Network modeling and analysis of normal and cancer gene expression data

Gaia Ceddia[0000−0001−9512−7781], Sara Pidò[0000−0003−1425−1719], and Marco Masseroli[0000−0003−2574−1174]

Politecnico di Milano - Dipartimento di Elettronica, Informazione e Bioingegneria, Piazza Leonardo da Vinci 32, Milan, Italy
first.last@polimi.it

Abstract. Network modelling 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.

Keywords: Gene networks, microRNA, gene expression profiles, complex networks, similarity networks, co-expression.

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 [1]. 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 [1].

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. However, its limitation lies on the fact that causality and direction of the gene interactions are ignored in the computation [4]. Zhang et al. [5] performed a *Weighted Gene Co-expression Network Analysis (WGCNA)* providing interesting communities of genes; nonetheless they carry several false positives. Some methods tried to handle the over-connectivity of co-expression networks by comparing the network structures among cancer types [6]. Other methods for the construction of gene networks include Bayesian network approaches, as well as regression and differential equation based models [1]. 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 consider three different similarity measures for the construction of gene co-expression networks and we innovatively deal with the over-connectivity of similarity gene networks by using three statistical thresholding steps. In particular, we focus on co-expression networks built by computing Pearson's correlation, Euclidean distance and inverse Covariance metrics. The first similarity 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 [7], 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. To our knowledge, this study is a novel approach for comparing different similarity co-expression networks using human datasets; other attempts were done on *S. cerevisiae* and *S. pombe* organisms [8]. In addition, we integrate long RNA and miRNA expression data as done in Pian et al. [9], although we innovatively

take advantage of three similarity measures to compare the overall differences of the gene networks in normal and cancer data by using the strength analysis. The novel use of strength comparisons lead us to find some relevant miRNAs by clearly displaying their dysregulation between normal and cancer Euclidean distance and the Pearson's correlation networks.

For the considered datasets, we integrate 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 [10]. 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, and thereby to negatively regulate mRNA translation [10]. 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 [11].

By focusing on whole gene co-expression networks in normal and cancer conditions we decide not to only select differentially expressed (DE) genes. DE genes are the ones showing statistically significant changes in read counts, or expression levels, between two experimental conditions. However, not significant DE genes, or genes with small changes in their expression levels, may play an important role due to the interaction of their products with other proteins and gene products; thus, our method is purely based on network comparison without any prior biological assumption on DE genes.

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. The whole process is represented in Figure 1.

2.1 Data Extraction and Pre-processing

We consider 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 all gene expression quantification. For each miRNA-Seq and RNA-Seq dataset of each tumor type in TCGA, we compute the number of normal and cancer condition samples from patients. In our datasets, each patient

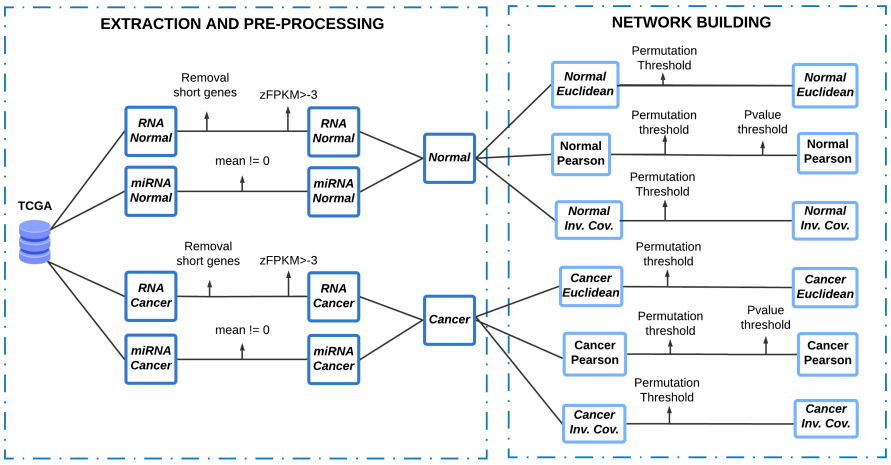

Fig. 1. Defined workflow that starts with the extraction and the pre-processing of TCGA data in order to build the gene co-expression networks.

corresponds to one sample, thus, in this study, the term patient and sample have the same meaning. Figure 2 shows the number of normal samples and the ratio between the number of normal and cancer samples for all tumor types in the TCGA repository. KIRC results as one of the tumor types with the highest ratio and number of normal samples, providing balanced normal and cancer datasets. Thus, we choose KIRC because it has the highest number of normal samples

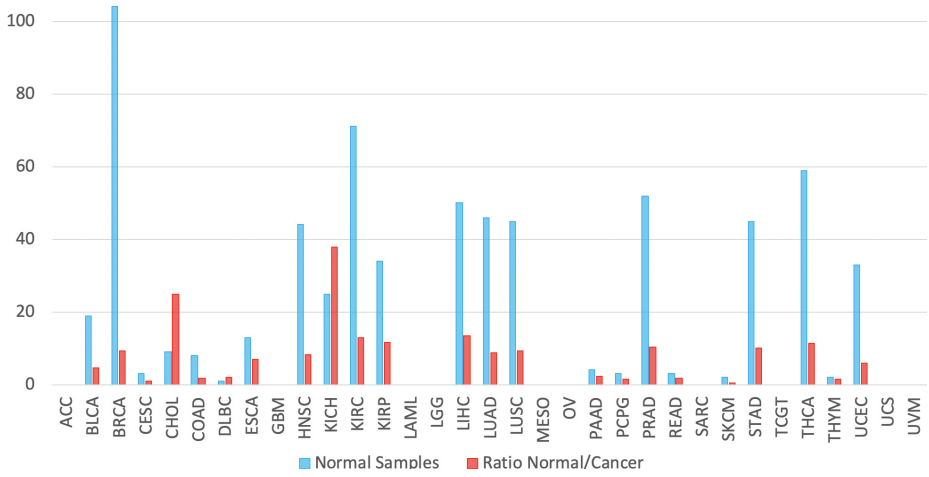

Fig. 2. Number of normal samples (blue bars) and ratio between the number of normal and cancer samples (red bars) for all tumor types in TCGA data.

after BRCA, but BRCA has low ratio of normal/cancer samples. As shown in Table 1, KIRC RNA-Seq dataset resulted to have 72 and 534 samples for normal and cancer conditions, respectively, and KIRC miRNA-Seq dataset 71 and 541 samples for normal and cancer conditions, respectively. Thus, we use these KIRC data for our analysis. After the selection and extraction of KIRC RNA-Seq and miRNA-Seq datasets performed with GMQL [12], we have 60,483 RNAs and 1,881 miRNAs for each sample, as shown in Table 2.

Table 1. Number of samples in TCGA KIRC data.

	Normal	**Cancer**	**Total**
RNA-Seq	72	534	606
miRNA-Seq	71	541	612
Common samples	**71**	**487**	**558**

Table 2. Number of RNA molecules in each sample during the filtering steps.

	RNAs	**miRNAs**	**Total**
Extraction with GMQL	60,483	1,881	62,364
Removal of RNAs of short genes	27,144	1,881	29,025
Removal of RNAs with zero mean	26,706	1,397	28,103
Removal of RNAs with zFPKM< -3	**12,792**	**1,397**	**14,189**

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, we remove them from the RNA-Seq dataset, and we select only data of protein coding and long non-coding genes (as reported in the second row of Table 2), which we integrate with the miRNA-Seq dataset ones, considering only common samples (as reported in the last row of Table 1).

We arrange these public gene expression data from the TCGA repository in the form of matrices; we assemble two RNA-Seq and two miRNA-Seq matrices (two for normal and two for cancer data) in which rows represent genes/miRNAs, columns represent samples and each matrix element represents an expression level. TCGA miRNA-Seq expression levels are available as reads per million miRNAs mapped (RPM); conversely, the expression levels in the TCGA RNA-Seq data are provided as fragments per kilobase per million mapped reads (FPKM). To integrate the two miRNA-Seq and RNA-Seq datasets, we transform miRNA expression data to be homogeneous with the RNA expression data; we convert RPM expression levels into FPKM ones by multiplying each element of the miRNA-Seq matrices by 1000 and dividing it by the double of the length of the corresponding miRNA [13].

After selecting the RNA molecules of interest for each dataset, i.e., protein coding, long non-coding genes and miRNAs, we delete miRNAs and RNAs

with null expression in all normal and cancer samples, as reported in the third row of Table 2. Furthermore, to separate biologically relevant genes from low-expression noisy ones, on the RNA-Seq data we apply the zFPKM normalization method [14]. For normal and tumoral cases separately, we compute the mean and the standard deviation of the log-transformed expression distribution of each gene across all KIRC samples, and we normalize 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 remove 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 [14].

Thus, we obtain 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 samples with KIRC tumor, respectively (Table 2). These two matrices contain all the relevant FPKM values needed to build then the desired networks.

2.2 Building the Networks

To build adjacency matrices describing gene networks, we consider three different similarity measures: Euclidean distance, Pearson's correlation and inverse Covariance. As mentioned in Section 1, we use 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 [15]:

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

We apply the Euclidean distance on each pair of genes/miRNAs in the datasets, considering the n samples 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 variable X and Y is defined as:

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

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 compute pairwise Pearson's correlation on each pair of genes/miRNAs in the datasets, and use the Pearson's coefficients to represent the weights of the edges connecting two nodes (i.e., genes/miRNAs) in the networks.

The inverse Covariance matrix, commonly referred to as *precision matrix*, displays information about the partial correlations of variables [16]. In the Covariance matrix, the (i,j)-th element represents the unconditional correlation between a variable i and a variable j [16]. 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 [16]. In other words, it gives the co-variation of two variables while conditioning on the potential influence of the other variables involved in the analysis, i.e., it removes the effect of other variables. Thus, the precision matrix allows obtaining direct co-variation between two variables by capturing partial correlations. If X is the data matrix containing k 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 $C \in \mathbb{R}^{k \times k}$, μ is the mean value of the variables, and $\top$ represents matrix transposition. In this study we consider genes/miRNAs as variables and samples as observations to compute the inverse of C, i.e., the precision matrix C^{-1}.

We build six different networks, three for the cancer and three for the normal conditions, based on the three similarity measures described. Networks are 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 randomize the expression data and compute again the similarity measures to obtain a reference null distribution [1]; we do 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 derive relevant associations in the networks [1]. In other words, we identify the limit values of each permuted distribution and use them as thresholds in the correspondent real distribution. Table 3 shows how the number of edges changes after each filtering step. E.g., Figure 3 (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 are considered as representing the only relevant associations, and links whose values range from -0.2 to 0.4 are deleted.

Furthermore, since the networks created with Pearson's correlation are very dense, we use the computed p-value of the Pearson's statistic to further threshold them. We sort the computed Pearson's p-values and we only consider the network edges associated with the 99^{th} percentile of the first ten percent of these p-values (i.e., the 0.1% of the edges of the fully-connected network). The third column of Table 3 shows the number of edges after the p-value threshold.

Table 3. Number of edges in the networks. Step 1 represents the creation of networks phase, Step 2 is the filtering phase by the permutation method and Step 3 filters non-relevant edges from Pearson's networks by p-value analysis.

	Step 1	**Step 2**	**Step 3**
Normal Pearson	194,140,866	39,392,104	3,959,308
Cancer Pearson	200,789,224	22,728,618	2,249,150
Normal Euclidean	201,247,218	141,846	141,846
Cancer Euclidean	201,313,190	113,492	113,492
Normal Inverse Covariance	194,073,019	14,300	14,300
Cancer Inverse Covariance	200,788,919	14,230	14,230

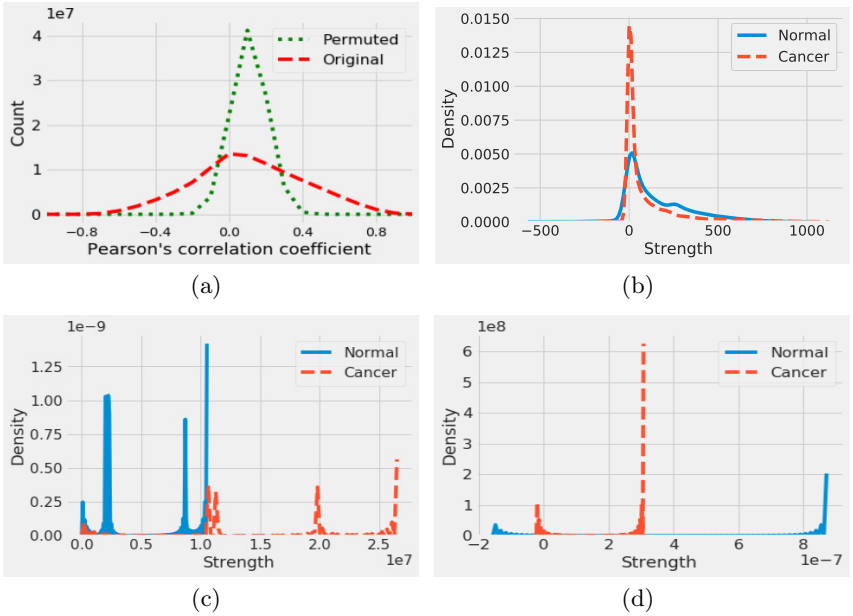

Fig. 3. (a) Red dashed line represents the distribution of Pearson's correlation coefficients for the gene/miRNA expression dataset in normal condition. Dotted green line represents the distribution of the average Pearson's correlation coefficients on 10 permuted repetitions of the gene/miRNA expression 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. Density is the proportion of network nodes having certain strengths.

3 Results

The six constructed networks have same nodes and different edges/weights, depending on the similarity measure used for each network construction (as shown in the fourth row of Table 2 and in the third column of Table 3). 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 Figure 3 (b), where the x-axis represents the strength values and the y-axis is the proportion of network 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, or relevantly lowered, their correlation with other genes/miRNAs. We perform 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 (180 genes out of 12,792). We find this gene set significantly enriched for several KEGG pathways related to cancer, particularly for *metabolic pathways*, as shown in Table 4; indeed, KIRC is known as a metabolic disease [17].

Table 4. Results of KEGG gene set enrichment analysis: first column contains the term name of KEGG pathways, second column reports the term ID, and the third column contains the adjusted p-values, which is the correction of p-values performed by [18]

KEGG pathway	ID	Adj. p-value
Metabolic pathways	**01100**	5.697×10^{-27}
Gastric cancer	05226	1.128×10^{-5}
Pathways in cancer	05200	7.181×10^{-5}
Proteoglycans in cancer	05205	3.244×10^{-4}
Transcriptional misregulation in cancer	05202	7.710×10^{-4}
Hepatocellular carcinoma	05225	1.743×10^{-3}

MiRNAs having high/low strength in normal condition and almost 0 strength in cancer are 9 (out of 1,397), including *hsa-mir-192*, *hsa-mir-194-1* and *hsa-mir-194-2*, which are well known miRNAs involved in cancer [19]. Out of the other 6, *hsa-mir-1266* is associated with epithelial tissue diseases, *hsa-mir-210*, *hsa-mir-218-1* and *hsa-mir-218-2* are known to be involved in breast cancer, *hsa-mir-934* is up-regulated in papillary renal cell carcinoma, and *hsa-mir-22* acts as an oncogenic mirna in renal cell carcinoma [20–24].

3.2 Euclidean Networks

Figure 3 (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 cancer (red dashed line) or normal (blue full line) condition, respectively. Figure 3 (c) shows higher values of strength in

cancer compared to the strengths in the normal network, i.e., $[4.0\times10^3, 2.65\times10^7]$ vs. $[1.5\times10^3, 1.0\times10^7]$, 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. In the literature *hsa-mir-10b* is known to be associated with *Non-Alcoholic Fatty Liver Disease* and *Bladder Cancer* [19]. Also *hsa-mir-30a* has been studied for its involvement in cancer development, in particular for its potential role as a diagnostic or prognostic marker of gliomas [25]. *Hsa-mir-143* has been associated with *Burkitt Lymphoma* and *Diffuse Large B-Cell Lymphoma* [19]. All three miRNAs are related to *MicroRNAs in cancer* pathway [19]. Moreover, as mentioned in Section 3.1, *hsa-mir-22* has been studied for its ability to repress cancer progression in *clear cell renal cell carcinoma* [24]. Instead, *hsa-mir-10a* has one of the highest strength in the normal network and low strength in cancer, with FPKM values over-expressed but not significantly in normal condition compared to cancer, where its regulatory activity could be disrupted. Biologically, *hsa-mir-10a* is associated with several diseases, including *renal cell carcinoma* [26]. Moreover, it is involved in two relevant pathways: *Proteoglycans in cancer* and *MicroRNAs in cancer* [19].

3.3 Inverse Covariance Networks

The inverse Covariance networks show different strength distributions in normal and cancer conditions, as presented in Figure 3 (d). The dependencies between pairs of genes/miRNAs conditioned for all the other genes/miRNAs, here used as edge weights of the inverse Covariance networks, are lower in cancer than in normal network. However, Figure 3 (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 a real dependency between genes/miRNAs in either condition.

3.4 Network Comparison

The strength analysis performed allows 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 distance 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 distance networks, i.e., it has very distant FPKM expression values from any 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). Gong et al. [24] found that hsa-mir-22 targets directly PTEN in *renal cell carcinoma*; thus, the increase of

expression levels in hsa-mir-22 leads to the downregulation of the PTEN protein, indicating an oncogenic effect of the miRNA. However, this miRNA has 15 targets out of 250 significantly related to the KEGG *Pathways in cancer*; thus, its loss of correlations in the Pearson's correlation cancer network may cause a cascade of dysregulation in cancer. 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 distance network, and it has very low strength in the cancer Euclidean distance network. Moreover, its strength values in Pearson's correlation networks are very 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, but not statistically significant. The antitumor role of hsa-mir-10a has been studied in Arai et al. [26] for its interaction with the SKA1 oncogene, explaining the computed downregulation in normal condition. Moreover, 21 targets out of 463 of hsa-mir-10a are significantly involved in the KEGG *Pathways in cancer*; thus, the reported change in correlations between normal and cancer condition may represent abnormal co-regulations of the miRNA-RNAs interaction network.

4 Conclusions

In this study we propose an unsupervised data-driven framework based on complex networks to better represent and understand gene/miRNA relationships and interactions based on gene expression data. We implement a novel pipeline to compute the gene co-expression networks that comprises the pre-processing and the construction phases[1]. Normally, these steps are taken for granted; indeed, it is very difficult to find a complete and efficient workflow.

To this aim, we preprocess the public gene expression data of kidney renal clear cell carcinoma from the TCGA project, and we compute 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 differentially expressed; yet, they display important features in both Euclidean and Pearson's correlation networks. According to Euclidean distance 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. Interestingly, in miRNet[2] hsa-mir-10a

[1] https://github.com/DEIB-GECO/GeneNetFusion/blob/master/preprocessing.py
[2] https://www.mirnet.ca/miRNet

and hsa-mir-22 share an interaction network of 12 genes enriched in a particular KEGG pathway called *Pathways in cancer*. Among them ERBB2, PIK3CG, PTEN and XIAP are known in the literature to have a relevant role in KIRC development [27–30]. Dysregulated miRNAs play an important role in cancer initiation and progression involving their targets [31]; they have also shown great potential as novel diagnostic/prognostic biomarkers of cancer [32]. 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 deeper 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 **1860**(1), 41–52 (2017)
2. Y. R. Wang and H. Huang: Review on statistical methods for gene network reconstruction using expression data. J Theor Biol **362**, 53–61 (2014)
3. E. de Silva and M. P. Stumpf: Complex networks and simple models in biology. J R Soc Interface **2**(5), 419–430 (2005)
4. M. M. Saint-Antoine and A. Singh: Network inference in systems biology: recent developments, challenges, and applications. Curr Opin Biotechnol **63**, 89–98 (2020)
5. B. Zhang and S. Horvath: A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol **4**(1), Article17 (2005)
6. M. A. Care, D. R. Westhead and R. M. Tooze: Parsimonious Gene Correlation Network Analysis (PGCNA): a tool to define modular gene co-expression for refined molecular stratification in cancer. NPJ Syst Biol Appl **5**(1), 1–17 (2019)
7. J. N. Weinstein, E. A. Collisson, G. B. Mills et al.: The Cancer Genome Atlas pan-cancer analysis project. Nat Genet **45**(10), 1113–1120 (2013)
8. R. Deshpande, B. VanderSluis and C. L. Myers: Comparison of profile similarity measures for genetic interaction networks. PloS One **8**(7), e68664 (2013)
9. C. Pian, G. Zhang, S. Wu et al.: Discovering the 'Dark matters' in expression data of miRNA based on the miRNA-mRNA and miRNA-lncRNA networks. BMC bioinformatics **19**(1), 379 (2018)
10. D. P. Bartel: MicroRNAs: target recognition and regulatory functions. Cell **136**(2), 215–233 (2009)
11. J. M. Perkel: Visiting "noncodarnia" (2013)
12. M. Masseroli, A. Canakoglu, P. Pinoli et al.: Processing of big heterogeneous genomic datasets for tertiary analysis of next generation sequencing data. Bioinformatics **35**(5), 729–736 (2019)
13. G. P. Wagner, K. Kin and V. J. Lynch: Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory Biosci **131**(4), 281–285 (2012)

14. T. Hart, H. K. Komori, S. LaMere et al.: Finding the active genes in deep RNA-seq gene expression studies. BMC Genomics **14**(1), 778 (2013)
15. H. Anton and C. Rorres: Elementary linear algebra. John Wiley & Sons (1994)
16. N. G. Van Kampen: Stochastic processes in physics and chemistry, vol. 1. Elsevier (1992)
17. W. M. Linehan, R. Srinivasan and L. S. Schmidt: The genetic basis of kidney cancer: a metabolic disease. Nat Rev Urol **7**(5), 277–285 (2010)
18. J. Reimand, M. Kull, H. Peterson et al.: g:profiler — a web-based toolset for functional profiling of gene lists from large-scale experiments. Nucleic Acids Res **35**(suppl_2), W193–W200 (2007)
19. G. Stelzer, N. Rosen, I. Plaschkes et al.: The GeneCards suite: from gene data mining to disease genome sequence analyses. Curr Protoc Bioinformatics **54**(1), 1–30 (2016)
20. N. S. Seifeldin, S. B. El Sayed and M. K. Asaad: Increased Micro RNA-1266 levels as a biomarker for disease activity in psoriasis vulgaris. Int J Dermatol **55**(11), 1242–1247 (2016)
21. B. Pasculli, R. Barbano, M. Rendina et al.: Hsa-miR-210-3p expression in breast cancer and its putative association with worse outcome in patients treated with Docetaxel. Sci Rep **9**(14913), 1–9 (2019)
22. W. Luo, L. Wang, M.-H. Luo et al.: hsa-mir-3199-2 and hsa-mir-1293 as novel prognostic biomarkers of papillary renal cell carcinoma by Cox ratio risk regression model screening. J Cell Biochem **118**(10), 3488–3494 (2017)
23. Q. Li, F. Zhu and P. Chen: miR-7 and miR-218 epigenetically control tumor suppressor genes RASSF1A and Claudin-6 by targeting HoxB3 in breast cancer. Biochem Biophys Res Commun **424**(1), 28–33 (2012)
24. X. Gong, H. Zhao, M. Saar et al.: mir-22 regulates invasion, gene expression and predicts overall survival in patients with clear cell renal cell carcinoma. Kidney cancer **3**(2), 119–132 (2019)
25. K. Wang, Z. Jia, J. Zou et al.: Analysis of hsa-mir-30a-5p expression in human gliomas. Pathol Oncol Res **19**(3), 405–411 (2013)
26. T. Arai, A. Okato, S. Kojima et al.: Regulation of spindle and kinetochore-associated protein 1 by antitumor mir-10a-5p in renal cell carcinoma. Cancer Sci **108**(10), 2088–2101 (2017)
27. H. Kędzierska, P. Popławski, G. Hoser et al.: Decreased expression of SRSF2 splicing factor inhibits apoptotic pathways in renal cancer. Int J Mol Sci **17**(10), 1598 (2016)
28. M. Liontos, E.-A. Trigka, P. Korkolopoulou et al.: Expression and prognostic significance of VEGF and mTOR pathway proteins in metastatic renal cell carcinoma patients: a prognostic immunohistochemical profile for kidney cancer patients. World J Urol **35**(3), 411–419 (2017)
29. W.-c. Que, H.-q. Qiu, Y. Cheng et al.: PTEN in kidney cancer: A review and meta-analysis. Clin Chim Acta **480**, 92–98 (2018)
30. S. Reuter, S. Prasad, K. Phromnoi et al.: Thiocolchicoside exhibits anticancer effects through downregulation of NF-κB pathway and its regulated gene products linked to inflammation and cancer. Cancer Prev Res **3**(11), 1462–1472 (2010)
31. G. Sharma, P. Dua and S. Mohan Agarwal: A comprehensive review of dysregulated miRNAs involved in cervical cancer. Curr Genomics **15**(4), 310–323 (2014)
32. H. Lan, H. Lu, X. Wang et al.: MicroRNAs as potential biomarkers in cancer: opportunities and challenges. Biomed Res Int **2015**, 15–31 (2015)