Identification of disease-miRNA networks across different cancer types using the SWIM software

Giulia Fiscon (1,2), Federica Conte (1,2), Lorenzo Farina (3), Marco Pellegrini (4), Francesco Russo* (5) and Paola Paci* (1,2)

(1) Institute for Systems Analysis and Computer Science Antonio Ruberti, National Research Council, Rome, Italy.
(2) SysBio Centre for Systems Biology, Italy.
(3) Department of Computer, Control, and Management Engineering Antonio Ruberti, Sapienza University of Rome, Italy.
(4) Institute of Informatics and Telematics, National Research Council, Pisa, Italy.
(5) Novo Nordisk Foundation Center for Protein Research, Translational Disease Systems Biology, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark.

*Corresponding authors: francesco.russo@cpr.ku.dk, paola.paci@iasi.cnr.it

1. Introduction

The idea that the greater complexity of higher eukaryotes arises from the portion of the genome called non-coding RNAs (ncRNAs) is becoming increasingly widespread [Costa 2008, Mattick 2009]. Indeed, ncRNAs are of growing interest, as they have been found to be important regulators of gene expression in development, physiology, and, when dysfunctional, in the presence of disease. This variegated class of RNA species encompasses the well-known microRNAs (miRNAs), as well as the most recently acknowledged long non-coding RNAs (lncRNAs). Discovered first, miRNAs have been intensively studied and much is now known about their biological functions, as opposed to lncRNAs that constitutes a new, potentially fascinating, territory to be explored yet. miRNAs are single-stranded short RNAs (∼22 nucleotides long) that post-transcriptionally regulate gene expression by translation inhibition or degradation of their target mRNAs [Filipowicz 2008, Bartel 2009, Mercer 2009, Koziol 2010]. Virtually, all biological processes have been proved to involve miRNA regulation, including development, metabolism, cell proliferation, differentiation and apoptosis [Lee 19993, Brennecke 2003]. Accordingly, altered miRNA expression characterizes many human diseases and mounting evidence strongly links specific miRNAs to tumor initiation, progression and metastasis [Lu Nature 2005, Iorio 2007, Gaur 2007, Ma Nature 2007, Garzon 2009, Spizzo 2009, Reddy 2015, Peng 2016, Catalanotto 2016].

However, a disease is rarely a direct consequence of an abnormality in a single gene or molecular component, rather nearly always the result of multiple pathobiological pathways that interact through an
interconnected network. A network is a set of nodes and edges, where nodes are linked together if an interaction occurs between them. Networks can be obtained from any sort of information, such as protein interaction networks, whose nodes are proteins that are linked to each other by physical interactions [Rual 2005, Stelzl 2005]; metabolic networks, whose nodes are metabolites that are linked if they participate in the same biochemical reactions [Jeong, 2000, Fell 2000, Duarte 2007]; regulatory networks, whose directed links represent regulatory relationships between a transcription factor and a gene [Carninci 2005]; co-expression networks, in which transcripts with similar co-expression patterns are linked [Stuart 2003].

Network-based approach has proved to be effective in analyzing complex data from various sources, as witnessed by the large number of top-ranking research on the topic. In this context, Paci et al. [Paci et al, 2017 SR] has developed a sophisticated network-based algorithm, denoted as SWitchMiner (SWIM) able to identify key genes in a network of interactions by defining appropriate “roles” to genes according to their local/global positioning in the overall network. The latter property being of crucial importance, given that, recently, it has been shown that genes associated with a disease are localized in specific neighborhoods, or “disease modules”, within the interactome (i.e., the overall network map of molecular components and their interactions) [Barabasi 2011]. So far, SWIM has provided noteworthy results in uncovering key players in biologically relevant scenarios, including viticulture [Palumbo 2014] and oncology [Paci et al. 2017, Fiscon et al. 2018].

2. Material

SWIM is a software with a user-friendly Graphical User Interface (GUI) downloadable from Supplementary information of [Paci et al, 2017]. It implements and integrates a network-based model able to unveil a small pool of genes, called switch genes, which are likely to be critically associated with drastic changes between two conditions (e.g., pathological and physiological) in many biological settings. The algorithm implemented by SWIM encompassed several steps depicted in Figure 1.

3. Methods

3.1 Steps performed by the SWIM method

3.1.1 Step 1: Pre-processing phase

Starting from gene expression data matrix and denoting with A and B the two given conditions between which searching for switch genes, this step requires the selection of two specific thresholds for removing genes whose expression across the total samples S (S = samples in the condition A + samples in the condition B) is mostly zero or change very little. The first threshold regards the maximum number of samples out of S allowed to be equal to zero. The second threshold concerns the minimum variation - measured by the Inter Quartile Range (IQR) percentile - allowed for each gene across the S samples.

3.1.2 Step 2: Filtering phase

This step requires the selection of two specific thresholds for removing genes whose expression between the two given conditions (A and B) does not change enough or does it without statistical significance. Considering the logarithm of the ratio between the average expression of samples in condition A and the average expression of samples in condition B (log fold-change), the first threshold allows to remove the genes falling behind, in absolute value, a fixed cutoff on the log fold-change. The second threshold concerns the smallest probability (p-value) for which the data allow to reject the null hypothesis (i.e., the means of the two distributions – normal and cancer - are identical) of the Student’s t-test. The obtained p-values are
adjusted according to False Discovery Rate (FDR) method [Benjamini and Hochberg, 1995] and thus the threshold refers to the FDR values. At end of this phase, the differentially expressed genes between conditions A and B have been identified.

3.1.3 Step 3: Building the correlation network

This step requires the selection of a threshold for building the correlation network where two nodes are connected if the absolute value of the Pearson correlation between their expression profiles exceeds a given cutoff. This threshold should reflect a right balance between the number of edges and the number of connected components of the network: the number of edges should be as small as possible in order to have a manageable network (pointing towards a higher threshold) and the number of connected components should be as small as possible in order to preserve the integrity of the network (pointing towards a smaller threshold).

3.1.4 Step 4: Finding communities in the network

To find communities in the network, SWIM makes use of the k-means algorithm [Hartigan et al., 1973], a method of cluster decomposition whose aim is to partition n objects (i.e., the nodes of the co-expression network) into N clusters. The goal of the clustering is expressed by an objective function that depends on the proximities of the nodes to the cluster centroids. As objective function, which measures the quality of a clustering, SWIM uses the Sum of the Squared Error (SSE). As measurement of the proximities of two nodes, SWIM makes use of the following metrics:

\[ \rho = \text{Pearson correlation coefficient between the expression profiles of the nodes } x \text{ and } y. \]

\[ \text{dist} = 0 \quad \rho = +1 \]

two nodes are close in the network (_______) if they are highly correlated (_______) on the contrary they

\[ \text{dist} = 2 \quad \rho = -1 \]

are far apart in the network (________) if they are highly anti-correlated (________).

To determine the “correct” number of clusters SWIM uses an approach - named Scree plot- that evaluates the behavior of the SSE function to vary the number of clusters. Then, the position of an elbow in the scree plot - i.e., where the “cliff” reaches a bottom plateau - determines an appropriate number of clusters [Lisboa et al., 2013].

3.1.5 Step 5: Building the heat cartography map

Once the modular structure of the complex network has been found, roles have to be assigned to each node. This is done by dividing the plan according to two parameters, the clusterphobic coefficient \( K_\pi \) and the global within-module degree \( z_s \). The clusterphobic coefficient \( K_\pi \) measures the “fear” of being confined in a cluster, in analogy with the claustrophobic disorder. A high value of \( K_\pi \) denotes nodes having much more external than internal links. The global within-module degree \( z_s \) measures how “well connected” each node is to other nodes in its own community. In the following, the formal definitions of these parameters:

\[ z_s^i = \frac{k_{in}^i - \bar{k}_C}{\sigma_C} \quad K_\pi = 1 - \left( \frac{k_{in}^i}{k_i} \right)^2 \]
where \( k_i^p \) is the number of links of node \( i \) to nodes in its module \( C \), \( k_i \) is the total degree (i.e., number of links emanating from a node) of node \( i \), and \( \bar{k}_C \) and \( \sigma_C \) are the average and standard deviation of the total degree distribution of the nodes in the module \( C \).

This definition of \( z_g \) quantifies how much a node is a hub (i.e., degree exceeding 5 [Han et al., 2004]) in its community and thus represents a measure of local connectivity. On the contrary, the parameter \( K_\pi \) evaluating the ratio of internal to external connections of a node represents a measure of global connectivity. The values of these two parameters define, in the plan identified by \( K_\pi \) and \( z_g \), a cartography made up by seven regions corresponding to seven different roles of nodes in the network [Guimera and Amaral, 2005]. Then, SWIM colors nodes in the cartography according to the Average Pearson Correlation Coefficient (APCC) between the expression profiles of each node and its nearest neighbors [Han et al., 2004]. This representation of the network is called “heat cartography map” [Paci 2017 SR].

By computing the APCC of expression over all interaction partners of each hub in protein-protein interaction (PPI) networks in yeast, the authors in [Han et al., 2004] concluded that hubs fall into two distinct categories: date hubs that display low co-expression with their partners (low APCC) and party hubs that have high co-expression (high APCC). In most gene expression networks [Palumbo 2014, Paci 2017 SR], the distribution of APCCs appears to be trimodal where, similarly to PPI networks, two peaks represent low (date hubs) and high (party hubs) positive APCC values, but with the addition of a new, third peak which is characteristic of gene expression networks and represents negative APCC values. The nodes populating this peak are defined “fight-club hubs” [Palumbo 2014].

3.1.6 Step 6: Identification of switch genes

Looking at the heat cartography map, SWIM identifies “switch genes” as the subset of the fight-club hubs that mainly interact outside their community (region R4 of the heat-cartography map). In particular, they satisfy the following topological and expression features:

1. being not a hub in their own cluster \( (z_g < 2.5) \)
2. having many links outside their own cluster \( (K_\pi > 0.8) \)
3. having a negative average weight of their incident links \( (APCC < 0) \)

3.1.7 Step 7: Evaluating network connectivity and robustness

At the end of step 6, SWIM gives the opportunity to perform further analyses regarding the evaluation of network robustness (i.e. resilience to errors) by studying the effect on the network connectivity of removing nodes by decreasing degree. In particular, SWIM evaluates the effect on the average shortest path (i.e. the shortest path between two nodes is the minimum number of edges connecting them and the average shortest path of a network is the average of the shortest paths for all possible pairs of network nodes) of removing randomly chosen nodes, switch genes, fight-club hubs, date and party hubs.

4. SWIM application for unveiling disease-specific miRNAs

In [Paci et al 2017 SR], SWIM has been successfully applied to a large panel of cancer datasets obtained from The Cancer Genome Atlas (TCGA) [Weistern 2013], in order to highlight switch genes that could be critically associated with the drastic changes in the physiological state of cells or tissues induced by the cancer development. Tumor expression data were from high-throughput RNA- and miRNA-sequencing and correspond to level 3 data (i.e., normalized expression data). For both RNA- and miRNA-sequencing
experiments, only cancer datasets including at least 7 patients with tumor and matched-normal samples (i.e., the matched-normal tissue is defined as the tissue that is adjacent to the tumor and taken from the same patient) were retained for the analysis. A total of 2108 samples relative to 14 tumor types and referring to 540 unique cancer patients were analyzed.

The authors systematically applied SWIM to all the 14 tumor types and an approximately trimodal distribution of the hub APCC values was recovered in each human cancer co-expression network (Figure 2).

From this multi-cancer analysis, the authors discovered switch genes in all cancer datasets (Figure 3) and found that most of the studied tumors shared approximately 100 switch genes implicated in cell cycle regulation. From a functional analysis of their "negative" (i.e., anti-correlated) neighbors, it emerged that switch genes were negative regulators of cancer metabolism. Interestingly, a DNA-binding motifs enrichment analysis of their promoter regions had shown that those switch genes could in turn be controlled by growth factors such as E2F and NFY [Paci et al 2017 SR].

Switch genes identified by SWIM in this multi-cancer analysis encompassed protein coding genes, long non-coding, and miRNAs, recovering many known key cancer players but also many new potential biomarkers not yet characterized in cancer context. Motivated by the universally recognized role of miRNAs as key negative regulators in many intracellular processes as well as in carcinogenesis, here we investigated the microRNAs acting as switch genes in the 14 human cancers of TCGA and we identified both disease-specific miRNAs (Figure 4A) and miRNAs common to different cancer types (Figure 4B).

To better investigate both disease-specific and common miRNAs in the large panel of analyzed cancer datasets, we built a network consisting of two sets of nodes: one set corresponds to the human cancer types under study, whereas the other set corresponds to all miRNAs identified by SWIM as switch genes in each tumor type (Figure 5). A disease and a miRNA are then connected by a link if the miRNA acts as switch gene in that disease. We called this network “miRNA- diseasome”, in analogy with the human diseasome network of [Kwang-Il Goh 2007, Barabasi 2011], where nodes represent disorders, and a link is placed between two disorders if they share at least one gene in which mutations are associated with both disorders.

This representation allows to highlight tumors with the highest number of miRNAs, which are microRNAs specific of each tumor and which ones are shared among different cancer types.

Evidence of multi-cancer microRNAs that are shared by more tumors is a key finding to highlight recurrent deregulated patterns in different cancer types; while identifying disease-specific microRNAs could help to unveil novel therapeutic targets for treating specific diseases.

We then used the Hamming distance to group cancer types based on their degree of microRNA similarity (Figure 6). With the result of this analysis it became possible to point out a cluster of two cancer datasets (i.e. bladder urothelial carcinoma and uterine corpus endometrial carcinoma) sharing a substantial number of microRNA acting as switch genes (N = 37, Figure 6).

Finally, we constructed the co-expression network of disease-specific microRNAs and their negative nearest neighbors. In the specific case of co-expression networks of breast invasive carcinoma (brca), head and neck squamous cell carcinoma (hnsc), and kidney renal clear cell carcinoma (kirc) (Figure 7), we found well-known cancer-specific miRNAs [Volinia 2012, Lopez 2018, Cizeron-Clairac 2015, Fan 2018]. In particular, miR-190-b has been linked to breast invasiveness and prognosis [Volinia 2012, Cizeron-Clairac 2015], miRNA-122 promotes proliferation, invasion, and migration of renal cell carcinoma [Fan 2018]; miR-1293 has been found in the characteristic miRNA expression signature of a broad range of head and neck squamous cell carcinoma subtypes [Lopez 2018]. Interestingly, by exploring the negative first nearest neighborhood, we found predicted (based on TargetScan [Agarwal 2015]) and experimentally validated
(based on miRTarBase [Chou 2015]) targets, as well as new potential mirna-targets interactions worthy for further investigation (Figure 7).

**Figures**

**Figure 1. Flowchart of SWIM.** The figure depicts the steps performed by SWIM and detailed in the text.
Figure 2. Averaged Pearson Correlation Coefficient in 14 human cancers of TCGA. In each panel the dashed curve is the estimated probability density using a smoothing algorithm with a Gaussian kernel of the APCC for each hub (i.e. node with degree greater than 5) of the correlation network obtained using the RNA-sequencing data of expression available on TCGA data portal for the 14 tumor we studied (see legend); while the solid blue line is the corresponding best fitting obtained by using a Gaussian mixture models (GMM) composed of k multivariate normal density components, where k is a positive integer ranging from one to five. The best fit is obtained by using Bayesian information criteria (BIC). In many cases, the distribution of APCCs is trimodal, with two peaks (not always apparent) representing low and high positive APCC values (mirroring the date and party hub distributions in PPI networks) and a third peak (always existing) representing negative APCC values (fight-club hubs). Following [Han 2004], an arbitrary date/party threshold value of APCC equal to 0.5 was defined in order to optimally separate the two types of hubs.
Figure 3. Heat cartography map for 14 human cancers of TCGA. In each panel the cartographic representation of the gene expression network - obtained using the RNA-sequencing data of expression available on TCGA data portal - is represented for each tumor we studied (see legend). The x and y axes correspond to the within-module degree $z_g$ and the clusterphobic coefficient $K_\pi$, respectively. $z_g$ is a normalized measure of intra-module communications of each node, while $K_\pi$ represents the mode of communication between nodes in different modules. Each point represents a node in the network colored according to its APCC value. The plane identified by $K_\pi$ and $z_g$ is divided into seven regions each defining a specific node role. Switch genes correspond to nodes colored in blue and falling in R4 region.
Figure 4. miRNAs acting as switch genes: multi-cancer switch miRNAs (A) and disease-specific switch miRNAs (B). The distribution of miRNAs acting as switch genes across 14 human cancers of TCGA is rendered in a heat map where magenta (white) color indicates whether a given miRNA listed in the row is (is not) a switch gene for the tumors listed in the columns. In particular, Panel A refers to miRNAs acting as switch genes that are shared by more than one cancer type (multi-cancer miRNAs) and Panel B refers to miRNAs that has been found in only one tumor (disease-specific miRNAs). In the histogram of panel A, each grey bar represents the number of cancer types where a given miRNA listed in the corresponding row has been identified as switch gene. Abbreviations. blad: bladder urothelial carcinoma; brca: breast invasive carcinoma; chol: cholangiocarcinoma; coad: colon adenocarcinoma; hnsc: head and neck squamous cell carcinoma; kich: kidney chromophobe; kirc: kidney renal clear cell carcinoma; kirp: kidney renal papillary cell carcinoma; lihc: liver hepatocellular carcinoma; luad: lung adenocarcinoma; lusc: lung squamous cell carcinoma; prad: prostate adenocarcinoma; thca: thyroid carcinoma; ucec: uterine corpus endometrial carcinoma.

Figure 5. miRNA-diseasome. The network is composed of two sets of nodes with different size: the larger ones correspond to the analyzed human cancer types from TCGA, whereas the smaller ones correspond to all microRNAs acting as switch genes. A link is placed between a tumor type and a microRNA if the microRNA has been identified by SWIM as switch gene for that tumor. Different colors are associated to different tumor types. Disease-specific miRNAs are colored according to the tumor type to which they belong. MicroRNAs shared by more than one tumor are colored in grey. Abbreviations. blad: bladder urothelial carcinoma; brca: breast invasive carcinoma; chol: cholangiocarcinoma; hnsc: head and neck squamous cell carcinoma; kich: kidney chromophobe; kirc: kidney renal clear cell carcinoma; kirp: kidney renal papillary cell carcinoma; lihc: liver hepatocellular carcinoma; luad: lung adenocarcinoma; lusc: lung squamous cell carcinoma; prad: prostate adenocarcinoma; thca: thyroid carcinoma; ucec: uterine corpus endometrial carcinoma.
Figure 6. Comparative analysis of microRNAs that are switch genes in the large panel of TCGA datasets. The distance matrix computed based on Hamming distance of binary-encoded (1 = present; 0 = absent) annotation of microRNAs across cancers is rendered in a symmetrical heat map where decreasing distance is rendered in a white to red scale. Dendrogram on columns and rows of the distance matrix indicate cancer clustering based on Hamming distance. Abbreviations. blad: bladder urothelial carcinoma; brca: breast invasive carcinoma; chol: cholangiocarcinoma; hnscc: head and neck squamous cell carcinoma; kic: kidney chromophobe; kirc: kidney renal clear cell carcinoma; kirkp: kidney renal papillary cell carcinoma; lihc: liver hepatocellular carcinoma; luad: lung adenocarcinoma; lusc: lung squamous cell carcinoma; prad: prostate adenocarcinoma; thca: thyroid carcinoma; ucec: uterine corpus endometrial carcinoma.

Figure 7. Co-expression networks of disease-specific microRNAs and their negative nearest neighbor in three tumor types. The networks report the negative nearest neighbors of the disease-specific microRNAs acting as switch genes for breast invasive carcinoma (panel A), head and neck squamous cell carcinoma (panel B), and kidney renal clear cell carcinoma (panel C). In each panel, disease-specific microRNAs are highlighted in magenta, predicted targets by TargetScan in cyan, and experimentally validated targets by miRTarBase in blue. Grey nodes correspond to novel interactions. Abbreviations. brca: breast invasive carcinoma; hnscc: head and neck squamous cell carcinoma; kic: kidney renal clear cell carcinoma.
References


Cizeron-Clairac, Geraldine, et al. "MiR-190b, the highest up-regulated microRNA in ERα-positive compared to ERα-negative breast tumors, a new biomarker in breast cancers?" *BMC cancer* 15,1 (2015): 499.


