A Comparison of Community Identication Algorithms for Regulatory Network Motifs

Douglas Oliveira and Marco Carvalho

Abstract— **In the recent years high throughput data about biological processes has become available and thus opened a wide range of possibilities of research in multi-disciplinary areas, like network science. An idea that has been widely accepted is the fact that no life can exist without complex systems formed by interacting macromolecules. Rather than a single gene being responsible for a single phenotype (central dogma), it has been shown that the interaction between several genes is responsible for a given phenotype, a concept called System Biology. Identifying patterns of interactions (motifs) in these complex networks has attracted the attention in the scientific community, given that these networks are often very dense and dynamic. In this work we focus on a particular kind of biological network, a regulatory network where each node is a transcription factor and two nodes are connected if one of them encodes a transcription factor to another one that is regulated by this transcription factor. We focus on a specific kind of motif, a dense overlapping region (DOR) that claims that a set of genes regulated by different transcription factors are more overlapping than expected at a random network. We use different community identification algorithms in order to identify which algorithm best suits to the task of identification of this particular motif.**

I. Introduction

According to [1] most of the interesting accomplishments achieved in biological research has been in genomics. One example is the genome sequencing of many species, including the human genome, which has created many possibilities for a better understanding of the function of many genes from large-scale sequencing processes. We currently have a good understanding of life at the molecular level, and recognize that we need to see gene structures not only in isolation but also as sets, and how they interact with one another [2].

By accepting the concept of system biology, we are not denying the importance of reductionist approaches. Reductionist approaches are just limited concerning the function of presenting a comprehensive picture of life [1]. One fact that supports the idea of system biology is that individual cells when separated from their neighbors lose many of their functional and structural attributes [3].

The notion of systems biology dates back from hundreds of years ago when the word organism was initially used to describe living animals and plants as organizations, where each part is reciprocally end and means. Many advantages have rise with this new approach like, for example, evolutionary mechanisms can be better understood in light of complex molecular systems [4].

D. Oliveira and M. Carvalho are with Florida Institute of Technology, 150 W. University Blvd, Melbourne, FL, USA doliveira2011@my.fit.edu, mcarvalho@cs.fit.edu

With the current availability of terabytes of data in many domains, including biological processes, communications, and social interactions, a variety of research actives have started to focus on modeling and identification of global network properties and characteristics. These include the small world property [5] and scale-free networks [6]. One of the first networks structures analyzed with this approach was the network representing scientific collaborations and co-publications [7]. While important, such global metrics must be augmented with the understanding of basic structural elements, the building blocks of the network. These building blocks are often referred to as network motifs [8] and represent recurring structures and patterns of connections.

In [8] the authors present several different kinds of motifs normally found in different types of networks. In their work, the authors justify the presence of the motifs to the way in which the network was designed. More specifically in biological networks the work of [9] identifies three major patterns that are significantly present in the network. Among them, a motif called dense overlapping regulons (DOR), requires special attention. The motif is defined as a layer of overlapping interactions that is much more dense than the corresponding structures in randomized networks. The result is a structure characterized by loosely connected and internally dense regions of interactions. These regions are often called communities.

There are many community identification algorithms in literature. In general, such algorithms rely on the partition of the data into a certain number of communities (groups, subsets or categories) [10]. There is no clear definition of a community, but most authors characterize a community by its internal homogeneity and the external separation [11]. In this work we evaluate the results of four community identification algorithms aiming to identify which bets suits for the identification of DOR motifs in a regulatory network.

II. Related Work

Gene expression data is obtained through microarray experiments [16] and is commonly used for study of biological networks. Community identification algorithms have been widely applied in these kinds of datasets, for example for the construction of coexpression networks [12]. In a coexpression network each node represents a gene, and two nodes are connected if their expression levels are similar [13].

The work of [14] shows results of clustering 118 genes using a hierarchical community identification algorithm in which members of the same clusters tend to participate in common processes. In a later work [15], the authors

978-1-4799-3163-7/13/$31.00 ©2013 IEEE

apply a systematic set of statistical algorithms and clustering algorithms to identify transcriptional regulatory sub-networks in yeast without any a priori knowledge of its structure. Their results uncovered new regulons and their putative cis-regulatory elements.

Changes in the modules in a coexpression network have also been associated with the possibility of cancer. In the work of [17] and [18] cluster algorithms have identified changes in expression patterns in individuals with and without cancer.

Coexpression networks allow the use of mining cluster algorithms like Self-Organizing Maps [19] and K-Means[20], once it is possible to calculate the distance between any two nodes based on the expression values. These algorithms are not valid in regulatory networks. Examples of the use of these algorithms in coexpression networks can be seen in [21] and [22].

The work of [10] presents a theoretical survey of clustering algorithms accordingly to their applicability. In that paper, the authors conclude that there is no clustering algorithm that can be universally used to solve all problems, as new technologies have generated greater volumes of more complex data, requiring more powerful clustering algorithms. A second conclusions highlighted by the authors is that the preprocessing and post-processing of the data are as important as the own clustering algorithm.

III. The Regulatory Network

In order to evaluate community identification algorithms in a regulatory network we choose a specie that is considered one of the best characterized, the bacteria *Escherichia coli K-12*. The data is available in the RegulonDB, a relational database which provides information on transcriptional regulation that has been manually curated from original scientific publications [23]. Each node is a transcription factor and two nodes are connected if one of them encodes a transcription factor to another one that is regulated by this transcription factor. The network can be seen in Figure 1.

In the Figure 1 we can see that the network is not entirely connected. From the 164 nodes, 26 have only one connection to itself (autoregulation), 10 of them are connected to a single other node and we have one triad which nodes have none other connection beyond the triad. In the end, 125 nodes form the giant component of the network. The nodes are colored accordingly to its degree. As illustrated in the figure, most nodes have only a few edges while only a few nodes (*crp, h-ns* and *fnr*) have more than a few edges. This behavior of the network lead us to investigate if the degree distribution followed a power law distribution [6]. If the degree distribution of the network follows a power law distribution, the network is called Scale Free. Graphically scale free networks present a long tail, where the probability of a node have a degree k is inversely proportional to k, formally:

$$P(k) \approx k^{-\alpha} \tag{1}$$

Where the exponent α is a positive constant and normally assume values between two and four [6]. The long tail

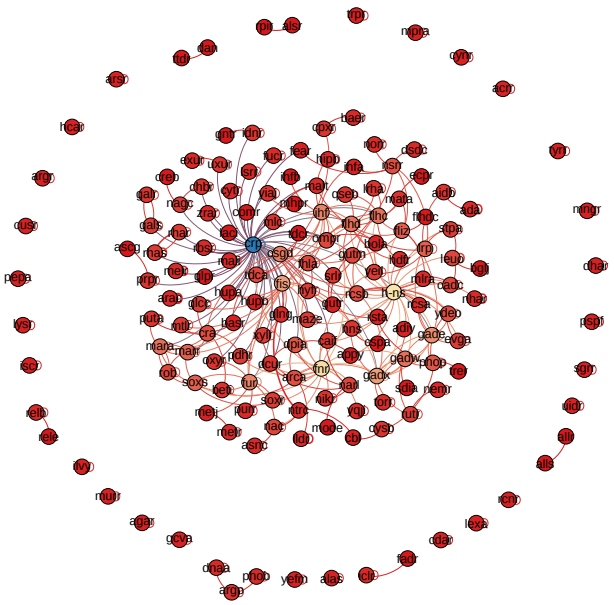

Fig. 1. Regulatory Network of Escherichia coli K-12.

behavior of the data can be seen in Figure 2 and when plotted in a log-log chart (inset) it approaches to a linear regression. The linear regression of the data is considered high assuming value of 0.91 and the inclination of the line (exponent) is -2.87. Thus we can affirm that a very small set of nodes is responsible for trigger the production the majority of the transcriptional factors in the network, once they are connected to a large percentage of the network.

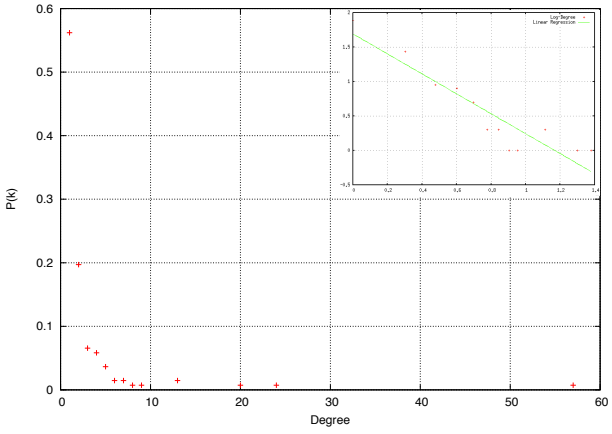

Fig. 2. Degree distribution of the regulatory network of E. coli. The inset shows the distribution in a log-log plot.

In order to have a general understanding of the network we also evaluated if the network can be classified as a small world network. In a small world network the average path length of the nodes is short, in fact it should be O(log n) where n is the number of nodes of the network. Not only the average path length needs to be short but also the network needs to be clustered [5]. The measure of

TABLE I

RESULTS OF THE CLUSTERING COEFFICIENT (C) COMPARED TO THE CLUSTERING COEFFICIENT OF A RANDOM NETWORK OF THE SAME SIZE (C_{rand}) AND THE AVERAGE PATH LENGTH (l) COMPARED TO THE LOG (n), WHERE n IS THE NUMBER OF NODES OF THE NETWORK.

Network	C	C_{rand}	l	log (n)
Directed	0.1	0.048	2.38	2.21
Undirected	0.31	0.094	3.1	2.21

clustering of a network is done by calculating the fraction of possible triangles in the neighborhood of each node [24]. To evaluate if the network is clustered we need to compare its clustering coefficient to the clustering coefficient of an equivalent randomized network, in other words, a network with the same number of nodes and edges, but with an edges choice randomized. As shown in Table I the clustering coefficient in the real network is larger than the clustering coefficient in the randomized network in both situations, if we consider the network directed or undirected. We can also see that the average path length is very close from log (n), where n is the number of nodes of the network. Thus we can conclude that the network has small world characteristics.

IV. COMMUNITY IDENTIFICATION ANALYSIS

In this paper we use four well known community detection algorithms for the identification of motifs in the regulatory network of E. coli. The goal is to identify which algorithm is better suited for the task. The first algorithm was proposed by Girvan and Newman [25] and it is based on consecutive removal of edges in the network until the communities are formed. The second algorithm is based on the optimization of clustering keeping in mind that in a community the number of edges intracommunity is higher than the number of edges intercomunities [26]. The third algorithm was design with the idea that if a set of nodes in a network forms a community then would be possible for a click (fully connected graph) of size k to 'walk' inside of the community [27]. The last algorithm is inspired in swarm intelligence, where each node locally chooses its community trying to minimize its entropy [28]. In this section we further describe and apply each of the proposed algorithms to the E. coli dataset.

A. Betweenness Centrality Method

The betweenness centrality method was proposed to mitigate some of the shortcomings in hierarchical clustering algorithms. Differently from methods that try to build a metric to identify which nodes are close to each other, this method focuses on the edges that are more 'central' in the network. The choice of the central edges is based on the betweenness centrality of an edge. The concept of betweenness centrality was first proposed by Freeman [29]. In that work, Freemen defines the betweenness centrality of a node i as the number of shortest paths between any pairs of nodes in the network that pass through i, thus measuring the influence of the node in the network. In the work of [25] this concept was extend to edges, having in mind that the edges with highest betweenness centrality are those who tend to exist inter-comunities. By progressively eliminating edges with high betweenness centrality, the network will eventually acquire a community structure.

Before using this algorithm, we need to eliminate those nodes in the network that do not belong to the giant component of the network, once these nodes already form a isolated community (even if it has only one or two nodes). As we can see in Figure 3 this algorithm did not yield the expected results for this particular network, create large communities of nodes. In this network most paths are short, having a maximum length of three or four. Another characteristic of this network is that the hubs of the network are hubs because of outgoing edges, thus most of shortest path do not pass through that node or its edges, making the edges located in periphery strong candidates for removal due to their high centrality measure.

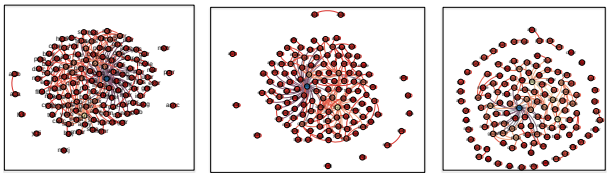

Fig. 3. Results of edges removal for, respectively, 10%, 20% and 40% of the edges in the network.

This algorithm is very dependent of the structure of the network to provide satisfactory results. Networks that have a more 'natural' community structure are more suited for this particular algorithm, which is not the case of the complex structure of a regulatory network.

B. Modularity Optimization Method

The modularity optimization method is a heuristic one; it gives an approximate solution to the optimization problem for community identification. The method, proposed by Blonde et al in [26], divides the network in communities by evaluating a metric of the quality of the community called modularity. The modularity of a community is a scalar value between -1 and 1, that measures the density of links intra-communities in comparison to links inter-communities.

The algorithm is composed of two main phases, in the first one each node is assigned to a community and interactively it will change its community if this change will increase the modularity. In case of having more than one option that will increase the modularity, the node will move to the community with the highest modularity gain. The second phase consists of treating each community formed in the first phase as a single node, edges inside the community are treated as self loops. In this phase the merge of entire communities is considered instead of only nodes. Both phases are successively executed until no gain in modularity is achieved.

Accordingly to the authors the steps of the algorithm are intuitive and easy to implement, another advantage is that

the outcome is unsupervised. We used the implementation of the tool Gephi [30]. As we can see in Figure 4 the method creates five communities, represented by the different colors in the network.

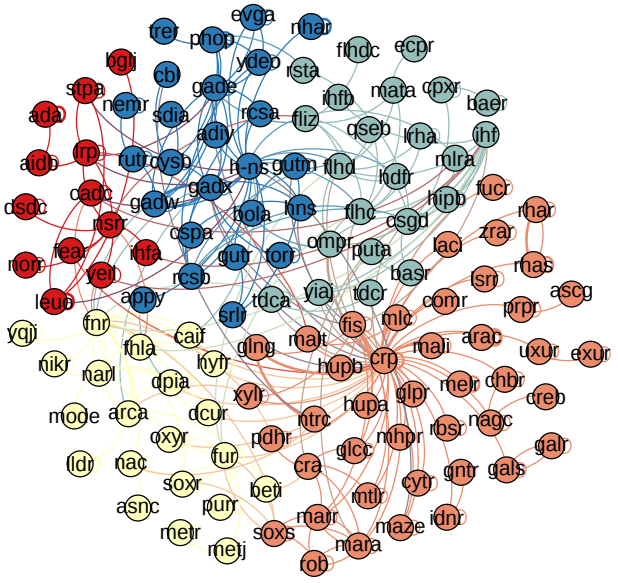

Fig. 4. Results of community identification using modularity optimization.

We can notice that only the red and cyan communities do not have a hub that connect most of its members. The remaining communities have a central hub, like the *h-ns* node in the blue community and the *crp* node in the orange one. What is interesting in this result is the amount of edges inter-communities, for example we can notice a large amount of edges connecting the cyan and orange community, what can indicate the existence of a dense overlapping motif.

C. Clique Percolation Method

All the methods presented so far were exclusive ones, in other words, a node can only participate of a single community. In this method each node has a membership number that is the number of communities that the node belongs to. The basic principle of the algorithm is that a typical community consist of several fully connected sub-graphs that tend to share many of their nodes, this kind of community is called a *k-clique-community* [27]. Formally a *k-clique-community* is the set of nodes that can be reached from each other through a series of adjacent k-cliques. It is possible to relaxing this definition by allowing incomplete k-cliques but this is equivalent to lowering the values of *k*.

One advantage of this method is that it offers the possibility of looking at the network from a higher level of organization and locate the communities that play a key role within the web of communities (those with many overlapping nodes). In this work we used the implementation provided by the tool CFinder [31].

The Figure 5 presents the results of this algorithm. In the left side of the figure, the network is divided in nine

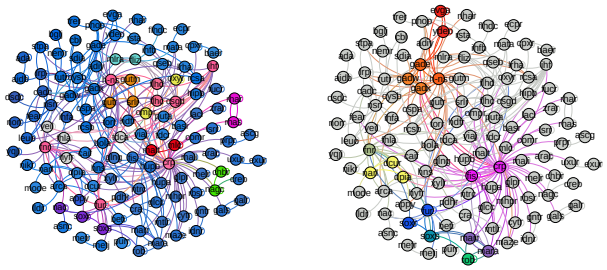

Fig. 5. Results of community identification using clique percolation. The network in left has the size of the clique equals to three. The network in the righ has size of the clique equals to four.

different communities that share a clique of size three. In this scenario we notice that the majority of nodes belongs to a single community while the remaining communities have just a couple nodes. An interesting fact in this community structure is that the hubs of the network are those that make the connections between the communities, for example, the nodes *crp* and *fnr*.

In the right side of the figure we have the results using a clique of size four. The nodes colored in gray were not assigned to any community, only the others. We notice that only five communities were identified with a minimum number of nodes in each (four). This result indicates the difficulty of having a clique of size four inside of regulatory networks.

D. SOCIAL

In the algorithm proposed by [28] each node is responsible for determining the community to which it belongs. This decision is based on the concept of Shannon entropy [32] and is made locally, which means that a node only needs information about its immediate neighbors. The algorithm was inspired in swarm intelligent systems, where each individual in the system interact only with its immediate neighborhood, thus obtaining high efficiency and robustness [33]. The concept of a node's entropy encapsulates the certainty of each node regard to its current community. At each iteration if a node have neighbors in different communities it calculates the entropy of joining the communities of its neighbors and has a higher probability of joining that gives a lower entropy. This step is repeated until the entropy of the entire network stabilizes.

Accordingly to its creators this algorithm is well suited for dynamic networks where nodes join and leave constantly. The decentralized design of the algorithm does not require extensive recalculation of the communities in case of a new node joining the network or a node leaving the network, all the changes are made locally.

The results of the algorithm can be seen in Figure 6. It was identified 80 different groups, which 45 nodes form isolated groups (groups with one node only). This happens due to the fact that some nodes does not have outgoing edges only incoming ones, thus they do not have any neighbor to consider joint its community. The remaining groups only

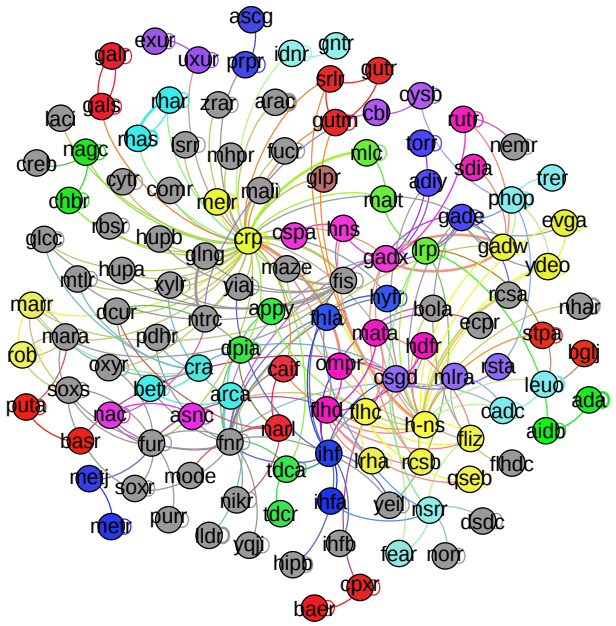

Fig. 6. Results of community identification using the algorithm SOCIAL.

have two or three nodes. We believe that large communities have not appeared in this network because this network has not an usual community structure, like for example social networks.

V. Conclusions and Future Work

In this work we analyzed the regulatory network of *Escherichia coli K-12*. The network presented characteristics of a scale free network, in which the degree distribution follows a power law. It also has small world properites, displaying high clustering coefficient and short average path length, which is common in signal processing networks [9].

We have also explored different community identification algorithms for motif identification. First we applied an algorithm based on edges betweenness centrality. The algorithm presented poor results, likely because the edges with highest centrality were mostly located in the periphery of the network. A second method based on modularity optimization was then applied. The second method presented good results, especially because we wanted to identify the communities that have a large number of edges connecting them, an indicator of a DOR motif. We also perform a community identification algorithm that uses the concept of clique percolation and due to the sparse structure of the network the balance between the communities was very low, leading to few communities with a large number of nodes and a large number communities with few nodes. As we increase the size of cliques, the results only get worse, because it is more rare to find cliques of a bigger size. Finally we used an swarm based algorithm for community identification, it has identified a large amount of groups with a couple of nodes inside. This behavior is primarily due to the fact that the directional edges of the network did not allowed the nodes consider some neighbors that have incoming edges to it.

Although some results of the algorithms used in this work are not useful in our initial task of identifying a DOR motif, one of its contributions relates to the fact that it explores the behavior of well known community identification algorithms in networks that are quite different from social networks; which were the networks in which those algorithm were designed and first tested.

As part of our future work, we intend to further combine coexpression information with regulatory information aiming to have a more realistic network and then compare how the clustering results of this hybrid network would differ from those obtained in this work. We also intend to investigate the behavior of these clustering algorithms in other species, preferable those who have well known interactions, like *Saccharomyces cerevisiae*, the yeast. We also want to analyze the behavior of such algorithms in a undirected regulatory network, which we presume that will generate quite different results, and thus it can give us a different perspective of the regulatory system of the studied specie.

References

[1] Marc Vidal, "A biological atlas of functional maps review," *Cell*, vol. 104, pp. 333–339, 2001.

[2] Leland H Hartwell, John J Hopfield, Stanislas Leibler, and Andrew W Murray, "From molecular to modular cell biology," *Nature*, vol. 402, pp. C47–C52, 1999.

[3] Marc Vidal, "A unifying view of 21st century systems biology," *FEBS letters*, vol. 583, no. 24, pp. 3891–3894, 2009.

[4] Dawn Anne Thompson and Aviv Regev, "Fungal regulatory evolution: cis and trans in the balance," *FEBS letters*, vol. 583, no. 24, pp. 3959, 2009.

[5] D. Watts and S. Strogatz, "The small world problem," *Collective Dynamics of Small-World Networks*, vol. 393, pp. 440–442, 1998.

[6] R. Albert and A.L. Barabási, "Statistical mechanics of complex networks," *Reviews of modern physics*, vol. 74, no. 1, pp. 47, 2002.

[7] Mark EJ Newman, "The structure of scientific collaboration networks," *Proceedings of the National Academy of Sciences*, vol. 98, no. 2, pp. 404–409, 2001.

[8] Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon, "Network motifs: simple building blocks of complex networks," *Science Signaling*, vol. 298, no. 5594, pp. 824, 2002.

[9] Shai S Shen-Orr, Ron Milo, Shmoolik Mangan, and Uri Alon, "Network motifs in the transcriptional regulation network of escherichia coli," *Nature genetics*, vol. 31, no. 1, pp. 64–68, 2002.

[10] Rui Xu, Donald Wunsch, et al., "Survey of clustering algorithms," *Neural Networks, IEEE Transactions on*, vol. 16, no. 3, pp. 645–678, 2005.

[11] Pierre Hansen and Brigitte Jaumard, "Cluster analysis and mathematical programming," *Mathematical programming*, vol. 79, no. 1-3, pp. 191–215, 1997.

[12] Michael B Eisen, Paul T Spellman, Patrick O Brown, and David Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proceedings of the National Academy of Sciences*, vol. 95, no. 25, pp. 14863–14868, 1998.

[13] Joshua M Stuart, Eran Segal, Daphne Koller, and Stuart K Kim, "A gene-coexpression network for global discovery of conserved genetic modules," *Science*, vol. 302, no. 5643, pp. 249–255, 2003.

[14] Saeed Tavazoie, George M Church, et al., "Quantitative whole-genome analysis of dna-protein interactions by in vivo methylase protection in e. coli," *Nature biotechnology*, vol. 16, no. 6, pp. 566–571, 1998.

[15] Saeed Tavazoie, Jason D Hughes, Michael J Campbell, Raymond J Cho, and George M Church, "Systematic determination of genetic network architecture," *Nature genetics*, vol. 22, no. 3, pp. 281–285, 1999.

[16] Michael B Eisen, Patrick O Brown, et al., "Dna arrays for analysis of gene expression," *Methods in enzymology*, pp. 179–204, 1999.

[17] Todd R Golub, Donna K Slonim, Pablo Tamayo, Christine Huard, Michelle Gaasenbeek, Jill P Mesirov, Hilary Coller, Mignon L Loh, James R Downing, Mark A Caligiuri, et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *science*, vol. 286, no. 5439, pp. 531–537, 1999.

[18] Avrum Spira, Jennifer E Beane, Vishal Shah, Katrina Steiling, Gang Liu, Frank Schembri, Sean Gilman, Yves-Martine Dumas, Paul Calner, Paola Sebastiani, et al., "Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer," *Nature medicine*, vol. 13, no. 3, pp. 361–366, 2007.

[19] Teuvo Kohonen, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.

[20] Anil K Jain and Richard C Dubes, *Algorithms for clustering data*, Prentice-Hall, Inc., 1988.

[21] Pablo Tamayo, Donna Slonim, Jill Mesirov, Qing Zhu, Sutisak Kitareewan, Ethan Dmitrovsky, Eric S Lander, and Todd R Golub, "Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation," *Proceedings of the National Academy of Sciences*, vol. 96, no. 6, pp. 2907–2912, 1999.

[22] Ralf Herwig, Albert J Poustka, Christine Müller, Christof Bull, Hans Lehrach, and John O'Brien, "Large-scale clustering of cdna-fingerprinting data," *Genome research*, vol. 9, no. 11, pp. 1093–1105, 1999.

[23] Heladia Salgado, Martin Peralta-Gil, Socorro Gama-Castro, Alberto Santos-Zavaleta, Luis Muñiz-Rascado, Jair S García-Sotelo, Verena Weiss, Hilda Solano-Lira, Irma Martínez-Flores, Alejandra Medina-Rivera, et al., "Regulondb v8. 0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more," *Nucleic acids research*, vol. 41, no. D1, pp. D203–D213, 2013.

[24] Charalampos E Tsourakakis, "Fast counting of triangles in large real networks without counting: Algorithms and laws," in *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*. IEEE, 2008, pp. 608–617.

[25] Michelle Girvan and Mark EJ Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences*, vol. 99, no. 12, pp. 7821–7826, 2002.

[26] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, pp. P10008, 2008.

[27] Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, no. 7043, pp. 814–818, 2005.

[28] Ben Collingsworth and Ronaldo Menezes, "A self-organized approach for detecting communities in networks," in *Intelligent Distributed Computing VI*, pp. 29–39. Springer, 2013.

[29] Stanley Wasserman and Katherine Faust, *Social network analysis: Methods and applications*, vol. 8, Cambridge university press, 1994.

[30] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy, "Gephi: An open source software for exploring and manipulating networks," in *International AAAI conference on weblogs and social media*. AAAI Press Menlo Park, CA, 2009, vol. 2.

[31] Balázs Adamcsek, Gergely Palla, Illés J Farkas, Imre Derényi, and Tamás Vicsek, "Cfinder: locating cliques and overlapping modules in biological networks," *Bioinformatics*, vol. 22, no. 8, pp. 1021–1023, 2006.

[32] Claude Elwood Shannon and Warren Weaver, "A mathematical theory of communication," 1948.

[33] Marco Mamei, Ronaldo Menezes, Robert Tolksdorf, and Franco Zambonelli, "Case studies for self-organization in computer science," *Journal of Systems Architecture*, vol. 52, no. 8, pp. 443–460, 2006.