}{\sum_i Z_i^2}. \tag{2}$$

The set $\{S_i\}$ is termed the *sequence profile.*

This method suffers from several drawbacks. Consider, for example, an arrangement such as the network

*Electronic address: bstacey@mit.edu; URL: http://web.mit.edu/bstacey/www

of neurons in an animal brain, where it is difficult for nodes separated by a large physical distance to connect. In this case, we might find groups of neighboring nodes connected more strongly than chance expectations only because they are situated near each other, and not for any reason having to do with biological selection. A "toy model" has been constructed, in which nodes are stochastically connected to proximate neighbors (connections being formed with a probability that falls off as a Gaussian). Analyzing the resulting networks shows the same prevalent motifs as Milo *et al.* find in the *C. elegans* neural connection structure[3]. Such toy models do, however, show motifs which do *not* appear in "real" networks, indicating that the sequence-profile method is not entirely fragile [4].

A subtlety arises when we consider that many proteins consist of multiple *domains,* each one of which looks much like the others. How do Z and S behave when we "double" a protein, taking one fundamental domain and reproducing it two, three or more times? Neglecting the "edge effects" caused by one domain's amino acids being physically close to another domain, we can approximate this situation by having d disjoint copies of the same network. (For the moment, I presuppose that all domains are identical in nature.) As will be verified empirically below, for d sufficiently large, each Z_i scales as

$$Z_i(d) \approx dZ_i f_i, \tag{3}$$

where f_i is a constant of order 1.

This has an interesting effect upon the normalized variable S. Substituting Eq. (3) into Eq. (2) shows that

$$S_i = \frac{dZ_i f_i}{d^2 \sum_i Z_i^2 f_i^2} = \frac{1}{d} \frac{Z_i f_i}{\sum_i Z_i^2 f_i^2}. \tag{4}$$

If some particular Z_j is larger than the others, it will come to dominate the sum. S_j will grow to asymptotically approach 1, while all other S_i for $i \neq j$ will shrink correspondingly to 0.

3. DATA ANALYSIS

Before examining the motif compositions of proteins *en masse,* it is instructive to make a few observations on the three proteins studied in [2]. These three proteins are an oxioreductase (PDB code 1AOR), a serine protease inhibitor (1EAW) and an immunoglobin (1A4J). Applying the algorithm discussed in Section 2, the PDB files for these three proteins were processed into networks, along with 1EQR, an aspartyl-tRNA synthetase from *E coli.* Figure 1 shows the degree profiles for these networks. The popularity of the scale-free network model makes it inevitable to attempt fitting these histograms with a power-law decay; they do not fit terribly well.

Each of these proteins consists of multiple, dissimilar domains. While it may be reasonable to consider the entire protein as a unit, as done in [2] (since natural selection presumably acts upon the entire protein), it is also possible to regard each subunit separately. As shown in Figure 2, these networks also display broad peaks with fairly sharp rises and slower decays.

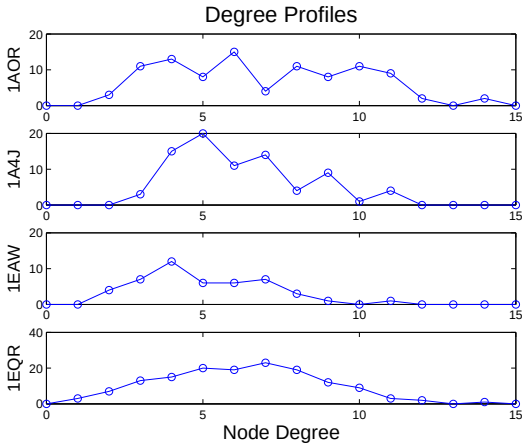

FIG. 1: Degree profiles for four entire proteins.

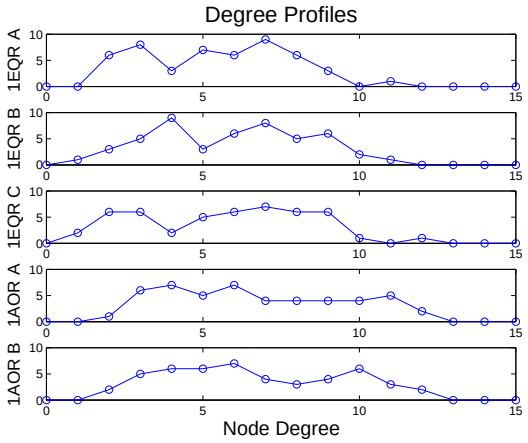

FIG. 2: Sample degree profiles for protein domains.

The next step is to compute the *sequence* profile for these proteins. Computations were performed using the MFinder 1.1 software [5]. To compare results with Milo *et al.,* I focused upon the six motifs listed in Figure 3. For each of the 512 proteins in my dataset, I used a Python program to download the PDB file from Brookhaven, extract the geometrical structure information, and convert that geometry into a network. This code and my dataset are available for closer examination [6].

The results are shown in Figures 4-6. A normalized Z-score of zero indicates, as per Eq. (2), that motif i is as common in the real protein as it is in the randomized network ensemble. The figures indicate that motifs 3, 5 and 6 are comparatively more frequent, in all three proteins, than the other motifs in question. (This is the result quoted in [2].) The same qualitative behavior is

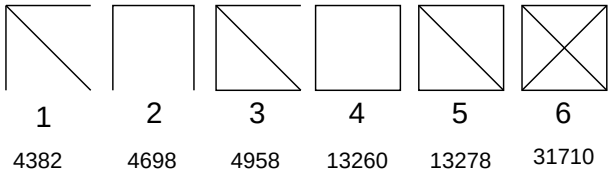

FIG. 3: Six four-node motifs with their associated MFinder identification codes (derived from the connection matrix).

seen in the sequence profiles of the protein subunits.

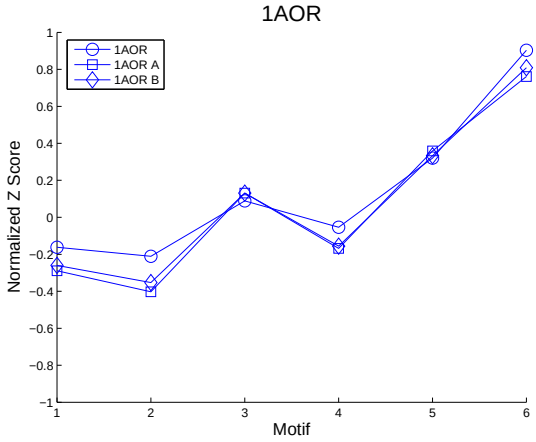

FIG. 4: Sequence profile of the oxioreductase 1AOR, from the hyperthermophilic archaea *Pyrococcus furiosis*.

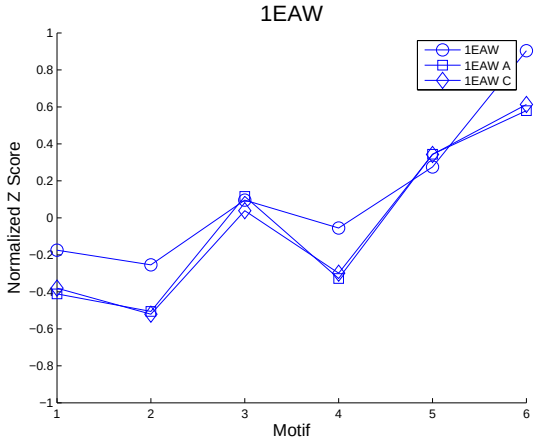

FIG. 5: Sequence profile of the *H. sapiens* serine protease inhibitor 1EAW.

Another topic of interest, as suggested above, is the influence of duplicating protein domains upon the Z- and S-scores. Figure 7 shows these scores computed for networks derived from 1EQR and duplicated up to eight times. (In all cases, the domains were treated as disjoint units, with no interactions.)

Finally, now that we have developed some understanding of these operations, we can apply them to a large dataset. Figure 8 illustrates the algorithm of Section 2 applied to 512 proteins from Sander and Hobohm's November 2004 list of non-homologous sequences [7]. Sequences were chosen to have a minimum of 150 amino-acid residues, and so that no two randomly chosen proteins have more than 25% sequence identity. All sequences less than 100 residues long have only one domain.

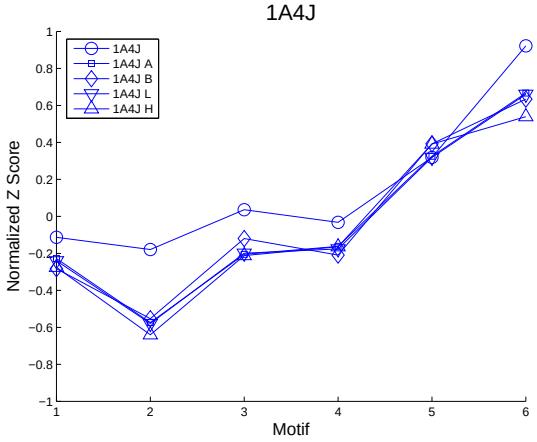

FIG. 6: Sequence profile of the *Mus musculus* Diels-Alder catalytic antibody germline precursor, 1A4J.

Again, motifs 3, 5 and 6 have $S_i > 0$, bearing out the pattern observed with the three initial proteins. These histograms can be fit decently well to single or double Gaussians, with the parameters listed in Table I. Even including the statistical error, the centers of distributions 5 and 6 are well over one standard deviation above 0, indicating that the abundances of these motifs are, overall, robust.

4. CONCLUSIONS

Overall, the behavior of the larger dataset upholds the claim advanced based upon the first three proteins: motifs 3, 5 and 6 of Figure 3 occur more often than chance expectation, while motifs 1, 2 and 4 appear underrepresented. Examining the overall distribution of S_i, I find that motifs 5 and 6 (the complete 4-graph and the complete 4-graph less one edge) are the most surely overrepresented, although the tendency for S_6 to approach 1 suggests that the duplication of domains is somewhat influencing the results. Since the same tendencies are apparent in the three proteins studied in [2], it is likely the same factors (favoring motif 6, at least) influence the results of that paper as well.

[1] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, Science **298**, 824 (2002).
[2] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon, Science **303**, 1538 (2004).
[3] Y. Artzy-Randrup et al., Science **305**, 1107c (2004).
[4] R. Milo et al., Science **305**, 1107d (2004).
[5] U. Alon et al., *MFinder program* (2002), `http://www.weizmann.ac.il/mcb/UriAlon/groupNetworkMotifSW.html`.
[6] B. Stacey (2005), `\http://web.mit.edu/~bstacey/www/8.592/final`.
[7] U. Hobohm, M. Scharf, R. Schneider, and C. Sander, Protein Science **1**, 409 (1992), November 2004 update from `ftp.heidelberg.de`.

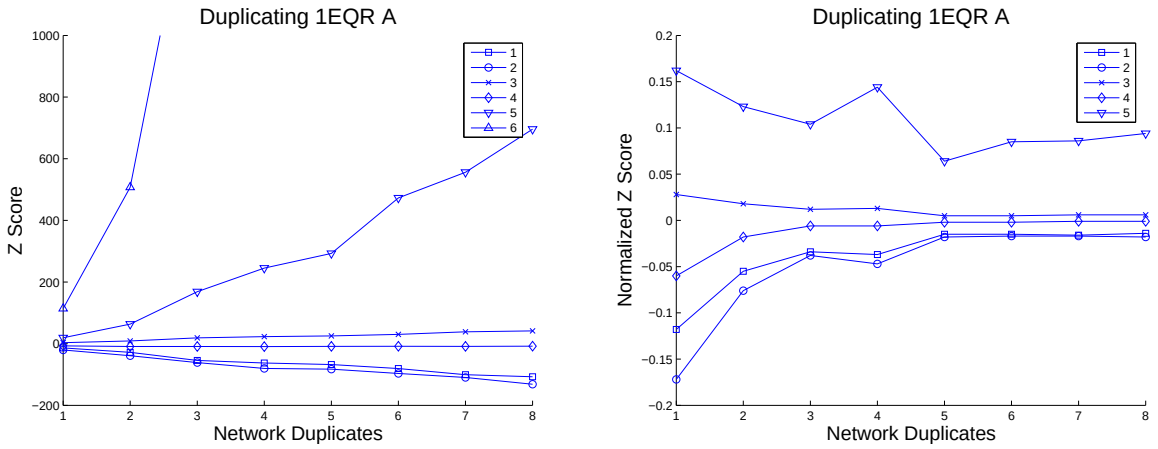

FIG. 7: Raw Z_i and normalized S_i scores for a fundamental domain of 1EQR.

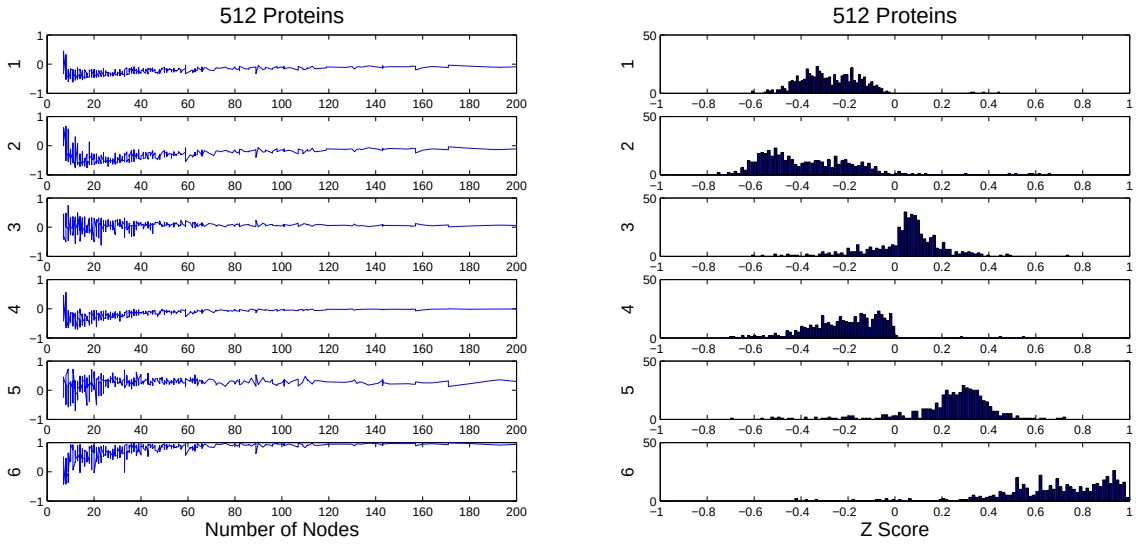

FIG. 8: Overview of normalized scores for a set of 512 sequences. The plots for motif 6 clearly show the effect of one S-score becoming dominant, as discussed after Eq. (4).

Motif	Height 1	Center 1	Width 1	Height 2	Center 2	Width 2	χ^2
1	17.5 ± 1.0	-0.35 ± 0.01	0.09 ± 0.01	13.7 ± 1.0	-0.17 ± 0.01	$0.08x \pm 0.01$	0.7
2	20 ± 1.3	-0.53 ± 0.01	0.095 ± 0.005	11.7 ± 0.8	-0.26 ± 0.01	0.15 ± 0.01	0.6
3	22 ± 2	0.072 ± 0.006	0.127 ± 0.007	—	—	—	1.4
4	15.4 ± 1.0	-0.25 ± 0.01	0.145 ± 0.01	21 ± 2	-0.17 ± 0.01	0.09 ± 0.01	0.9
5	25 ± 2	0.29 ± 0.05	0.13 ± 0.05	—	—	—	1.0
6	17 ± 2	0.92 ± 0.01	0.058 ± 0.008	11.2 ± 0.7	0.68 ± 0.02	0.23 ± 0.02	0.85

TABLE I: Double-Gaussian fits with Poisson noise of the histograms in Figure 8. The residuals from these fits do not show any obvious patterns; as to why the histograms are Gaussian at all, I can only wave my hands and invoke the Fuzzy Central Limit Theorem.