Quantifying Systemic Evolutionary Changes by Color Coding Confidence-Scored PPI Networks

Phuong Dao, Alexander Schönhuth, Fereydoun Hormozdiari, Iman Hajirasouliha, S. Cenk Sahinalp, and Martin Ester

School of Computing Science, Simon Fraser University, 8888 University Drive Burnaby, BC, V5A 1S6, Canada

Abstract. A current major challenge in systems biology is to compute statistics on *biomolecular network motifs*, since this can reveal significant systemic differences between organisms. We extend the "color coding" technique to weighted edge networks and apply it to PPI networks where edges are weighted by probabilistic confidence scores, as provided by the *STRING* database. This is a substantial improvement over the previously available studies on, still heavily noisy, binary-edge-weight data. Following up on such a study, we compute the expected number of occurrences of non-induced subtrees with $k \leq 9$ vertices. Beyond the previously reported differences between unicellular and multicellular organisms, we reveal major differences between prokaryotes and unicellular eukaryotes. This establishes, for the first time on a statistically sound data basis, that evolutionary distance can be monitored in terms of elevated systemic arrangements.

Keywords. Biomolecular Network Motifs, Color Coding, Evolutionary Systems Biology, Protein-Protein Interaction Networks,

1 Introduction

A current major issue in evolutionary systems biology is to reliably quantify both organismic complexity and evolutionary diversity from a systemic point of view. While currently available biomolecular networks provide a data basis, the assessment of network similarity has remained both biologically and computationally challenging. Since currently available network data is still incomplete, simple edge statistics, for example, do not apply. Moreover, recent research has revealed that many biomolecular networks share global topological features which are robust relative to missing edges, which rules out many more straightforward approaches to the topic (see e.g. [13] for a related study on global features such as degree distribution, k-hop reachability, betweenness and closeness). On the more sophisticated end of the scale of such approaches would be attempts to perform and appropriately score alignments of the collection of all systemic subunits of two organisms. However, the development of workable scoring schemes in combination with related algorithms comes with a variety of obvious, yet unresolved, both biological and computational issues. Clearly, any such scoring schemes would already establish some form of condensed, systemic evolutionary truth by themselves.

This explains why recent approaches focused on monitoring differences between biomolecular networks in terms of *local structures*, which likely reflect biological arrangements such as functional subunits. A seminal study which reported that statistically overrepresented graphlets, i.e. small subnetworks, are likely to encode pathway fragments and/or similar functional cellular building blocks [17] sparked more general interest in the topic. In the meantime, to discover and to count *biomolecular network motifs* has become a thriving area of research which poses some intriguing algorithmic problems. As is summarized in the comprehensive review [6], such approaches are supported by various arguments.

In this paper, following up on a series of earlier studies, we will focus on physical protein-protein interaction (PPI) network data. Related studies focused on determining the number of all possible "induced" subgraphs in a PPI network, which already is a very challenging task. Przulj et al. [19] devised algorithms with which to count induced PPI subgraphs on up to $k = 5$ vertices. Recently developed techniques improved on this by counting induced subgraphs of size up to $k = 6$ [13] and $k = 7$ [11]. However, the running time of these techniques all increase exponentially with k. To count subgraphs of size $k \geq 8$ required novel algorithmic tools. A substantial advance was subsequently provided in [1] which introduced the "color coding" technique for counting non-induced occurrences of subgraph topologies in the form of bounded treewidth subgraphs, which includes trees as the most obvious special case. Counting non-induced occurrences of network motifs is not only challenging but also quite desirable since non-induced patterns are often correlated to induced occurrences of denser patterns. Trees in particular can be perceived as the backbones of induced dense and connected subgraphs and there is abundant evidence that dense and connected (induced) subgraphs reflect functional cellular building blocks (e.g. [26]). See also [7] for a successful approach to count all patterns of density at least 0.85 in several PPI networks as well as in synthetic networks from popular generative random models. See also [7] for a successful approach to count all patterns of density at least 0.85 in several PPI networks as well as in synthetic networks from popular generative random models.

While these studies successfully revealed differences between PPI networks of uni- and multicellular organisms, a binary edge has remained a notoriously noisy datum. However, none of the studies considered PPI networks with weighted edges where edge weights reflect the confidence that the interactions are of cellular relevance instead of being experimental artifacts. Weighted network data have recently become available and have already been employed for other purposes (see e.g. [22] and the references therein for a list of weighted network data sources). One of the main reasons for the lack of network motif studies on such data might be that to exhaustively mine biomolecular networks with probabilistic edge weights poses novel computational challenges.

In this paper, we show how to apply the "color coding" technique to networks with arbitrary edge weights and two different scoring schemes for weighted subgraphs. Edge weights are supposed to reflect our confidence in the interactions, as provided by the *STRING* database, and we will apply a scoring scheme which reflects our expectation[1] in entire subgraphs to be present or not. *STRING* is a major resource for assessments of

[1] Expectation is meant to be in the very sense of probability theory, by interpreting confidence scores as probabilities

protein interactions and/or associations predicted by large-scale experimental data (in the broad sense, including e.g. literature data) of various types (see [15] for the latest issue of STRING and the references therein for earlier versions). Here, we only employ experimentally determined physical protein interactions in order to both follow up on the recent discussions and to avoid inconsistencies inherent to using several types of protein interactions at once. Clearly, statistics on such networks will establish substantial improvements over studies on binary network data in terms of statistical significance and robustness.

We compute the expected number of non-induced occurrences (E-values) of tree motifs G' ("treelets") with k vertices in a network G with n vertices in time polynomial in n, provided $k = O(\log n)$. Note that, in binary edge weight graphs, computation of the number of expected occurrences and counting occurrences is equivalent when interpreting an edge to be an interaction occurring with probability 1. This provides the basis on which we can benchmark our results against previous studies. We use our algorithm to obtain normalized treelet distributions, that is the sum of the weights of non-induced occurrences of different tree topologies of size $k = 8, 9$[2] normalized by the total weight of all non-induced trees of size $8, 9$ for weighted PPI networks. We analyze the prokaryotic, unicellular organisms (*E.coli*, *H.pylori*), *B. subtilis* and *T. pallidum*, which are all quite similar, the eukaryotic unicellular organism *S.cerevisiae* (Yeast), and a multicellular organism (*C.elegans*). Beyond the previously reported similarities among the prokaryotic organisms, we were able to also reveal strong differences between Yeast and the prokaryotes. As before, statistics on C.elegans are still different from all other ones. As a last point, we demonstrate that our weighted treelet distributions are *robust* relative to reasonable amounts of network sparsification as suggested by [12].

To summarize, we have presented a novel randomized approximation algorithm to count the weight of non-induced occurrences of a tree T with k vertices in a weighted-edge network G with n vertices in time polynomial with n, provided $k = O(\log n)$ for a given error probability and an approximation ratio. We prove that resulting weighted treelet distributions are robust and sensitive measures of PPI network similarity. Our experiments then confirm, for the first time on a statistically reliable data basis, that uni- and multicellular organisms are different on an elevated systemic cellular level. Moreover, for the first time, we report such differences also between pro- and eukaryotes.

Related Work

Flum and Grohe [10] showed that the problem of counting the number of paths of length k in a network is $\#W[1]$-complete. Thus it is unlikely that one can count the number of paths of length k efficiently even for small k. The most recent approaches such as [5, 23] offer a running time of $O(n^{k/2+O(1)})$. As mentioned before, [19, 11, 13] describe practical approaches to counting all induced subgraphs of at most $k = 5, 6$ and $k = 7$ vertices in a PPI network.

[2] We recall that there are 23 resp. 47 different tree topologies on 8 resp. 9 nodes, see e.g. [18].

Approximate counting techniques have been devised which were predominantly based on the color coding technique as introduced by Alon et al. [3]. Color coding is based on assigning random colors to the vertices of an input graph and, subsequently, counting only "colorful" occurrences, that is, subgraphs where each vertex has a different color. This can usually be done in polynomial time. Iterating over different color assignments sufficiently many, but still a polynomial number of times yields statistically reliable counts. In the seminal study [3], color coding was used to detect (but not to count) simple paths, trees and bounded treewidth subgraphs in unlabelled graphs.

Scott et al. [20], Shlomi et al. [21] and Huffner et al. [14] designed algorithms for querying paths within a PPI network. More recently, Dost et al. [9] have extended these algorithms in the QNet software to allow searching for trees and bounded treewidth graphs. Arvind and Raman [4] use the color coding approach to count the number of subgraphs in a given graph G which are isomorphic to a *bounded treewidth graph* H. The framework which they use is based on approximate counting via sampling [16]. However, even when $k = O(\log n)$, the running time of this algorithm is *super-polynomial* with n, and thus is not practical. Alon and Gutner [2] derandomized the color coding technique by the construction of balanced families of hash functions. Recently, Alon et al. [1] presented a randomized approximation algorithm that, given an additive error ϵ and error probability δ, with success probability $1 - \delta$, outputs a number within ϵ times the number of non-induced occurrences of a tree T of k vertices in a graph G of n vertices running in time $O(|E| \cdot 2^{O(k)} \cdot \log(1/\delta) \cdot \frac{1}{\epsilon^2})$. Note that if $k = O(\log n)$ and ϵ, δ are fixed, this results in a polynomial time algorithm.

Note that all the previous works tried either to count exactly the total weight of non-induced occurrences of a pattern in a given graph or to approximate the occurrences where the weights of all edges are 1. The exact counting methods, due to parameterized complexity as mentioned earlier, give exponential running time even for paths of $k = O(\log n)$ vertices. We will give approximate based counting methods that offer polynomial running time given that patterns are trees of $k = O(\log n)$ vertices, a fixed approximation factor and an error probability.

2 Methods

In the following, let $G = (V, E)$ be a graph on n vertices and $w : E \to \mathbb{R}$ be an edge-weight function. Let T be a tree on k vertices where, in the following, $k = O(\log n)$. We define $\mathcal{S}(G, T)$ to be the set of non-induced subgraphs of G which are isomorphic to T and let $E(H)$ to be the edges of such a subgraph $H \in \mathcal{S}$. We extend w to weight functions on the members of $\mathcal{S}(G, T)$ by either defining

$$w(H) = \prod_{e \in E(H)} w(e) \quad \text{or} \quad w(H) = \sum_{e \in E(H)} w(e) \tag{1}$$

Note that if $w(e)$ is interpreted as the probability that e is indeed present in G then, assuming independence between the edges, $w(H)$ of the first case is just the probability that H is present in G. In the following, we will focus on the first case. Proofs for the second choice of $w(H)$ can be easily obtained, mutatis mutandis, after having

replaced multiplication by addition in the definition of $w(H)$. Finally, let $w(G,T) = \sum_{H \in \mathcal{S}(G,T)} w(H)$ be the total weight of non-induced occurrences of T in G. We would like to provide reliable estimates $\hat{w}(G,T)$ on $w(G,T)$. Note that $w(G,T)$ is the number of expected occurrences of T in G due to the linearity of expectation.

Consider Fig. 1. Here, T is a star-like tree on 4 vertices. There are two subgraphs H and H' in G which are isomorphic to T; therefore, $w(G,T) = w(H) + w(H')$. In the case that the weight of a subgraph in G is calculated as the product of the weights of its edges, we have $w(G,T) = w(H) + w(H') = 0.5 \times 0.4 \times 0.3 + 0.7 \times 0.8 \times 0.9$. In the other case, we have $w(G,T) = w(H) + w(H') = (0.5 + 0.4 + 0.3) + (0.7 + 0.8 + 0.9)$.

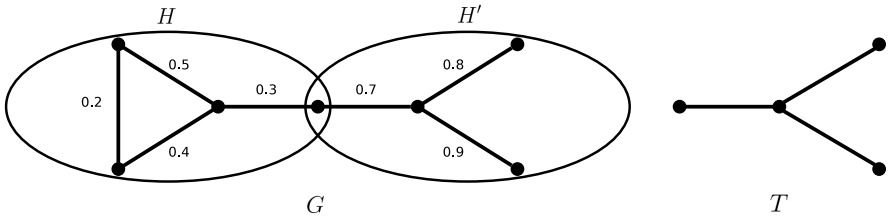


Fig. 1. An example of counting the total weight of non-induced subgraphs in a given network G which are isomorphic to a query tree T.

In the following, in order to estimate $w(G,T)$ by color coding, we will randomly assign k colors to the vertices of G where k is the size of T. Therefore, we introduce the notations $[k] = \{1, ..., k\}$ for the set of k colors and $\mathcal{S}(G,T,[k])$ for the set of all non-induced subgraphs of G which are *colorful* in terms of $[k]$, that is occurrences of T where each vertex has been assigned to a different color.

The following algorithm APPROXWEIGHTEDOCCUR, when given an approximation factor ϵ and an error probability δ, computes an estimate $\hat{w}(G,T)$ of $w(G,T)$ efficiently in n and k, given that $k = O(\log n)$ such that with probability $1 - 2\delta$, $\hat{w}(G,T)$ lies in the range $[(1-\epsilon)w(G,T), (1+\epsilon)w(G,T)]$.

Algorithm 1 APPROXWEIGHTEDOCCURR (G, T, ϵ, δ)

$G = (V,E)$, $k \leftarrow |V(T)|$, $t \leftarrow \log(1/\delta)$, $p \leftarrow k!/k^k$, $s \leftarrow 4/(\epsilon^2 p)$
for $i = 1$ to t **do**
  $Y_i \leftarrow 0$
  for $j = 1$ to s **do**
    Color each vertex of G independently and uniformly at random with one of k colors
    $X \leftarrow$ total weight of colorful subgraphs of G which are isomorphic to T
    $Y_i \leftarrow Y_i + X$
  end for
  $Y_i \leftarrow Y_i / s$
end for
$Z \leftarrow$ median of $Y_1 \ldots Y_t$
Return Z/p as the estimate $\hat{w}(G,T)$ of $w(G,T)$

The following lemmata give rise to a theorem that supports our claims from above.

Lemma 1. *The algorithm* ApproxWeightedOccurr *(G, T, ϵ, δ) returns $\hat{w}(G,T)$ such that with probability at least $1-2\delta$ we have $(1-\epsilon)w(G,T) \leq \hat{w}(G,T) \leq (1+\epsilon)w(G,T)$*

Proof. The proof proceeds quite similar to that of [1], after having replaced x_F there by x_H here where x_H is the indicator random variable whose value is $w(H)$ if H is colorful in a random coloring of G and 0 otherwise. Similarly, we then define $X = \sum_{H \in \mathcal{S}(G,T)} x_H$ which is the random variable that counts the total weight of colorful subgraphs of G which are isomorphic to T. The expected value of X then evaluates as

$$E(X) = E(\sum_{H \in \mathcal{S}(G,T)} x_H) = \sum_{H \in \mathcal{S}(G,T)} E(x_H) = \sum_{H \in \mathcal{S}} w(H)p = w(G,T)p \quad (2)$$

where $p = k!/k^k$ is the probability that the vertices of a subgraph H of size k are assigned to different colors. To obtain a bound on the variance $Var(X)$ of X, one observes that $Var(x_H) = E(x_H^2) - E^2(x_H) \leq E(x_H^2) = [w(H)]^2 p$. Moreover, the probability that both H and H' are colorful is at most p which implies

$$Cov(x_H, x_{H'}) = E(x_H x_{H'}) - E(x_H)E(x_{H'}) \leq E(x_H x_{H'}) \leq w(H)w(H')p.$$

Therefore, in analogy to [1], the variance of X satisfies $Var(X) = (\sum_{H \in \mathcal{S}} w(H))^2 p = w^2(G,T)p$. Since Y_i is the average of s independent copies of random variable X, we have $E(Y) = E(X) = w(G,T)p$ and $Var(Y_i) = Var(X)/s \leq w^2(G,T)p/s$. Again in analogy to [1], we obtain $P(|Y_i - w(G,T)p| \geq \epsilon w(G,T)p) \leq \frac{1}{4}$.

Thus, with constant error probability, Y_i/p is an ϵ-approximation of $w(G,T)$. To obtain error probability $1-2\delta$, we compute t independent samples of Y_i (using the first for loop) and replace Y_i/p by Z/p where Z is the median of Y_i's. The probability that Z is less than $(1-\epsilon)w(G,T)p$ is the probability that at least half of the copies of Y_i computed are less than Z, which is at most $\binom{t}{t/2}4^{-t} \leq 2^{-t}$. Similarly we can estimate the probability that Z is bigger than $(1+\epsilon)w(G,T)p$. Therefore, if $t = \log(1/\delta)$ then with probability $1-2\delta$ the value of $\hat{o}$ will lie in $[(1-\epsilon)w(G,T), (1+\epsilon)w(G,T)]$. $\diamond$

We still need to argue that given the graph G where each vertex is colored with one of k colors, we can compute the total weight of all non-induced colorful occurrences $w(G,T,[k])$ of T in G which refers to the variable X in the second for loop efficiently.

Lemma 2. *Given a graph G where each vertex has one of k colors, we can estimate $w(G,T,[k])$ in time $O(|E| \cdot 2^{O(k)})$.*

Proof. We pick a vertex ρ of T and consider T_ρ to be a rooted version of the query tree T with designated root ρ. We will compute $w(G,T_\rho,[k])$ recursively in terms of subtrees of T_ρ; so let $T'_{\rho'}$ be any subtree T' of T with designated root ρ'. Let $C \subset [k]$ and $\mathcal{S}(G,T'_{\rho'},v,C)$ be the set of all non-induced occurrences of $T'_{\rho'}$ in G which are rooted at v and colorful with colors from C and $w(v,T'_{\rho'},v,C) = \sum_{H \in \mathcal{S}(G,T'_{\rho'},v,C)} w(H)$ to be the total weight of all such occurrences. We observe that

$$w(G,T,[k]) = \frac{1}{q} \sum_{v \in G} w(G,T_\rho,v,[k]) \quad (3)$$

where q is equal to one plus the number of vertices ϱ in T for which there is an automorphism that ρ is mapped to ϱ. For example, if T in Figure 2 is rooted at ρ'', q is equal to 3. The key observation is that we can compute $w(G, T_\rho, v, [k])$ or the total weight of colorful non-induced subtrees rooted at v in G which are isomorphic to T_ρ in terms of total weight of colorful non-induced occurrences of subtrees of T_ρ in G. Let $T'_{\rho'}$ be an

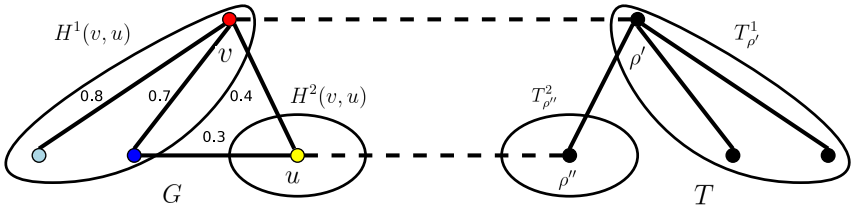


Fig. 2. Counting the total weight of colorful non-induced occurrences of T in G by counting the total weight of colorful non-induced occurrences of subtrees $T_{\rho'}$ and $T_{\rho''}$ in G.

arbitrary rooted subtree of T_ρ. We decompose $T'_{\rho'}$ into two smaller subtrees and count total weight colorful non-induced occurrences of these subtrees in G as follows. We choose a child ρ'' of ρ' and, by removing the edge between ρ' and ρ'' to decompose $T'_{\rho'}$ into two rooted subtrees $T^1_{\rho'}$ that do not contain ρ'' and $T^2_{\rho''}$ that do not contain ρ' for example Figure 2. Analogously, for every neighbor u of v in G, we denote a colorful copy of $T^1_{\rho'}$ at v by $H^1(v,u)$ and a colorful copy of $T^2_{\rho''}$ at u by $H^2(v,u)$. To obtain a copy H of $T'_{\rho'}$ in G by combining $H^1(v,u)$ and $H^2(v,u)$, $H^1(v,u)$ and $H^2(v,u)$ must be colorful for color sets $C_1(v,u), C_2(v,u)$ such that $C_1(v,u) \cap C_2(v,u) = \emptyset, C_1 \cup C_2 = C$ where the cardinality of C is the number of vertices of $T'_{\rho'}$. Finally, independent of the choice of u, we have

$$w(H) = w(H^1(v,u))w(H^2(v,u))w(vu) \tag{4}$$

To initialize the base case of single-vertex trees $T'_{\rho'}$, we set $w(G, T'_{\rho'}, v, \{i\}) = 1$ if the color of v is i; otherwise 0. In general, we have

$$\begin{aligned} w(G, T'_{\rho'}, v, C) &= \frac{1}{d} \sum_{u \in N(v)} \sum_{\substack{C_1 \cap C_2 = \emptyset \\ C_1 \cup C_2 = C}} \sum_{\substack{H^1 \in \mathcal{S}(G, T^1_{\rho'}, v, C_1) \\ H^2 \in \mathcal{S}(G, T^2_{\rho''}, u, C_2)}} w(H^1(v,u))w(H^2(v,u))w(vu) \\ &= \frac{1}{d} \sum_{u \in N(v)} \sum_{\substack{C_1 \cap C_2 = \emptyset \\ C_1 \cup C_2 = C}} w(G, T^1_{\rho'}, v, C_1)w(vu)w(G, T^2_{\rho''}, u, C_2). \end{aligned} \tag{5}$$

Note that $w(H)$ will, as in the summands, appear exactly d times across the different suitable choices of color sets C_1, C_2. For example, H in Fig. 2, rooted at v, is a colorful copy of a star-like rooted tree T with three leaves. There are three different ways by

which one can decompose H into a path of length 2, denoted by H_1 and a single node H_2, meaning that $d = 3$ in this case. Now observe that the weight of H $w(H)$ will appear three times as a summand in the above summation scheme, according to the three different decompositions of H. The proof for the case of additive weight schemes proceeds mutatis mutandis, after having replaced multiplication by addition and adjusted the base case ($w(G, T'_{\rho'}, v, \{i\} = 0$ regardless of the color of v). Let $N(v)$ be the set of neighbors of v, we have

$$w(G, T'_{\rho'}, v, C)$$
$$= \frac{1}{d} \sum_{u \in N(v)} \sum_{\substack{C_1 \cap C_2 = \emptyset \\ C_1 \cup C_2 = C}} n_2 w(G, T^1_{\rho'}, v, C_1) + n_1 n_2 w(vu) + n_1 w(G, T^2_{\rho''}, u, C_2) \quad (6)$$

where $n_1 = |\mathcal{S}(G, T^1_{\rho'}, v, C_1)|, n_2 = |\mathcal{S}(G, T^2_{\rho''}, u, C_2)|$ are the cardinalities of the respective sets of colorful copies of $T^1_{\rho'}$ resp. $T^2_{\rho''}$ rooted at v resp. u which can be computed efficiently [1] and parallely with $w(G, T^1_{\rho'}, v, C_1)$ and $w(G, T^2_{\rho''}, u, C_2)$.

Note that each $w(G, T'_{\rho'}, v, C)$ can be computed in $O(deg(v) \cdot 2^{O(k)})$ time where $deg(v)$ is the degree of v. Thus, the computation of total weight of colorful non-induced occurrences of T in G is in $O(|E|2^{O(k)})$ time. $\diamond$

Theorem 1. *The algorithm* APPROXWEIGHTEDOCCURR *(G, T, ϵ, δ) estimates the total weight of non-induced occurrences of a tree T in G with additive error ϵ and with probability at least $1 - 2\delta$ and runs in time $O(|E| \cdot 2^{O(k)} \cdot \log(1/\delta) \cdot \frac{1}{\epsilon^2})$ where $|E|$ is the number of edges in the input network.*

Proof. Now we only need to consider its running time. Notice that we need to repeat the color coding step and counting step $s \cdot t$ times and each iteration runs in time $O(|E| \cdot 2^{O(k)})$ where $|E|$ is the number of edges in the input network. Thus, since $p = k^k/k! = O(e^k) = O(2^{O(k)})$, the asymptotic running time of our algorithm evaluates as

$$O(s \cdot t \cdot |E| \cdot 2^{O(k)}) = O(|E| \cdot 2^{O(k)} \cdot \log(1/\delta) \cdot \frac{1}{\epsilon^2 p}) = O(|E| \cdot 2^{O(k)} \log(1/\delta) \cdot \frac{1}{\epsilon^2}). \quad (7)$$

3 Results

3.1 Data and Implementation

Weighted PPI Networks We downloaded PPI networks with confidence scores from the *STRING* database, version 8.0 [15] for the prokaryotic, unicellular organisms *E.coli, H.pylori, B.subtilis, T. pallidum*, the eukaryotic, unicellular organism *S.cerevisiae* (Yeast) and the eukaryotic, multicellular organism *C.elegans*. Edge weights exclusively reflect confidence in experimentally determined physical interactions to be functional cellular entities and not only experimental artifacts (see [25, 24] for detailed information). See Table 1 for some basic statistics about these networks.

Query Tree Topologies There are 23 and 47 possible tree topologies with 8 and 9 nodes respectively. We obtained the list of treelets from the Combinatorial Object

Table 1. Number of vertices, edges, in the studied PPI networks.

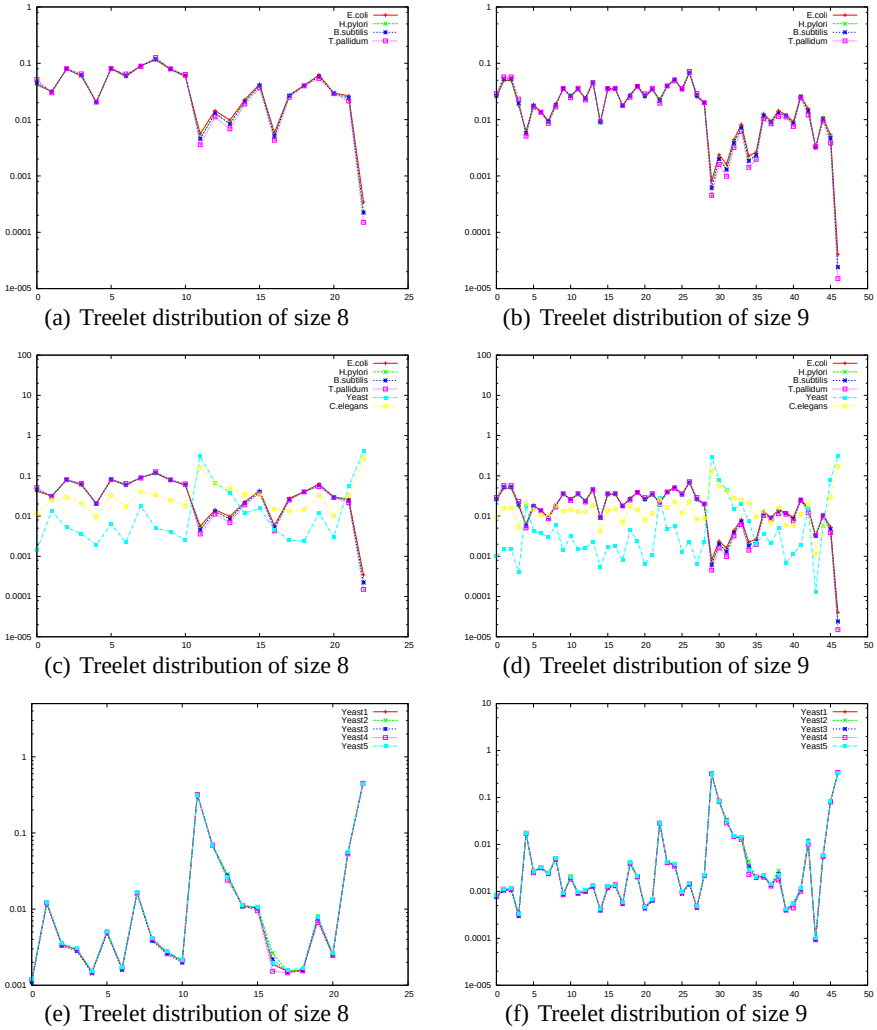
	E.coli	*H.pylori*	*B. subtilis*	*T. pallidum*	*S.cerevisiae*	*C.elegans*
Vertices	2482	1019	939	398	5913	5353
Edges	22476	9274	9184	4198	207075	43100

Server [18]. See our supplementary website [8] for diagrams of the respective tree topologies.

Implementation / Choice of Parameters We implemented our algorithm APPROX-WEIGHTOCCUR with the multiplicative weight scheme such that the weight of a query tree can be interpreted as its probability to be present in the network, as aforementioned. We set $\epsilon = 0.01$ as the approximation ratio and $\delta = 0.001$ as the error probability. Then we computed expected numbers of occurrences of all query trees of size 8 and 9 for each of the networks described above. By normalizing the occurrences of the different query trees of size 8 resp.9 over the 23 resp. 47 different query trees, we obtained size 8 resp. size 9 treelet distributions which we refer to as *normalized weighted treelet distributions*. The idea behind normalizing expected occurrences is for comparing PPI networks with different number of nodes and edges and to increase robustness with respect to missing data which still is a considerable issue in PPI network studies. We will demonstrate the robustness by an approved series of experiments [12] in the subsequent subsection 3.3. Experiments were performed on a Sun Fire X4600 Server with 64GB RAM and 8 dual AMD Opteron CPUs with 2.6 Ghz speed each.

3.2 Comparison of PPI networks

In order to be able to appropriately benchmark our results against previous findings we considered the same organisms that were examined in [1]. We also considered the two prokaryotic organisms *B.subtilis* (a Gram-negative bacterium commonly found in soil) and *T.pallidum* (a Gram-negative pathogen giving rise to congenital syphilis). The corresponding weighted treelet distributions are displayed in Fig. 3. The upper row of figures shows that the treelet distributions of the prokaryotic organisms are all similar. This is quite amazing since the weighted PPI networks have been determined in experiments which were independent of one another and without the integration of cross-species associations [15]. As can be seen in the middle row of Fig. 3, the treelet distributions of the Yeast PPI network is quite different from the ones of the prokaryotic organisms, which had not been observed in the boolean networks used in [1]. Still, there are obvious differences between the unicellular organisms and *C.elegans*, the multicellular model organism under consideration. It might be interesting to note that the greatest differences occur for the expected numbers of occurrences of tree topologies 23 resp. 47, which are the stars with 8 resp. 9 nodes. As a last point, note that global features such as degree and clustering coefficient distributions of these networks do not differ much (see the supplementary materials [8] for respective results).



(a) Treelet distribution of size 8 (b) Treelet distribution of size 9

(c) Treelet distribution of size 8 (d) Treelet distribution of size 9

(e) Treelet distribution of size 8 (f) Treelet distribution of size 9

Fig. 3. Normalized weighted treelet distributions of the prokaryotes *H.pylori*, *E.coli*, *B.subtilis*, *T. pallidum* PPI networks (top row) and of the prokaryotes *H.pylori*, *E.coli*, *B.subtilis*, *T. pallidum*, *S.cervisiae* (Yeast) and *C.elegans* PPI networks (middle row). and of five networks (a) size 8, (b) size 9 generated from the *S.cervisiae* Yeast PPI network as outlined in ssec. 3.3 (bottom row).

3.3 Robustness Analysis

In order to assess the reliability of the normalized weighted treelet distributions as a measure of weighted PPI network similarity one needs to ensure that they are robust w.r.t. small alterations to the network. This is motivated by that there still might be some insecurity about the amount and the quality of currently available PPI data. In this section, we evaluate the robustness of normalized weighted treelet distributions meaning that minor changes in the weighted PPI networks do not result in drastic changes in their normalized weighted treelet distributions.

Therefore, we used the random sparsification method which was proposed in [12] and was applied in earlier studies [1]. The method iteratively sparsifies networks by removing vertices and edges in a sampling procedure and specifically addresses the pecularities of experimentally generating PPI networks. It is based on two parameters, the bait sampling probability α_b and the edge sampling probability α_e which refer to sampling vertices and edges. As in [1], we set $\alpha_b = 0.7$ and $\alpha_e = 0.7$ and shrank the weighted PPI network of Yeast to five smaller networks accordingly. A comparison of the normalized weighted treelet distributions of the shrunken networks is displayed in the bottom row of Fig. 3. As can be seen, the normalized weighted treelet distributions are very similar to one another which confirms that normalized treelet distributions are robust w.r.t. experimentally induced noise and/or missing data.

4 Conclusions

To quantify organismic complexity and evolutionary diversity from a systemic point of view poses challenging biological and computational problems. Here, we have investigated *normalized weighted treelet distributions*, based on exploration of PPI network whose edges are assigned to confidence scores, which can be retrieved from the STRING database as an appropriate measure. As a theoretical novelty, we have extended the color coding technique to weighted networks. As a result, we were able to reveal differences between uni- and multicellular as well as pro- and eukaryotic organisms. Systemic differences based on local features in PPI networks between pro- and eukaryotes had not been reported before. In sum, our study reveals novel systemic differences and confirms previously reported ones on a substantially more reliable data.

References

1. N. Alon, P. Dao, I. Hajirasouliha, F. Hormozdiari, and S.C. Sahinalp. Biomolecular network motif counting and discovery by color coding. *Bioinformatics*, 24(Supp. 1):i241–i249, 2008.
2. N. Alon and S. Gutner. Balanced families of perfect hash functions and their applications. *Proc. ICALP*, pages 435–446, 2007.
3. Noga Alon, Raphael Yuster, and Uri Zwick. Color-coding. *J. ACM*, 42(4):844–856, 1995.
4. V. Arvind and V. Raman. Approximation algorithms for some parameterized counting problems. In *ISAAC '02: Proceedings of the 13th International Symposium on Algorithms and Computation*, pages 453–464, 2002.
5. A. Bjorklund, T. Husfeldt, P. Kaski, and M. Koivisto. The fast intersection transform with applications to counting paths. *http://arxiv.org/abs/0809.2489*, 2008.

6. G. Ciriello and C. Guerra. A review on models and algorithms for motif discovery in protein-protein interaction networks. *Briefings in Functional Genomics and Proteomics*, 2008.
7. R. Colak, F. Hormozdiari, F. Moser, A. Schönhuth, J. Holman, S.C. Sahinalp, and M. Ester. Dense graphlet statistics of protein interaction and random networks. *Proceedings of the Pacific Symposium on Biocomputing*, 14:178–189, 2009.
8. P. Dao, A. Schoenhuth, F. Hormozdiari, I. Hajirasouliha, S.C. Sahinalp, and M. Ester. Quantifying systemic evolutionary changes by color coding confidence-scored ppi networks. *Supplementary Materials*, 2009. http://www.cs.sfu.ca/ pdao/personal/weightedmotifsup.pdf.
9. Banu Dost, Tomer Shlomi, Nitin Gupta 0002, Eytan Ruppin, Vineet Bafna, and Roded Sharan. Qnet: A tool for querying protein interaction networks. In *RECOMB*, pages 1–15, 2007.
10. J. Flum and M. Grohe. The parameterized complexity of counting problems. *SIAM J. Comput.*, 33:892922, 2004.
11. Joshua A. Grochow and Manolis Kellis. Network motif discovery using subgraph enumeration and symmetry-breaking. In *RECOMB*, pages 92–106, 2007.
12. J. Han, D. Dupuy, N. Bertin, M. Cusick, and M. Vidal. Effect of sampling on topology predictions of protein-protein interaction networks. *Nature Biotech*, 23:839–844, 2005.
13. F. Hormozdiari, P. Berenbrink, N. Przulj, , and S.C. Sahinalp. Not all scale-free networks are born equal: The role of the seed graph in ppi network evolution. *PLoS Comput Biol*, 3(7), 2007.
14. Falk Huffner, Sebastian Wernicke, and Thomas Zichner. Algorithm engineering for color coding with applications to signaling pathways. *Algorithmica*, 52(2):114–132, 2008.
15. L.J. Jensen, M. Kuhn, M. Stark, S. Chaffron, C. Creevey, J. Muller, T. Doerke, P. Julien, A. Roth, M. Simonovic, P. Bork, and C. von Mering. String 8—a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Research*, 37(Database Issue):D412–D416, 2009.
16. Richard M. Karp and Michael Luby. Monte-carlo algorithms for enumeration and reliability problems. In *FOCS*, pages 56–64, 1983.
17. R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.
18. University of Victoria. Combinatorial object server. *http://www.theory.csc.uvic.ca/ cos*, since 1995.
19. Natasa Przulj, Derek G. Corneil, and Igor Jurisica. Modeling interactome: scale-free or geometric? *Bioinformatics*, 20(18):3508–3515, 2004.
20. J. Scott, T. Ideker, R. M. Karp, and R. Sharan. Efficient algorithms for detecting signaling pathways in protein interaction networks. *J Comput Biol*, 13(2):133–144, 2006.
21. Tomer Shlomi, Daniel Segal, Eytan Ruppin, and Roded Sharan. Qpath: a method for querying pathways in a protein-protein interaction network. *BMC Bioinformatics*, 7:199, 2006.
22. I. Ulitsky and R. Shamir. Identifying functional modules using expression profiles and confidence-scored protein interactions. *Bioinformatics*, 2009. doi:10.1093/bioinformatics/btp118.
23. V. Vassilevska and R. Wiliams. Finding, minimizing and counting weighted subgraphs. *Proceedings of the Symposium of the Theory of Computing (STOC)*, 2009. To appear.
24. C. von Mering and L.J. Jensen. News about the string and stitch databases. *http://string-stitch.blogspot.com/*, 2008.
25. C. von Mering, L.J. Jensen, B. Snel, S.D. Hooper, M. Krupp, M. Foglierini, N. Jouffre, M.A. Huynen, and P. Bork. String: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Research*, 33(Database Issue):D433–D437, 2005.
26. X. Zhu, M. Gerstein, and M. Snyder. Getting connected: analysis and principles of biological networks. *Genes and Development*, 21:1010–1024, 2007.