Network motifs come in sets: Correlations in the randomization process

Reid Ginoza*

Division of Natural Sciences and Mathematics, Bennington College, Bennington, Vermont 05201, USA

Andrew Mugler

Department of Physics, Columbia University, New York, New York 10027, USA

Received 25 January 2010; revised manuscript received 5 May 2010; published 22 July 2010

The identification of motifs—subgraphs that appear significantly more often in a particular network than in an ensemble of randomized networks—has become a ubiquitous method for uncovering potentially important subunits within networks drawn from a wide variety of fields. We find that the most common algorithms used to generate the ensemble from the real network change subgraph counts in a highly correlated manner, such that one subgraph’s status as a motif may not be independent from the statuses of the other subgraphs. We demonstrate this effect for the problem of three- and four-node motif identification in the transcriptional regulatory networks of E. coli and S. cerevisiae in which randomized networks are generated via an edge-swapping algorithm. We find strong correlations among subgraph counts; for three-node subgraphs these correlations are easily interpreted, and we present an information-theoretic tool that may be used to identify correlations among subgraphs of any size. Our results suggest that single-feature statistics such as $Z$ scores that implicitly assume independence among subgraph counts constitute an insufficient summary of the network.

I. INTRODUCTION

Identifying motifs has become a standard way to probe the functional significance of biological, technological, and sociological networks [1–7]. A motif is commonly defined as a subgraph whose number of appearances in a particular network is significantly greater than its average number of appearances in an ensemble of networks generated under some null model [7]. The typical null model prescribes an algorithm by which many randomized networks can be produced from the original network (see, e.g., Milo et al. [8] for a review and comparison of several such algorithms). Using an ensemble generated from the actual network often preserves features of the network that are desired for fair comparison, such as the degree distribution.

Many researchers have demonstrated that care must be taken when summarizing a network based on a motif analysis. For example, some have argued that motifs need not be of specific local importance but can instead arise as by-products of more global topological constraints [9–11]. In biological networks, it has been observed that motifs do not correspond to subgraphs deemed important under different metrics, such as functional enrichment or evolutionary conservation [9,10,12]. Other critiques have centered on the implementation of the null model for generating an ensemble of random networks; Arzty-Randrup and Stone, for example, have shown that without correction simple edge-swapping algorithms do not generate ensembles in which each random network appears with equal probability [13].

In this paper we focus on another concern in motif identification arising from the generation of the ensemble of random networks: the randomization process can induce correlations in subgraph counts, such that one subgraph’s status as a motif is not independent from the statuses of the other subgraphs. We demonstrate and interpret such correlations in a simple case and describe how a well-known measure from signal processing, mutual information [14,15], may be used to identify such correlations in general.

II. METHODS

Following Milo et al. [7], we perform three- and four-node motif detections on the transcriptional regulatory networks of E. coli (version 1.1) and S. cerevisiae, using their freely available network data and software (Mfinder version 1.2) [16]. Generation of randomized networks from the actual network is performed according to one of the three null models: an edge-swapping algorithm, an edge-matching algorithm, and a Monte Carlo algorithm, all described in detail in [8]. Because significance results are similar among models (cf. [8]), emphasis here is placed on the edge-swapping algorithm, a Markovian chain procedure that repeatedly swaps the target nodes between pairs of edges. A $Z$ score is computed for each subgraph, $Z=(n_0-\mu)/\sigma$, where $n_0$ is the subgraph count in the actual network, and $\mu$ and $\sigma$ are the mean and standard deviation of the subgraph count within an ensemble of at least 1000 randomized networks [7].

We quantify correlation between the counts of any two subgraphs over the course of a randomization process using mutual information [14,15]. Mutual information has been used successfully in the study of biological networks both as a statistical measure [17,18] and as a measure of functionality in the presence of intrinsic noise [19,20]. It captures correlation between two random variables even when a relationship exists that is nonlinear (unlike, e.g., the correlation coefficient) or nonmonotonic (unlike, e.g., Spearman’s rho).

In this study, the counts $n_i$ and $n_j$ of the $i$th and $j$th subgraphs at each iteration of the edge-swapping process are used to increment a count matrix from which the joint probability

*reid.ginoza@gmail.com
distribution \( p(n_i, n_j) \) is obtained by normalization, and mutual information \( I_{ij} \) is computed as

\[
I_{ij} = \sum_{n_i, n_j} p(n_i, n_j) \log_2 \frac{p(n_i, n_j)}{p(n_i)p(n_j)},
\]

(1)

where the logarithm has base 2 to give \( I_{ij} \) in bits, and \( p(n_i) = \sum_n p(n_i, n_j) \) and \( p(n_j) = \sum_n p(n_i, n_j) \).

Mutual information \( I_{ij} \) is bounded from below by zero [as seen in Eq. (1) when there is no correlation and the subgraph counts are independent of each other, i.e., \( p(n_i, n_j) = p(n_i)p(n_j) \)] and bounded from above by the smaller of the two variables’ entropies \( H_i \) and \( H_j \), where

\[
H_i = -\sum_{n_i} p(n_i) \log_2 p(n_i).
\]

In order to obtain a statistic that can be compared across all subgraph pairs, we normalize by the average entropy, defining

\[
a_{ij} = \frac{I_{ij}}{(H_i + H_j)/2}
\]

(3)

as our measure of correlation. Note that \( 0 \leq a_{ij} \leq 1 \), with \( a_{ij} = 0 \) when \( I_{ij} = 0 \) and \( a_{ij} = 1 \) when \( i = j \). We find qualitatively similar results (cf. Sec. III) when normalizing by the minimum, instead of the average, entropy.

### III. RESULTS

#### A. Interpretable correlation

Only four three-node subgraphs are present in the transcriptional network of *E. coli*, and a Z score analysis of the type performed in [7] reveals a curious effect. Specifically, with respect to ensembles generated via any of the edge-swapping, edge-matching, and Monte Carlo algorithms [8], the Z scores of three of the subgraphs (IDs 6, 12, and 36; where IDs are defined in Fig. 1) are either very close or equal to the negative of the Z score of the fourth subgraph (the feed-forward loop, ID 38), as shown in Fig. 2. In fact, as shown in Fig. 3, the absolute value of the difference in counts within the actual network and counts within a sample randomized network at each iteration of the edge-swapping algorithm is the same among all four subgraphs for the first 1000 iterations. The interpretation is simple: as detailed in Fig. 4, each time an edge of a feed-forward loop is swapped with an external edge, the feed-forward loop is destroyed and one of each of the other three subgraphs is created; using subgraph IDs we may denote this process as

\[
38 \rightarrow 6, 12, 36.
\]

Since this process accounts for the overwhelming majority of the changes in count of the latter three subgraphs, there is an extremely high correlation among the counts of all four sub-

---

FIG. 1. All possible three-node subgraphs, labeled as in Alon’s “motif dictionary” [16].

FIG. 2. The four three-node subgraphs that appear in the *E. coli* network. Z scores are calculated with respect to an ensemble of 1000 randomized networks, generated via the edge-swapping algorithm [8]. Plots show the count of each subgraph during the generation of one randomized network. Each iteration corresponds to one edge swap.
graphs in each randomized network, and the magnitudes of their Z scores are very close.

The plots in Fig. 2 illustrate two distinct regimes: (i) a regime at low iteration number, during which the subgraph count is changing roughly monotonically, and (ii) a regime at high iteration number, during which the subgraph count is fluctuating about a steady-state value. The first regime corresponds to the randomization of the real network, while the second regime corresponds to further exploration in the space of randomized networks. We emphasize that the correlations described above occur in both regimes (we have performed runs of up to ~100 000 iterations). That is, further randomization of an already random network, while (by definition) not revealing motifs, produces the same correlations that occur during randomization of the real network. The correlations are a product of the randomization scheme and are present even when motifs are not.

**B. Information-theoretic tool**

To quantify and extend the detection of correlations such as that just described, we use a normalized mutual information measure, as detailed in Sec. II. For the cases of three- and four-node subgraphs in both the *E. coli* and *S. cerevisiae* transcriptional networks, the measure \( a_{ij} [\text{cf. Eq. (3)}] \) is computed between all pairs of subgraphs \( i \) and \( j \) that appear during the randomization of a network via the edge-swapping algorithm. Figure 5 shows the matrices \( a_{ij} \); the row and column order is determined by summing along either direction and sorting, which tends to group together sets of subgraphs with high pairwise correlations.

During randomization of the *E. coli* network, a set of four three-node subgraphs (IDs 6, 12, 36, and 38; cf. Fig. 1) are highly correlated, as shown by the bright 4 × 4 square in Fig. 5(A). The high correlation is simply the result of the effect described in the previous section, in which any of the three swaps overwhelmingly converts a feed-forward loop (ID 38) into three other subgraphs (IDs 6, 12, and 36). In fact the same set of high correlations is seen during the randomization of the *S. cerevisiae* network, as shown by the upper left 4 × 4 square in Fig. 5(B). There are additional correlated sets in *S. cerevisiae*: subgraphs 14, 74, and 102 are highly correlated as indicated by the bright 3 × 3 square involving these IDs in Fig. 5(B), and subgraphs 74 and 108, as well as subgraphs 14 and 46, are correlated as indicated by the relatively bright entries at these coordinate pairs in Fig. 5(B). Respectively, these correlations are due to the effects [in the notation of Eq. (4)]

\[
102 \rightarrow 12, 14, 74, \quad (5)
\]

\[
108 \rightarrow 6, 74, 74, \quad (6)
\]

\[
46 \rightarrow 14, 14, 36, \quad (7)
\]

of which one may convince oneself with the aid of Fig. 1. Note that although subgraphs 14, 102, and 108 participate in the highly correlated effects described here, none changes in number significantly enough upon randomization to be labeled a motif in the *S. cerevisiae* network [7] (subgraphs 46 and 74 do not appear in the actual network, only during the course of the randomization).

Our analysis reveals correlations between counts of four-node subgraphs as well. As indicated by the bright blocks
and off-diagonal elements in Figs. 5(C) and 5(D), several sets of subgraphs are highly correlated during the randomization of both the \( E. coli \) and \( S. cerevisiae \) networks. Correlations are less easily interpreted in the four-node case than in the three-node case, but one must nonetheless remain aware of such artifacts of the randomization process when identifying subgraphs as motifs. We note that the bifan (ID 204), the four-node subgraph commonly identified as a motif in a variety of networks including both transcriptional networks studied here [7], does not exhibit particularly high correlation with any other subgraph under our measure in either the \( E. coli \) or \( S. cerevisiae \) network.

There are more principled ways to cluster the correlations in Fig. 5; when combined with a single-feature statistic such as the \( Z \) score, clustering can quantitatively reveal sets of subgraphs which are both (i) correlated during the randomization process and (ii) significantly under- or over-represented in the actual network, which we call “motif sets.” For example, Fig. 6 shows the results of a principal component analysis on the correlation matrices in Fig. 5, in which subgraphs are plotted according to their components in the eigenvectors of the correlation matrix corresponding to the eigenvalues with the first- and second-largest magnitudes, which tends to cluster subgraphs with high mutual correlations. Figures 6(A) and 6(B) both highlight the motif set that includes three-node subgraph IDs 6, 12, 36, and 38, in which the feed-forward loop is over-represented, and the three subgraphs with which it is correlated are under-represented (cf. Fig. 2). Figure 6(B) also highlights a case in which a set of subgraphs is correlated (and thus clusters in eigenvector space) but for which none of the subgraphs have large-magnitude \( Z \) scores, and so is not a motif set. Similarly, among four-node subgraphs in both the \( E. coli \) and \( S. cerevisiae \) networks, although the principal component analysis reveals several clusters [Figs. 6(C) and 6(D)], none are comprised of subgraphs which have large-magnitude \( Z \) scores,
i.e., there is correlation but no significant overabundance or underabundance.

We find results qualitatively similar to Figs. 5 and 6 when normalizing by the minimum, instead of the average, entropy in Eq. (3). The techniques we describe here can be extended to the detection of subgraphs of any size.

IV. DISCUSSION

By quantifying correlations among subgraph counts during three- and four-node motif detections in the transcriptional networks of *E. coli* and *S. cerevisiae*, we reveal that motifs can come in sets: the destruction of a subgraph during the randomization process can be highly correlated with the creation of one or more other subgraphs. The correlations are easily understood in the three-node case, and we present an information-theoretic tool to extract such correlations in general. The correlations are artifacts of the algorithm used to generate the ensemble of randomized networks; although we demonstrate their existence here in the context of only one randomization algorithm, the edge-swapping algorithm, they occur in other commonly used algorithms, as evidenced by mutually consistent effects on the Z scores.

Many correlations are near maximal (i.e., $a_{ij} = 1$; cf. Fig. 5), meaning that if the count of one subgraph changes, the count of the other nearly always changes as well. For these subgraph pairs, using a single-feature statistic such as the Z score, which implicitly assumes independence between two subgraphs’ statuses as motifs, is simply an insufficient description. Rather, we use here the correlations to supplant an identification of independent motifs with that of codependent motif sets. The abundance of the feed-forward loop, for example, in these networks and under these null models, is contextual: it is significantly over-represented if and only if

FIG. 6. Principal component analysis on the correlation matrices in Fig. 5. Subgraphs are plotted by their components in the eigenvectors of the correlation matrix corresponding to the eigenvalues with the first- and second-largest magnitudes (arbitrary units). Subgraphs are labeled as in Alon’s motif dictionary (Fig. 1 and [16]). Examples are shown in (A) and (B) in which there are high correlation (i.e., clustering in eigenvector space) and high significance (i.e., Z scores with large magnitude), which we call motif sets. An example is shown in (B) in which there is high correlation but low significance, corresponding to subgraphs that are correlated during the randomization process but are neither under- nor over-represented in the actual network.
three other subgraphs are significantly under-represented. We advocate that in future work our observations serve as the basis for more principled clustering of subgraphs based on correlations (e.g., by mixture modeling in which the state of the subgraph count is a mixture of several states, with counts conditionally independent given the state).

ACKNOWLEDGMENTS

We thank Chris H. Wiggins for his helpful suggestions and guidance with this work. AM was supported by NSF Grant No. DGE-0742450.

[1] S. Wuchty, Z. N. Oltvai, and A.-L. Barabási, Nat. Genet. 35, 176 (2003).
[2] H.-W. Ma, B. Kumar, U. Ditges, F. Gunzer, J. Buer, and A.-P. Zeng, Nucleic Acids Res. 32, 6643 (2004).
[3] O. Sporns and R. Kötter, PLoS Biol. 2, e369 (2004).
[4] J. Berg and M. Lässig, Proc. Natl. Acad. Sci. U.S.A. 101, 41689 (2004).
[5] E. Yeger-Lotem, S. Sattath, N. Kashtan, S. Itzkovitz, R. Milo, R. Y. Pinter, U. Alon, and H. Margalit, Proc. Natl. Acad. Sci. U.S.A. 101, 5934 (2004).
[6] J. Tsang, J. Zhu, and A. van Oudenaarden, Mol. Cell 26, 753 (2007).
[7] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, Science 298, 824 (2002).
[8] R. Milo, N. Kashtan, S. Itzkovitz, M. E. J. Newman, and U. Alon, e-print arXiv:cond-mat/0312028.
[9] Y. Artzy-Randrup, S. J. Fleishman, N. Ben-Tal, and L. Stone, Science 305, 1107c (2004).
[10] A. Mazurie, S. Bottani, and M. Vergassola, Genome Biol. 6, R35 (2005).
[11] A. S. Konagurthu and A. M. Lesk, BMC Systems Biology 2, 73 (2008).
[12] O. Meshi, T. Shlomi, and E. Ruppin, BMC Systems Biology 1, 1 (2007).
[13] Y. Artzy-Randrup and L. Stone, Phys. Rev. E 72, 056708 (2005).
[14] C. E. Shannon, Bell Syst. Tech. J. 27, 379 (1948).
[15] T. M. Cover and J. A. Thomas, Elements of Information Theory (Wiley, New York, 1991).
[16] Uri Alon, Molecular Cell Biology Lab, http://www.weizmann.ac.il/mcb/UriAlon/
[17] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. D. Favera, and A. Califano, BMC Bioinf. 7, S7 (2006).
[18] E. Ziv, M. Middendorf, and C. H. Wiggins, Phys. Rev. E 71, 046117 (2005).
[19] G. Tkacik, C. G. Callan, and W. Bialek, Proc. Natl. Acad. Sci. U.S.A. 105, 12265 (2008).
[20] P. Mehta, S. Goyal, T. Long, B. L. Bassler, and N. S. Wingreen, Mol. Syst. Biol. 5, 325 (2009).