Discovering Weighted Motifs in Gene co-expression Networks

Sarvenaz Choobdar, Pedro Ribeiro, and Fernando Silva
CRACS and INESC-TEC
Faculdade de Ciencias, Universidade do Porto, Portugal
Email:{sarvenaz,pribeiro,fds}@dcc.fc.up.pt

ABSTRACT

An important dimension of complex networks is embedded in the weights of its edges. Incorporating this source of information on the analysis of a network can greatly enhance our understanding of it. This is the case for gene co-expression networks, which encapsulate information about the strength of correlation between gene expression profiles. Classical unweighted gene co-expression networks use thresholding for defining connectivity, losing some of the information contained in the different connection strengths. In this paper, we propose a mining method capable of extracting information from weighted gene co-expression networks. We study groups of differently connected nodes and their importance as *network motifs*. We define a subgraph as a motif if the weights of edges inside the subgraph hold a significantly different distribution than what would be found in a random distribution. We use the Kolmogorov-Smirnov test to calculate the significance score of the subgraph, avoiding the time consuming generation of random networks to determine statistic significance. We apply our approach to gene co-expression networks related to three different types of cancer and also to two healthy datasets. The structure of the networks is compared using *weighted motif profiles*, and our results show that we are able to clearly distinguish the networks and separate them by type. We also compare the biological relevance of our weighted approach to a more classical binary motif profile, where edges are unweighted. We use shared Gene Ontology annotations on biological processes, cellular components and molecular functions. The results of gene enrichment analysis show that weighted motifs are biologically more significant than the binary motifs.

Keywords

Complex Networks, Network Motifs, Weighted Networks, Gene Co-expression Network

1. INTRODUCTION

SAC'15 April 13-17, 2015, Salamanca, Spain.

Gene co-expression networks (GCNs) are powerful structures to represent relationships between genes. In these networks, nodes represent genes, or more specifically, gene expression profiles on a microarray experiment. Network connections correspond to correlations between a pair of gene profiles under the chosen samples [12]. The classical approach for network construction is to use binary information to encode the relationships, with a pair of genes being either connected or not [30]. This implies the usage of a threshold filter to decide the presence of an edge. In this process we may lose information, since the correlation metrics are not binary, but rather produce real values that can express the strength of connections. This lead to a richer network setup where a gene co-expression is encoded by a weight value, rather than a boolean one [33]. The study of these types of networks is known as *Weighted Gene Co-Expression Network Analysis* (WGCNA).

An important goal of studying GCNs is to predict gene functions and disease biomarkers such as the discovery of cancer related genes [26, 34]. In this paper, we are interested in studying the structure of gene co-expression networks across healthy tissues and disease associated ones. In particular, we would like to answer questions such as: what does a healthy network look like? How different is a healthy network from a cancer related one? What makes one GCN different from another across different cancer types? Is there subnetworks (groups of densely connected nodes in the network) in cancer sample networks that do not appear in healthy networks?

We study these questions by comparing networks using *network motifs* as small connected subgraphs representing characteristic patterns of a network. In this paper we study motifs in weighted networks which require a different problem setting from that of the original definition of motifs by Milo et al.[23]. Milo et. al. define motifs based on *subgraph frequency* and over-representation, thus motifs are subgraphs that appear more frequently than what would be expected. In order to incorporate the weight information, we identify motifs by comparing the weight distribution of edges within a subgraph type to a random distribution of weights. Our method calculates a significance value for each type of subgraph in order to build a *weighted motif profile* that can act as a fingerprint of the network, revealing classes. The basic intuition is that the important functions of the networks are correlated with relevant subgraphs. Only the subgraphs appearing in the original network with a significantly different distribution than the random distribution are selected as motifs. In our definition, it is not the quantity of sub-

graphs but the quality of relations within a subgraph that is of interest. Our concept of weighted motifs is therefore well suited to applications where the strength of relations between entities is more important, as is the case in WGCNA.

The contributions of our paper can be summarized as follows:

- We define a new method for motif mining in weighted networks based on distribution of weights of edges. The significance of a subgraph is evaluated by a pairwise comparison of edge weights inside the subgraph and the whole network.
- We compare gene co-expression networks of normal tissues and cancer associated ones by their motif profiles. The proposed weighted motif is capable of distinguishing networks by their types.
- We use gene ontology terms enrichment analysis to assess the biological relevance of discovered motifs in each network. We use a functioning score to compare the binary and weighted motif profiles in terms of their capability to determine the functionality of the disease-associated genes.

The remainder of this paper is organized as follows. Section 2 gives an overview of related work. Section 3 defines weighted motif and describes the proposed motif mining method. Section 4 discusses the experimental results and the evaluation of the developed method. Finally, section 5 concludes the paper, summarizing what was found and giving some possible future directions.

2. RELATED WORK

2.1 Microarray Data and Gene Co-Expression Networks

DNA microarrays are a powerful experimental tool producing valuable information for the study of cell components at different levels of genes, proteins or metabolites. A conventional approach to microarray data analysis is to use statistical methods such as a T-test or a F-test to identify genes that are differently expressed between groups of samples [4]. These differences are utilized to identify genes, that are capable of correctly assigning samples to different groups. A major drawback of these approaches is that the selected genes are usually not functionally related and thus cannot reveal key biological functions or processes associated with each disease.

Besides statistical methods, there are many other research works studying the interactions betweens genes to find biological functions. Several approaches have been adopted to study the collective behavior of individual regulatory pathways. In these methods, instead of studying single genes to identify functionally related genes, the focus is more on global gene expression patterns. The goal of gene co-expression network analysis in these methods is to identify groups of genes that are highly correlated regarding expression levels across multiple samples [13, 14, 18, 26, 34, 35].

Biological functionality can emerge from the interactions between the constituents of a cell, and in particular between genes. It has been shown that on average each gene has interactions with four to eight other genes [2] and is associated with ten biological functions [21]. For constructing a co-expression network, correlation coefficients such as Pearson's are used to compare the expression profile of pairs of genes[33]. When two genes are correlated, they are connected in the network.

A major challenge in genes co-expression data analysis is on how to define the threshold for genes connectivity such that the network still holds the actual properties that we want to study. In WGCNA, the network is built based on the concept of scale-free networks since in many metabolic networks the scale-free phenomena is observed [33].

There are other studies on weighted gene co-expression networks, such as the one by Zhao et al. [36]. They developed a series of methods to study the networks using a hierarchical clustering based approach. They propose several metrics to assess the similarity of genes and identifying groups of genes associated with diseases such as ASPM in glioblastoma [13].

Another line of research on wighted networks is about finding modules and substructures in the networks. Nepusz et al. [24] introduced a clustering method (ClusterONE) for detecting potentially overlapping protein complexes from protein-protein interaction data. Their method is based on the concept of the cohesiveness score which measures how likely it is for a group of proteins to form a protein complex. The cohesiveness of a group of proteins is proportional to the total weight of edges contained entirely by the group of proteins, and the total weight of edges that connect the group with the rest of the network. They use a greedy growth process to find groups in a protein-protein interaction network that are likely to correspond to protein complexes.

Frequent subgraph mining algorithms, designed for binary networks are not applicable for weighted networks as the anti-monotone property does not hold in weighted networks. This property reduced the search space and speeds up candidate generation and isomorphism test in subgraph mining algorithms. To address this issue, Jiang et al. [16] proposed a number of strategies to control candidate generation, namely ATW-gSpan, AW-gSpan and UBW-gSpan. All three algorithm are weighted variations of the gSpan algorithm [32]. They designed weighted support measures based on average weights in subgraphs.

2.2 Network Motifs

The original definition of network motifs was introduced by Milo et. al in 2002 as *patterns of interconnections occurring in complex networks at numbers that are significantly higher than those in randomized networks* [23]. This definition implies a hypothesis test to find out if the frequency of subgraph in the original network is larger than its frequency in a randomized network. The primary definition only considers the overrepresented subgraphs as motifs but underrepresented subgraphs, or anti-motifs, were later also considered [22]. One of the main applications of motif discovery using any algorithm is to compare networks in different domains such as biological [22] or social [6, 7] networks.

We note that network motifs, by their definition, are different from frequent subgraphs [15, 32] or substructures [11]. For an unweighted network with binary connections (where two nodes are either connected or not), motif mining consists essentially in enumerating all subgraphs of specific sizes in a network, and finding those that appear more frequently than expected [23]. This subgraph enumeration leads to a higher computational complexity for motif mining algo-

rithms, when compared to frequent subgraph mining algorithms where pruning criterions, such as the anti-monotonicity property, are used to limit the search space and to improve efficiency. Another restricting issue in motif mining is the calculation of random frequency of subgraphs. From a statistical point of view, the random frequency of a subgraph is reliable only if a reasonable number of random networks is generated for this purpose. These properties, imposed by the definition of motifs, make it computationally hard to increase the size of subgraphs.

In a weighted network, one requires a measure different from the usual frequency. Saramaki et al. [29] used the average of weights to find motifs in a network. They define two measures, intensity and coherence, based on the average of weights in instances of a particular subgraph type. A subgraph is a motif if these measurements differ from random values. Choobdar et al. [7] defined an entropy-based measurement to assess the significance of subgraphs. A subgraph whose weight entropy is different from the random entropy is called a motif. Our work takes a different approach, as we will see, by using the distribution of weights in the subgraphs.

3. MOTIFS IN WEIGHTED NETWORKS

Our goal is to find *weighted motifs* as sets of differently connected genes in weighted co-expression networks and to use their relative importance as a fingerprint of the network. Before starting the mining process, we first establish which sets of connected nodes are we going to use as subgraphs and we describe which subgraphs we are going to consider for enumerating in the original graph. Furthermore, we design a scoring function to assess the significance of each of these subgraph types. Finally, we discuss how the new significance scoring function is incorporated in the motif mining process.

3.1 Subgraphs Types and Enumeration

For the purposes of this paper, we will consider as motif candidates all possible 29 types of undirected subgraphs from sizes 3 to 5, as depicted in Fig. 1. There is nothing intrinsic in our methodology that forbids us from using even larger sizes, with the exception of potentially being computationally expensive to enumerate all their occurrences.

In each subgraph type we divide its edges in classes of equivalence according to the subgraph symmetry. For instance, there is only one type of edge on the clique of 4 nodes (`4-6` type) since all edges are topologically equivalent. The same can be said for the star subgraph of 4 nodes (`4-1` type). However, in the linear chain of 4 nodes (`4-2` type) there are two different edge types: the one between the middle nodes and the one between a middle node and a leaf node.

We explain in section 3.3 how we use g-tries [28] for storing and searching for subgraph occurrences.

3.2 Motif Significance Measure

Following the definition of motifs in unweighted networks, we define a subgraph as motif if the weights of the edges in the subgraph follow a significantly different distribution than a "similar" random distribution. In classical unweighted network motifs, the original null model proposed involved the creation of random networks with the same degree sequence as the original network [23]. This is to guarantee that the motif is really a characteristic of the network and not just

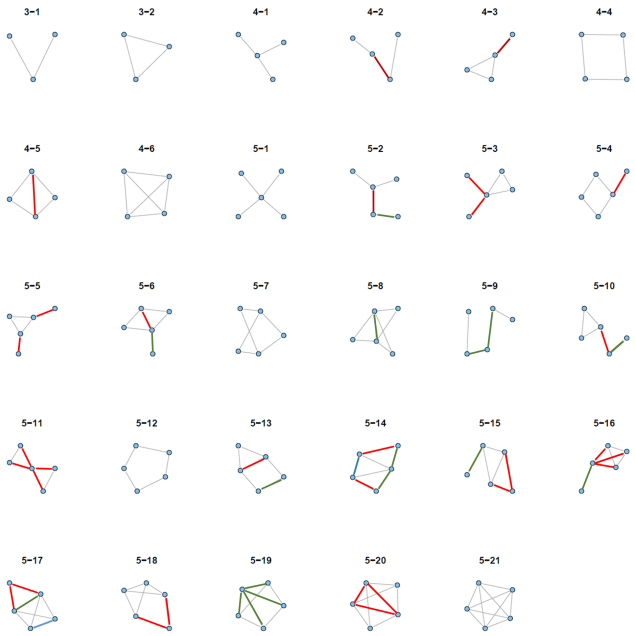

Figure 1: Set of subgraphs used for creating a motif profile of the network. Each motif is given an identification that we will use throughout the paper. Different topological classes of equivalence in the edges of a subgraph are distinguished by color and thickness.

a consequence of its global topological properties. Similarly, in our weighted case we also want to preserve certain global characteristics of the individual network being analysed, and thus we use the weight distribution over the whole network as a suitable random model.

Denoting the probability distribution of weights in a network by $P(w)$, the random distribution of weights in a subgraph with h edges is derived from $P_w(sg) = \prod^h P(w)$. Hence, a motif is a subgraph whose actual weight distribution in the subgraph sg is different from the random distribution, which uses the weights over the entire network.

There are several methods that one can use to compare the distributions of weights. Two notable examples are and Kulbeck-Leibler distance [17], or the Kolmogorov-Smirnov test [20]. We follow the univariate comparison where weight distributions are compared edge-wise. We use the two sample Kolmogorov-Smirnov test, which compares two samples regarding the location and shape of the empirical cumulative distribution functions of the two samples. For the univariate comparison, the actual weight distribution of every different edge type of subgraph sg is compared with the random distribution. The weighted motifs are those subgraphs for which the probability of holding a weight distribution different from the random distribution is higher than a significance value α. Hence a subgraph sg is a motif if:

$$max\{P(F_e(w_i) \neq F_r(w_i))|i \in E(sg)\} < \alpha \qquad (1)$$

where $F_e(w_i)$ and $F_r(w_i)$ are respectively the empirical and random distribution function of w_i, weights on edge i and $E(sg)$ is the set of edges in sg, that is, the set of classes of equivalence over all the edges of sg, as defined in the previous section.

Note that in this definition only subgraphs having different distributions over all edges are considered as motifs. An alternative would be to define motifs as subgraphs that have at least one edge with a different distribution. In either definition of motifs in weighted networks, the quality of relations within the subgraph is of interest to us, not its quantity in the network. This suits well for applications where the strength of connections is important, as it is the case in WGCNA.

We define the weighted score of subgraphs as follows:

$$w\text{-}score_k = argmax\{P(KS(w_i))|i \in E(sg)\} \quad (2)$$

where $KS(w_i)$ is the Kolmogorov-Smirnov statistic for distribution comparison of weights on edge i and it is equal to the maximum absolute difference between the empirical weight distribution and random distribution:

$$KS(w_i) = \max_{w \in w_i} |F_{empirical}(w) - F_{random}(w)| \quad (3)$$

and $P(KS(w_i))$ are the critical values, regarding the distribution of the KS statistic when $F_{empirical}(w_i) = F_{random}(w_i)$.

A *weighted motif profile* of the network can then be constructed as a feature vector containing the w-scores of all 29 subgraph types.

3.3 Weighted motif mining

The overall process for finding motifs of size k in a weighted network starts with finding all subgraphs of size k (storing the weight set over the edges for each subgraph type i), and then the weight distribution in occurrences of g_i^k is derived. This is a multivariate function whose dimension increases as the number of edges in subgraph increases. To find the weight distribution of a given subgraph, we use the stored weight sets while enumerating the instances of the subgraph in the original network.

We use g-tries [27] for storing and searching for subgraph occurrences. G-tries are multiway trees that are able to store a collection of subgraphs. Their basic principle is to identify common substructure. Subgraphs with the same parent g-trie node share the same topological structure with the exception of a single node and its connections, as is exemplified Figure 2.

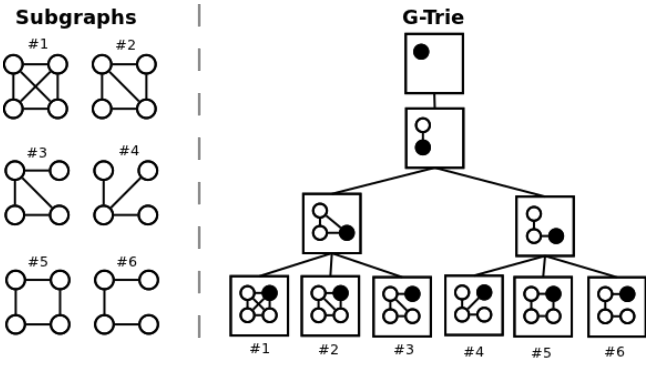

Figure 2: An example g-trie storing all possible undirected subgraphs of size 6. In each g-trie node, the black vertex is the new one being added, and the white vertices are the ones "inherited" from the parent g-trie nodes.

By using an efficient canonical labeling procedure and symmetry breaking conditions, g-tries allow the search at the same time for an entire set of subgraphs. This avoids the redundancy of searching several times for the same substructure that belongs to different subgraphs, as it would happen if we would search for each subgraph type individually, in a subgraph-centric algorithm such as Grochow and Kellis [10]. At the same time, g-tries also do isomorphism testing as we are traversing the g-trie tree, since when we are a at a leaf we can be certain that the subgraph found is of that type. This contrasts with network-centric methods such as ESU [31], which enumerate all connected sets of the desired number of vertices and postpone isomorphism tests to when an entire occurrence is found, not reusing information from previous isomorphisms found.

We modified the original g-tries algorithm so that we are able to store sets of edge weights for each subgraph type, instead of simple integer frequency. After discovering all occurrences of a subgraph g^k in the network, we evaluate the significance of each subgraph, measured trough a weighted score regarding equation 2 to compare the distribution of weights in the subgraph versus a random distribution.

We produce the *weighted motif profile* of the network by creating a feature vector containing all the individual weighted scores found. The end product of our methodology, the constructed profile, can be seen as a characteristic fingerprint of the network, which can be used, for instance, for network comparison purposes. An example of such an application is described in the results section.

4. EXPERIMENTAL RESULTS

4.1 Datasets

The NCBI Gene Expression Omnibus (GEO) is a very rich source for cancer microarray datasets. We queried GEO to retrieve data of various types of tumor biopsy samples. We selected microarray data for three cancer types, including lung cancer, breast cancer and neuroblastoma cancer, as depicted in Table 1). All the datasets include at least 30 samples in order to have reliable correlations between genes as mentioned in [19, 25]. We also retrieve two datasets of a normal "healthy" tissue microarray.

4.2 Weighted Network Construction

In the classic unweighted scenario, the co-expression network is constructed with nodes representing genes, and two nodes are connected if the corresponding genes are significantly co-expressed across chosen tissue samples. However, in such network construction it is important to know at what level of correlation two nodes must be connected to be biologically meaningful. Instead of a binary definition of connections between genes (connected=1, unconnected=0), we use a "soft thresholding" framework, as proposed by Zhang and Horvat et al. [33], to build weighted gene co-expression networks, where associated connections have a strength value. We should mention that our proposed motif mining method is independent of the network construction method and the input network can be built by any other method.

The similarity of genes, measured regarding their gene expression profiles, is used as the weight of connections in the network. Given two genes i and j, the similarity between

Table 1: The microarray datasets used for gene co-expression network construction.

GSE NO.	CancerType	SampleSize
GSE12460	neuroblastoma	64
GSE2570	neuroblastoma	38
GSE18864	breast cancer all types	84
GSE21653	medullary breast cancers	266
GSE10445	lung adenocarcinoma	72
GSE3141	lung	111
GSE10245	lung	58
GSE19804	lung	120
GSE10072	lung	107
GSE5056	lung	44
GSE1643	normal	40
GSE13564	normal	44

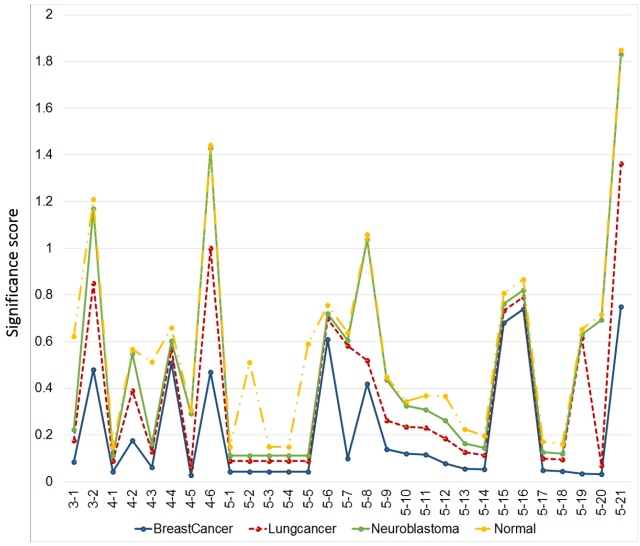

Figure 3: Weighted motif profiles of gene co-expression networks for each network type. The subgraphs score is the average for each network type .

them, s_{ij}, is defined as the absolute value of the Pearson correlation $s_{ij} = |cor(i, j)|$. Then, the similarity matrix by $S = [s_{ij}]$ is transformed to an adjacency matrix using a thresholding function defined as:

$$a_{ij} = |s_{ij}|^{\beta}$$

where a_{ij} is the weight of the connection between nodes i and j and β is the parameter chosen with the scale-free topology criterion. This is based on the fact that metabolic networks in all organisms have been suggested to be scale-free networks [5, 8, 9].

For each of the microarray dataset in Table 1, we build the adjacency matrix of all genes and then extract the network of 500 most connected genes in each dataset. We limit our study to this number of genes as our main concern in this paper is showing the applicability of our method and not computational issues. The larger the network, the longer the motif mining process will be.

4.3 Weighted Motif Results and Evaluation

We constructed weighted gene co-expression networks for all datasets using the method we already described. After that, we enumerated all 29 subgraphs types, stored the respective set of weights and we proceeded by computing the weighted score of each subgraph. Finally, we aggregated all the scores in one feature vector per network, creating an individual fingerprint for each co-expression network.

In order to evaluate the feasibility of our approach, we follow a network comparison scenario, by using the constructed weighted motif profiles to compare the structural patterns of healthy and disease-associated networks. Fig. 3 shows the average motif profiles we found on each type of network.

Fig. 4 is a heat map showing the similarity of gene co-expression networks for healthy tissues and cancer associated networks. The similarity of two networks is measured in terms of Euclidean distance of their weighted motif profiles. From this figure we can clearly see that the weighted motif profile is capable of distinguishing between different network classes. Each network type, including breast cancer, lung cancer, neuroblastoma and healthy tissues, are clearly separated into different groups.

Fig. 5 shows the most outstanding subgraphs (or motifs) in terms of differentiating gene co-expression networks. These subgraphs are selected based on principle component study over the motif profiles of networks. These subgraphs are those that make the most difference between network types regarding motif profiles in Fig. 3. Less dense subgraphs (`4-3`) and (`5-2`) are more significant in normal networks than the other types. Although in all cancer associated networks dense subgraphs like (`5-21`) are significant, there are some other types of subgraphs that distinguishes them between themselves. Subgraph (`5-7`) for breast cancer, subgraphs (`5-15`) and (`5-16`) for neuroblastoma and subgraph (tt 3-2) for lung cancer are outstanding.

In the next section, we study the significance of weighted network motifs in biological terms and compare binary motifs against our weighted motif profile.

4.4 Domain Based Evaluation

We use the domain-based metric to evaluate the discovered motif regarding their biological relevance. Every gene product is described in terms of their association to biological processes, cellular components and molecular functions. A biological process refers to entities at both the cellular and organism levels of granularity, cellular component refers to the localization of proteins inside the cell and molecular function refers to shared activities at the molecular level. The Gene Ontology (GO) [1] database provides vocabularies to describe functions of genes. We use GO term enrichment analysis to find out what function every motif is enriched in.

Only finding the relevant GO terms associated with a given gene list of each motif does not reveal the statisti-

[1]http://www.geneontology.org/GO.ontology.structure.shtml

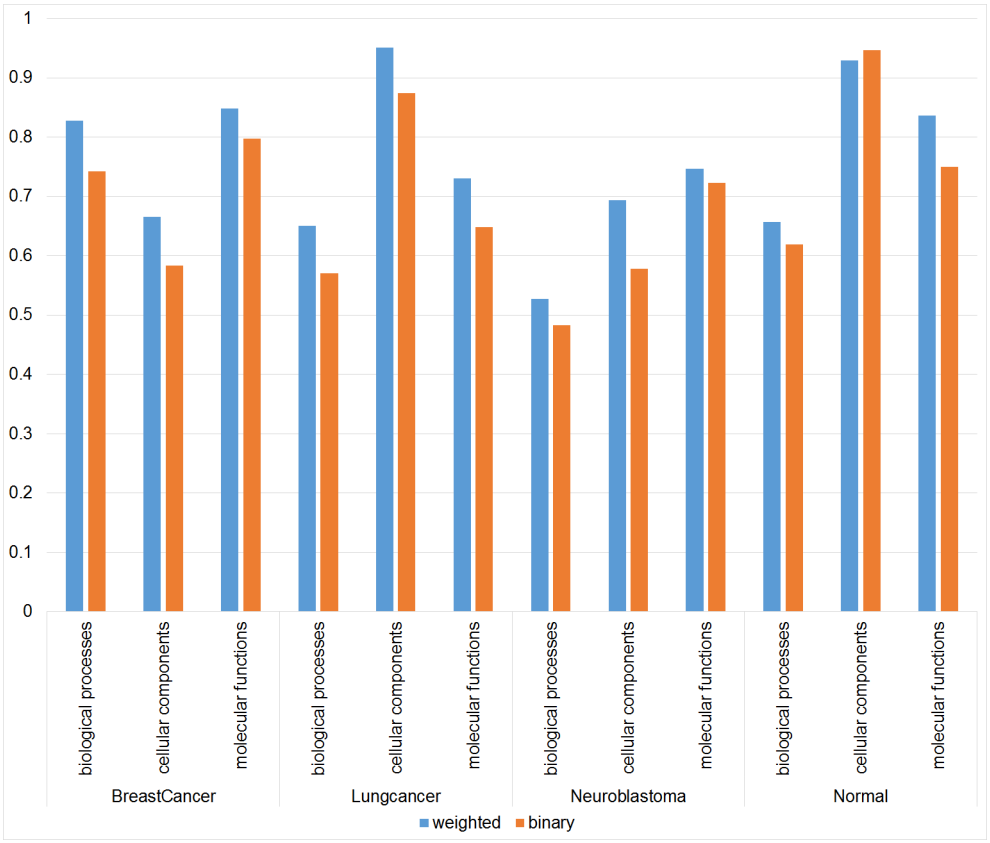

Figure 6: Domain base score of motif profiles for three types of cancer and an instance of normal gene co-expression networks .

cal and biological significance of a function. Hence, we use p-values to assess the chance of observing a particular GO term [1, 3]. If the set of genes participating in motif sg is of size n and m genes have a particular biological annotation then the probability of observing m or more genes, annotated with the same GO term out of n genes is given by:

$$p\text{-}value = \sum_{i=m}^{n} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}} \quad (4)$$

where N is the number of genes in the database and M is the number of genes that have the same annotation. In other words we are testing the hypothesis of a motif being associated to a particular biological annotation or not. Smaller p-values show that the association is not random and is biologically more significant than one with a higher p-value. We distinguish biologically significant motifs from non-significant ones using a cutoff value, then we compare different motif profiles (binary and weighted) regarding the scoring function:

$$\text{Motif profile score} = 1 - \frac{\sum_{i=1}^{n_S} \min(p_i) + n_I * cutoff}{(n_S + n_I) * cutoff} \quad (5)$$

where n_S and n_I are respectively the number of significant and insignificant motifs and $\min(p_i)$ denotes the smallest p-value of the significant motif i. A motif with a p-value less than a cutoff is significant. We used the recommended cut-off of 0.05 for all our validations.

The motif profiles (binary/weighted) are compared using the score function across three ontologies vocabularies namely biological, cellular and molecular. Fig. 6 shows the comparison between weighted and binary profiles of three cancer types and normal networks. We can see that the weighted profile of a network has higher biological score i.e. the number of motifs discovered by our weighted method are also biologically significant.

5. CONCLUSIONS

We proposed a novel method for motif mining in edge weighted networks. The weighted method assesses the quality and strength of the connection between objects. The subgraphs for which a relation between nodes is differently weighted than the whole network are considered as motifs. We use statistical testing to compare the weight distribution of edges in the whole network to edges inside a particular subgraph. This definition is well suited for applications such as gene co-expression networks where the goal is to find groups of genes differentially expressed. In the end we are able to construct a characteristic weighted fingerprint of a network.

We applied our method on several healthy and cancer related datasets to compare the gene networks in terms of their structural patterns, showing that our fingerprint is capable of distinguishing different types of networks. We also showed that the discovered weighted motifs are more biologically relevant when compared to the discovered traditional binary motifs.

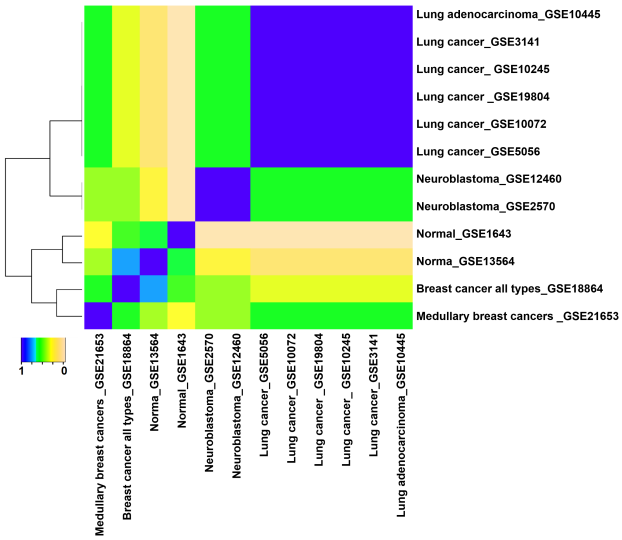

Figure 4: Similarity matrix of gene co-expression networks for datasets with 3 types of cancers and 2 healthy cases. The similarity is calculated by Euclidean distance of networks based on their weighted motif profiles.

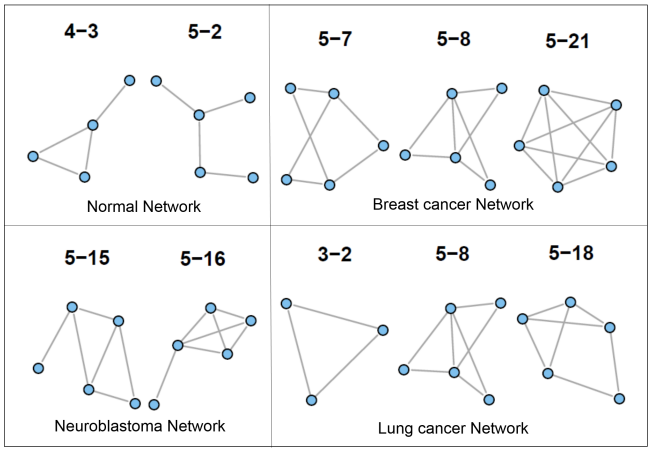

Figure 5: Discriminating subgraphs for each type of networks.

In the near future we would like to apply our technique to more scenarios where edge weight is important and where a weighted network is a suitable representation. We also intend to use a larger set of subgraphs as motif candidates to understand if we are able to gain even more information.

Acknowledgments

Sarvenaz Choobdar is funded by an FCT Research Grant (SFRH/BD/72697/2010). Pedro Ribeiro is funded by an FCT Research Grant (SFRH/BPD/81695/2011). This work is partially funded by projects (SIBILA NORTE-07-0124-FEDER-000059 and PEStFCOMP-01-0124-FEDER-037281) under ON.2, NSRF/ERDF, COMPETE and national funds, through FCT.

References

[1] V. Arnau, S. Mars, and I. Marín. Iterative cluster analysis of protein interaction data. *Bioinformatics*, 21(3):364–378, 2005.

[2] M. I. Arnone and E. H. Davidson. The hardwiring of development: organization and function of genomic regulatory systems. *Development*, 124(10):1851–1864, 1997.

[3] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[4] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[5] M. R. Carlson, B. Zhang, Z. Fang, P. S. Mischel, S. Horvath, and S. F. Nelson. Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. *BMC genomics*, 7(1):40, 2006.

[6] S. Choobdar, P. Ribeiro, S. Bulga, and F. Silva. Co-authorship network comparison across research fields using motifs. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2012.

[7] S. Choobdar, P. Ribeiro, and F. Silva. Motif mining in weighted networks. In *Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on*, pages 210–217. IEEE, 2012.

[8] J. Dong and S. Horvath. Understanding network concepts in modules. *BMC Systems Biology*, 1(1):24, 2007.

[9] P. S. Gargalovic, M. Imura, B. Zhang, N. M. Gharavi, M. J. Clark, J. Pagnon, W.-P. Yang, A. He, A. Truong, S. Patel, et al. Identification of inflammatory gene modules based on variations of human endothelial cell responses to oxidized lipids. *Proceedings of the National Academy of Sciences*, 103(34):12741–12746, 2006.

[10] J. Grochow and M. Kellis. Network motif discovery using subgraph enumeration and symmetry-breaking. In *Research in Computational Molecular Biology*, pages 92–106. Springer, 2007.

[11] C. Helma, S. Kramer, and L. De Raedt. The molecular feature miner molfea. In *Proceedings of the Beilstein-Institut Workshop*. May, 2002.

[12] S. Horvath and J. Dong. Geometric Interpretation of Gene Coexpression Network Analysis. *PLoS Comput Biol*, 4(8):e1000117+, Aug. 2008.

[13] S. Horvath, B. Zhang, M. Carlson, K. Lu, S. Zhu, R. Felciano, M. Laurance, W. Zhao, S. Qi, Z. Chen, et al. Analysis of oncogenic signaling networks in glioblastoma identifies aspm as a molecular target. *Proceedings of the National Academy of Sciences*, 103(46):17402–17407, 2006.

[14] H. Hu, X. Yan, Y. Huang, J. Han, and X. J. Zhou. Mining coherent dense subgraphs across massive biological networks for functional discovery. *Bioinformatics*, 21(suppl 1):i213–i221, 2005.

[15] J. Huan, W. Wang, and J. Prins. Efficient mining of frequent subgraphs in the presence of isomorphism. In *Proceedings of the Third IEEE International Conference on Data Mining*, ICDM '03, pages 549–, 2003.

[16] C. Jiang, F. Coenen, and M. Zito. Frequent sub-graph mining on edge weighted graphs. *Data Warehousing and Knowledge Discovery*, pages 77–88, 2010.

[17] S. Kullback. *Information theory and statistics*. Courier Dover Publications, 1968.

[18] H. Li, Y. Sun, and M. Zhan. Exploring pathways from gene co-expression to network dynamics. In *Computational Systems Biology*, pages 249–267. Springer, 2009.

[19] N. K. MacLennan, J. Dong, J. E. Aten, S. Horvath, L. Rahib, L. Ornelas, K. M. Dipple, and E. R. McCabe. Weighted gene co-expression network analysis identifies biomarkers in glycerol kinase deficient mice. *Molecular genetics and metabolism*, 98(1):203–214, 2009.

[20] F. J. Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.

[21] G. L. G. Miklos and G. M. Rubin. The role of the genome project review in determining gene function: Insights from model organisms. *Cell*, 86:521–9, 1996.

[22] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon. Superfamilies of evolved and designed networks. *Science*, 303(5663):1538–1542, 2004.

[23] 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.

[24] T. Nepusz, H. Yu, and A. Paccanaro. Detecting overlapping protein complexes in protein-protein interaction networks. *Nature methods*, 9(5):471–472, 2012.

[25] M. C. Oldham, S. Horvath, and D. H. Geschwind. Conservation and evolution of gene coexpression networks in human and chimpanzee brains. *Proceedings of the National Academy of Sciences*, 103(47):17973–17978, 2006.

[26] M. A. Pujana, J.-D. J. Han, L. M. Starita, K. N. Stevens, M. Tewari, J. S. Ahn, G. Rennert, V. Moreno, T. Kirchhoff, B. Gold, et al. Network modeling links breast cancer susceptibility and centrosome dysfunction. *Nature genetics*, 39(11):1338–1349, 2007.

[27] P. Ribeiro and F. Silva. G-tries: an efficient data structure for discovering network motifs. In *Proceedings of the 2010 ACM Symposium on Applied Computing*, pages 1559–1566, 2010.

[28] P. Ribeiro and F. Silva. G-tries: a data structure for storing and finding subgraphs. *Data Mining and Knowledge Discovery*, 2013.

[29] J. Saramaki, J.-P. Onnela, J. Kertesz, and K. Kaski. Characterizing motifs in weighted complex networks. *AIP Conference Proceedings*, 776(1):108–117, 2005.

[30] J. M. Stuart, E. Segal, D. Koller, and S. K. Kim. A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302:249–255, 2003.

[31] S. Wernicke. Efficient detection of network motifs. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 3(4):347–359, 2006.

[32] X. Yan and J. Han. gspan: Graph-based substructure pattern mining. In *Proceedings of the 2002 IEEE International Conference on Data Mining*, ICDM '02, pages 721–, 2002.

[33] B. Zhang and S. Horvath. A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, 4(1):1128, 2005.

[34] J. Zhang, K. Huang, Y. Xiang, and R. Jin. Using frequent co-expression network to identify gene clusters for breast cancer prognosis. In *Bioinformatics, Systems Biology and Intelligent Computing, 2009. IJCBS'09. International Joint Conference on*, pages 428–434. IEEE, 2009.

[35] J. Zhang, K. Lu, Y. Xiang, M. Islam, S. Kotian, Z. Kais, C. Lee, M. Arora, H.-w. Liu, J. D. Parvin, et al. Weighted frequent gene co-expression network mining to identify genes involved in genome stability. *PLoS Computational Biology*, 8(8):e1002656, 2012.

[36] W. Zhao, P. Langfelder, T. Fuller, J. Dong, A. Li, and S. Hovarth. Weighted gene coexpression network analysis: state of the art. *Journal of biopharmaceutical statistics*, 20(2):281–300, 2010.