Computational analysis and comparison of gene networks from TCGA normal and cancer data

Gaia Ceddia*,[1], Sara Pidò[1], Marco Masseroli[1]

(1) Politecnico di Milano - Dipartimento di Elettronica, Informazione e Bioingegneria,
Piazza Leonardo da Vinci 32, Milan, Italy, first.last@polimi.it
* corresponding author

Keywords: Gene networks, microRNA, gene expression profiles, complex networks.

Abstract. Network modeling is an important approach to understand cell behaviour. It has proven its effectiveness in understanding biological processes and finding novel biomarkers for severe diseases. In this study, using gene expression data and complex network techniques, we propose a computational framework for inferring relationships between RNA molecules. We focus on gene expression data of kidney renal clear cell carcinoma (KIRC) from the TCGA project, and we build RNA relationship networks for either normal or cancer condition using three different similarity measures (Pearson's correlation, Euclidean distance and inverse Covariance matrix). We analyze the networks individually and in comparison to each other, highlighting their differences. The analysis identified known cancer genes/miRNAs and other RNAs with interesting features in the networks, which may play an important role in kidney renal clear cell carcinoma.

1 Scientific Background

Network biology covers a wide range of scales, from molecular interactions in the cell to intercellular communications and connections between organisms. At the cell level, high-throughput next-generation sequencing technology is generating an enormous amount of genomic data from which qualitative and quantitative relationships between RNA molecules can be inferred [1]. In particular, gene expression data provide information about the synthesis of functional gene products, either proteins or not; using mathematical and statistical techniques, from gene expression data we can generate biological networks, where genes are the network nodes and interactions between gene products are the edges in the network graph. This process, named network inference or reverse engineering, has given important insights on complex biological processes and disease mechanisms within the cell [2]. Network inference has the advantage of being efficient and inexpensive compared to experimental lab validation; thus, complex network techniques and algorithms have been increasingly deployed to understand inferred biological networks.

A complex network is a graph with non-trivial topological features [3], i.e., the patterns of connection between its elements are neither purely regular nor purely random. All biological processes can be modeled as networks, since they occur thanks to interactions among molecules. In biology, the most studied complex networks are gene networks, where typically genes encode for proteins; their interrelated activity determines protein abundance and related processes [3].

Most of the approaches used for inferring edges in gene networks are based on similarity (co-expression) measures. Co-expression measurement is based on the “guilt by association” definition, where genes with similar expression profiles are functionally associated due to their presumable co-regulation [2]. Thus, several different measures have been considered to assess co-expression, including Pearson's correlation and Euclidean distance. Pearson's correlation is the most common co-expression measure in the literature [2]. It has the benefit of being scalable, i.e., it can be efficiently computed for large numbers of genes, and it is not sensitive to linear transformations or different normalizations. Other methods for the construction of gene networks include Bayesian network approaches, as well as regression and differential equation based models. Bayesian networks are applied to represent conditional dependencies between genes given their expression levels, using a directed acyclic graph structure [1]. However, this procedure is applicable only to small networks, i.e., only a modest number of genes must be involved. Instead, regression and differential equation models are used for inferring gene regulatory networks, i.e., they assume that a particular subset of gene expression profiles is the most informing subset of all to predict expression profiles of target genes [1].

Here, we focus on co-expression networks built by computing Pearson's correlation, Euclidean distance and inverse Covariance metrics. The first measure is calculated to capture the scale-free similarity of gene expression profiles, the second one to take into account the scale of different gene expression profiles, and the third one as a multivariate analysis representing conditional independence between variables. Using expression data from the TGCA project [4], we build two different gene co-expression networks for normal or cancer cells, respectively; normal and cancer gene networks are computed for each similarity measure, and comparison analyses are performed among them.

For the considered datasets, we integrated messenger RNA (mRNA), microRNA (miRNA) and long non-coding RNA (lncRNA) expression profiles and we computed the co-expression networks among them; thus, our study is not limited to protein coding RNAs. MicroRNAs are small non-coding RNA molecules containing between 19 and 25 nucleotides, which work for RNA silencing and post-transcriptional regulation of gene expression [5]. The predominant function of miRNAs is to regulate protein translation by binding to complementary sequences in the 3' untranslated region (UTR) of target messenger RNAs (mRNAs), and thereby to negatively regulate mRNA translation [5]. A single miRNA can target hundreds of mRNAs, using base-pairing with complementary sequences within mRNA, and influence the expression of many genes often involved in a functional interaction pathway. However, miRNAs can also target lncRNAs, which are made of more than 200 nucleotides and are not translated into proteins. In this case, lncRNAs act as decoys for miRNAs silencing, allowing the translation of target mRNAs [6].

2 Materials and Methods

In this section, we explain our extraction and pre-processing pipeline for TCGA gene expression data and how we build pair networks for normal and cancer conditions, respectively, using three different similarity measures for each condition, resulting in a total of six networks.

2.1 Data extraction and pre-processing

We considered both RNA-Seq and miRNA-Seq public data for the human GRCh38 assembly from the TCGA repository. GRCh38 miRNA-Seq data contains miRNA quantification (i.e., the calculated expression for all reads aligning to a particular miRNA) and is derived from the sequencing of microRNAs, whereas GRCh38 RNA-Seq data contains gene expression quantification. For each miRNA-Seq and RNA-Seq dataset of each tumor type in TCGA, we computed the number of normal and cancer condition patients. For kidney renal clear cell carcinoma (KIRC) the ratio between the number of normal and cancer samples from patients resulted the lowest among all tumors in the TCGA repository, providing balanced datasets. KIRC RNA-Seq dataset resulted to have 72 and 527 patients for normal and cancer conditions, respectively, and KIRC miRNA-Seq dataset 71 and 545 patients for normal and cancer conditions, respectively. Thus, we used these KIRC data for our analysis.

Since RNA-Seq is designed for long gene sequencing, expression quantifications of short genes (i.e., shorter than 200 bp) can be considered as measure errors indeed. Thus, from the RNA-Seq dataset we removed them and selected only data of protein coding and long non-coding genes, which we integrated with the miRNA-Seq dataset ones keeping only those of patients common in both datasets.

We arranged these public gene expression data from the TCGA repository in the form of matrices, thus we constructed RNA-Seq and miRNA-Seq matrices in which rows represent genes/miRNAs and columns represent patients. Each element of the TCGA miRNA-Seq matrices is the expression level computed as reads per million miRNA mapped (RPM); conversely, the expression levels in the TCGA RNA-Seq matrices are provided as fragments per kilobase million (FPKM). To integrate the two miRNA-Seq and RNA-Seq datasets, we transformed miRNA expression data to be homogeneous with the RNA expression data; we converted RPM expression levels into FPKM ones by multiplying each element of the miRNA-Seq matrices by 100 and dividing it by the double of the length of the corresponding miRNA.

Furthermore, to separate biologically relevant genes from low-expression noisy ones, on the RNA-Seq data we applied the zFPKM normalization method [7]. For normal and tumoral cases separately, we computed the mean and the standard deviation of the log-transformed expression distribution of each gene across all KIRC samples and we normalized each logarithmic FPKM value of a gene by subtracting the gene computed mean and dividing the obtained value by the gene standard deviation (i.e., zFPKMs are Z-scores of log(FPKMs)). Then, we removed those genes with mean of their zFPKM distribution smaller than -3.0 in both normal and cancer conditions; this threshold separates expression levels of active genes from background genes as shown in [7].

After removing the background genes, we also deleted miRNAs with null expression in all normal and cancer samples from patients; thus, we obtained two matrices, one for normal and one for cancer data, each with 12,792 long RNAs (either coding or non-coding) and 1,397 miRNAs, and regarding 71 normal and 487 patients with cancer, respectively. These two matrices contain all the relevant FPKM values needed to build the desired networks.

2.2 Building the networks

To build the adjacency matrices of the gene networks, we considered three different similarity measures: Euclidean distance, Pearson's correlation and inverse Covariance. As mentioned in Section 1, we used these three different similarity measures to find scale-free, scale-dependent and multivariate similarities, respectively.

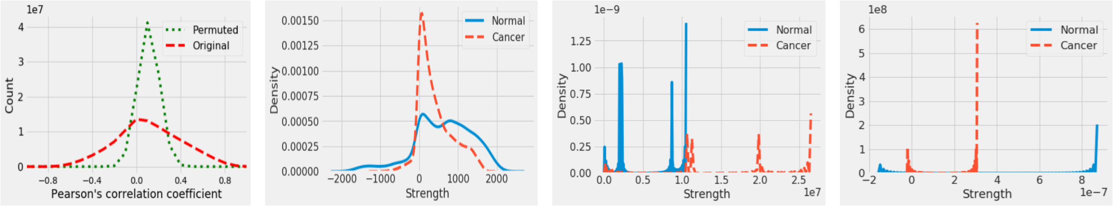
The Euclidean distance between two points is the length of the path connecting them. If $p = (p_1, p_2, ..., p_n)$ and $q = (q_1, q_2, ..., q_n)$ are two points in an Euclidean n-space, then their distance d is given by the Pythagorean formula [8]:

$$\mathbf{d} = \sqrt{\sum_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{q_i} - \mathbf{p_i})^{\mathbf{2}}} \tag{1}$$

We applied the Euclidean distance on each pair of genes/miRNAs in the datasets, considering the n patients in the datasets as the Euclidean n-dimensional space.

In statistics, the Pearson's correlation coefficient is a measure of the linear correlation between two variables X and Y (Eq. 2) [1]. Its values range between -1 and +1, where -1 indicates total negative linear correlation, 0 no linear correlation, and +1 total positive linear correlation. The Pearson's correlation between varable X and Y is defined as:

$$\rho_{\mathbf{X},\mathbf{Y}} = \frac{\mathrm{cov}(\mathbf{X}, \mathbf{Y})}{\sigma_{\mathbf{X}} \sigma_{\mathbf{Y}}} = \frac{\mathbf{E}[(\mathbf{X} - \mu_{\mathbf{X}})(\mathbf{Y} - \mu_{\mathbf{Y}})]}{\sigma_{\mathbf{X}} \sigma_{\mathbf{Y}}} \tag{2}$$



(a) (b) (c) (d)

Figure 1: (a) Red dashed line represents the distribution of Pearson's correlation coefficients for the gene/miRNA expression dataset in normal condition. Dotted green line represents the distribution of the average Pearson's correlation coefficients on 10 permuted repetitions of the gene/miRNA expression dataset in normal condition; (b)-(d) Strength distributions in normal and cancer networks are shown in blue full line and red dashed line, respectively, for the networks built with each of the similarity measures considered, i.e., Pearson's correlation (b), Euclidean distance (c) and inverse Covariance (d), respectively.

where $cov(X, Y)$ is the covariance of the two variables X and Y, i.e., the joint variability of X and Y, σ_X and σ_Y are the standard deviations of X and Y, respectively, and $cov(X, Y)$ can be expressed as the expected product of X and Y deviations from their individual expected values (i.e., their means μ_X and μ_Y, respectively). In our study we computed pairwise Pearson's correlation on each pair of genes/miRNAs in the datasets, and used the Pearson's coefficients to represent the weights of the edges in the networks.

The inverse Covariance matrix, commonly referred to as *precision matrix*, displays information about the partial correlations of variables [9]. In the Covariance matrix, the (i,j)-th element represents the unconditional correlation between a variable i and a variable j [9]. The inverse Covariance matrix instead represents conditional dependence, such that its (i,j)-th element is equal to zero if i and j are conditionally independent [9]. In other words, it gives the covariation of two variables while conditioning on the potential influence of the others involved in the analysis, i.e., it removes the effect of other variables. Thus, the precision matrix allows to obtain direct covariation between two variables by capturing partial correlations. If X is the data matrix containing d variables and n observations, the Covariance matrix can be expressed as follows:

$$\mathbf{C} = \frac{\mathbf{1}}{\mathbf{n-1}} \sum_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{X_i} - \mu)(\mathbf{X_i} - \mu)^\top \tag{3}$$

where $C \in \mathrm{I\!R}^{d \times d}$, μ is the mean value of the variables, and $\top$ represents matrix transposition. In this study we considered genes/miRNAs as variables and patients as observations to compute the inverse of C, i.e., the precision matrix C^{-1}. We built six different networks, three for the cancer and three for the normal patients, based on the three similarity measures described. Networks were first built as fully connected graphs for all gene/miRNA pairs, where similarity coefficients are used as weights of the network node associations. Then, we randomized the expression data and computed again the similarity measures to obtain a reference null distribution [1]; we did so by computing the average null distribution on 10 permuted repetitions of the gene/miRNA expression dataset. From the comparison between real and average permuted distributions of each similarity measure, we derived relevant associations in the networks [1]. In other words, we identified the limit values of each permuted distribution and used them as thresholds in the correspondent real distribution. E.g., Fig. 1 (a) shows that the average permuted distribution for the normal Pearson's correlation has values ranging from -0.2 to 0.4; thus, values of the real normal distribution greater than 0.4 and smaller than -0.2 were considered as representing relevant associations.

3 Results

The six constructed networks have same nodes and different edges/weights, depending on the similarity measure used for each network construction. We focused our unsupervised analysis on the computation of each node *strength*, i.e., the sum of the total weighted connections of each gene/miRNA, in each of the six networks.

3.1 Pearson's correlation networks

Strength distributions of Pearson's correlation networks for normal and cancer condition are shown in Fig. 1 (b), where the x-axis represents the strength values and the y-axis is the proportion of nodes having certain strengths. Interestingly, the proportion of nodes with strength around 0 gets higher in cancer condition (red dashed line), meaning that in cancer many genes/miRNAs have lost their correlation with other genes/miRNAs. We performed a gene set enrichment analysis on the set of genes whose strength changes from high/low in the normal network to almost 0 in the cancer network (156 genes out of 12,792) and found it significantly enriched for the KEGG *metabolic pathways* (p-value 1.65×10^{-25}); indeed, KIRC is known as a metabolic disease.

MiRNAs having high/low strength in normal condition and almost 0 strength in cancer were 5 (out of 1,397): hsa-mir-192, hsa-mir-194-1 and hsa-mir-194-2, which are well known miRNAs involved in cancer, as well as hsa-mir-22 and hsa-mir-378c.

3.2 Euclidean networks

Fig. 1 (c) shows the strength distribution for the nodes of the Euclidean networks, i.e., the networks built using the Euclidean distance as similarity measure between each pair of genes/miRNAs in the processed KIRC dataset in cancer (red dashed line) or normal (blue full line) condition, respectively. Fig. 1 (c) shows higher values of strength in cancer compared to the strengths in the normal network, i.e., $[4.0 \times 10^3, 2.65 \times 10^7]$ vs. $[1.5 \times 10^3, 1.0 \times 10^7]$, respectively. The y-axis scale permits the identification of a set of outlier nodes having high values of strength in both normal and cancer conditions, i.e., hsa-mir-10b, hsa-mir-30a, hsa-mir-22 and hsa-mir-143; these miRNAs maintain high Euclidean distances with all the other genes/miRNAs in the dataset from normal to cancer condition. Instead, hsa-mir-10a has one of the highest strength in the normal network and low strength in cancer, with FPKM values over-expressed in normal condition compared to cancer, where its regulatory activity could be disrupted.

3.3 Inverse Covariance networks

Also the inverse Covariance networks show different strength distributions in normal and cancer conditions, as presented in Fig. 1 (d). The dependencies between pairs of genes/miRNAs conditioned for all the other genes/miRNAs, here used as weights of the inverse Covariance networks, are lower in cancer than in normal network. However, Fig. 1 (d) shows that inverse Covariance values in both normal and cancer networks are very close to 0; this means that, even if inverse Covariance coefficients have greater values in normal than in cancer, they do not represent real dependency between genes/miRNAs in either condition.

3.4 Network comparison

The strength analysis performed allowed us to identify relevant RNAs to be further investigated. For example, hsa-mir-22 has an interesting behaviour in both Pearson's correlation networks and Euclidean networks. It has high values of Pearson's correlation coefficients with all the other genes/miRNAs in normal condition, however it does not maintain these high correlations in cancer. It also has one of the highest value of strength in both Euclidean networks, i.e., it has very distant FPKM expression values from each other gene/miRNA in the network both in cancer and normal condition; furthermore, these Euclidean distances get wider in cancer, where hsa-mir-22 doubles its strength compared to the one in the normal network, with its FPKM mean value increasing in cancer (to 396,490 from 332,072 in the normal condition). These features together make hsa-mir-22 a miRNA of interest for the analysis of gene/miRNA interactions in KIRC. Another interesting miRNA is hsa-mir-10a; it is one of the outliers with high value of strength in the normal Euclidean network, and it has very low strength in the cancer Euclidean network; moreover, its strength values in Pearson's correlation networks are significantly different from normal to cancer condition (1,128 vs. 380, respectively). Thus, in normal condition this miRNA has FPKM expression values distant from those of the other genes/miRNAs, but highly correlated with them, whereas in cancer they get closer to the ones of the other genes/miRNAs and their correlation to them decreases. Hsa-mir-10a has 290,026 and 140,536 mean FPKM values in normal and cancer condition, respectively; thus, it is over-expressed in normal condition.

4 Conclusions

In this study we proposed an unsupervised data-driven framework based on complex networks to better represent and understand gene/miRNA relationships and interactions based on gene expression data. To this aim, we preprocessed the public gene expression data of kidney renal clear cell carcinoma from the TCGA project, and we computed three different similarity measures between genes/miRNAs to get different normal and cancer network representations. Comparative analysis of the six networks obtained lead us to identify two interesting miRNAs: hsa-mir-22 and hsa-mir-10a. They are not significantly differentially expressed; yet, they display important features in both Euclidean and Pearson's correlation networks. According to Euclidean networks, hsa-mir-22 has highly different expression from other genes/miRNAs in both normal and cancer conditions, and hsa-mir-10a only in normal condition; however, based on Pearson's correlation networks, from normal to cancer condition both miRNAs lose many correlations with other genes/miRNAs, i.e., they co-regulate with a lower number of genes/miRNAs. Dysregulated miRNAs play an important role in cancer initiation and progression, they have also showed great potential as novel diagnostic/prognostic biomarkers of cancer [10]. Our findings support this assumption and stress the importance of understanding the function of miRNAs as gene suppressors. Future work will further explore the created networks with ad hoc network algorithms, and will investigate the role of miRNAs in the networks.

Acknowledgments

This research is funded by the ERC Advanced Grant project 693174 GeCo (Data-Driven Genomic Computing), 2016-2021.

References

[1] M. Banf and S.Y. Rhee. "Computational inference of gene regulatory networks: Approaches, limitations and opportunities". *Biochim Biophys Acta Gene Regul Mech*, vol. 1860, pp. 41–52, 2017.

[2] R.Y.X. Wang and H. Huang. "Review on statistical methods for gene network reconstruction using expression data". *J Theor Biol*, vol. 362, pp. 53–61, 2014.

[3] E. de Silva and M.P.H. Stumpf. "Complex networks and simple models in biology". *J R Soc Interface*, vol. 2, no. 5, pp. 419–430, 2005.

[4] J.N. Weinstein, E.A. Collisson, G.B. Mills, K.M. Shaw, B.A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, J.M. Stuart. "The Cancer Genome Atlas pan-cancer analysis project". *Nat Genet*, vol. 45, no. 10, pp. 1113–1120, 2013.

[5] D.P. Bartel. "MicroRNAs: target recognition and regulatory functions". *Cell*, vol. 136, no. 2, pp. 215–233, 2009.

[6] J.M. Perkel. "Visiting "noncodarnia"". *BioTechniques*, vol. 54, no. 6, pp. 301–304, 2013.

[7] T. Hart, H.K. Komori, S. LaMere, K. Podshivalova and D.R. Salomon. "Finding the active genes in deep RNA-seq gene expression studies". *BMC Genomics*, vol. 14, no. 778, 2013.

[8] A. Howard. "Elementary Linear Algebra". John Wiley & Sons, New York, pp. 170–171, 1994.

[9] N.G. van Kampen. "Stochastic processes in physics and chemistry". North-Holland, New York, 1981.

[10] L. Huiyin, L. Haiqi, W. Xian and J. Hongchuan. "MicroRNAs as potential biomarkers in cancer: opportunities and challenges". *BioMed research international*, vol. 2015, 2015.