Pacific Symposium on Biocomputing 15:397-408(2010)

# CRITICAL ANALYSIS OF TRANSCRIPTIONAL AND POST-TRANSCRIPTIONAL REGULATORY NETWORKS IN MULTIPLE MYELOMA

MARTA BIASIOLO[1*], MATTIA FORCATO[2*], LINO POSSAMAI[3], FRANCESCO FERRARI[1], LUCA AGNELLI[4], MARTA LIONETTI[4], KATIA TODOERTI[4], ANTONINO NERI[4], MASSIMO MARCHIORI[3], STEFANIA BORTOLUZZI[1], SILVIO BICCIATO[2]

[1]*Dipartimento di Biologia, Università di Padova, via U. Bassi 58/ B, 35121, Padova, Italy*
[2]*Dipartimento di Scienze Biomediche, Università di Modena, via G. Campi 287, 41125, Modena, Italy*
[3]*Dipartimento di Matematica Pura ed Applicata, Università di Padova, via Trieste 63,35121, Padova, Italy*
[4]*Dipartimento di Scienze Mediche, Università di Milano, Ematologia 1, CTMO, Fondazione IRCCS Ospedale Policlinico, Via F. Sforza, 35, 20122 Milano*

Network analysis has emerged as a powerful approach to understand complex phenomena and organization in social, technological and biological systems. In particular, it is increasingly recognized the role played by the topology of cellular networks, the intricate web of interactions among genes, proteins and other molecules regulating cell activity, in unveiling the biological mechanisms underlying the physiological states of living organisms. In this study, critical analysis of network components has been applied to inspect the transcriptional and post-transcriptional regulatory networks reconstructed from mRNA and microRNA expression data of multiple myeloma (MM) samples. Specifically, the importance of a gene as a putative regulatory element has been assessed calculating the drop in the network performance caused by its deactivation instead of quantifying its degree of connectivity. The application of critical analysis to transcriptional and post-transcriptional regulatory networks allowed inferring novel regulatory relations potentially functional in multiple myeloma.

## 1. Introduction

Systems biology elevates the study from the single entity level (e.g., genes, proteins) to higher hierarchies, such as entire genomic regions, groups of co-expressed genes, functional modules, and networks of interactions. The functioning and development of a living organism is controlled by the networks of relations among its genes (as well as proteins and small molecules) and the signals regulating each gene (or set of genes), therefore understanding how elementary biological objects act together and interact in the general context of a genome is fundamental to the advancement of science. As such, the scientific attention is focusing more and more on the critical levels of biological organization and their emerging properties rather than on the single components of the system [1]. However, despite the significant advances in genome sequencing and in transcription, protein and metabolite profiling, still significant limitations hamper the global understanding of regulatory phenomena. The control of biochemical processes is hierarchical and originates at the level of transcription (induction-repression mechanism and mRNA degradation), moving on to translation (protein activation and proteolysis) and enzyme activity through signaling cascades. The presence of several feedback loops among these regulatory processes makes their organization and functioning very complex. This level of complexity can, at least in part, be addressed using methods that allow the reverse engineering and the reconstruction of regulatory networks. In this context, the availability of high-throughput genomic data, coupled with bioinformatics tools for their analysis, represents a promising starting point in the identification of molecular interaction networks which will allow turning genomic researches into accurate biological hypotheses.

Microarray experiments have been extensively used to detect patterns in gene expression that stem from regulatory interactions and lately have been applied to analyze the transcriptional activity of microRNAs (miRNAs), i.e. small non-coding RNAs that regulate the post-transcriptional mRNA stability by binding 3' target sites. The availability of a sufficient number of matched miRNA-mRNA expression profiles represents a further opportunity to deepen the study of transcriptional regulation. However, standard methodologies for the analysis of gene expression profiles, which aim at identifying relevant genes from the statistical analysis of the microarray signals, seem to be severely limited in unveiling the mechanisms governing the transcriptional cascade. Although proving their effectiveness e.g., in identifying expression signatures for cancer diagnosis, most computational tools have fallen short of representing a systematic method for understanding how the transcriptional regulation process takes place.

* These authors equally contributed.

Bioinformatics and computational biology need to overcome this limitation and develop approaches to identify regulatory networks, through the integration of multiple types of data. These computational methods should help deciphering how the transcribed elements of genomes impact the molecular mechanisms of functional utilization.

Network analysis has emerged as a powerful approach to understand complex phenomena and organization in social, technological and biological systems [2-4]. In particular, it is increasingly recognized the role played by the topology of cellular networks, the intricate web of interactions among genes, proteins and other molecules regulating cell activity, in unveiling the function and the evolution of living organisms [5-9]. Gene networks, in this respect, present a unique opportunity to employ this new type of approach [10-11]. In general, transcriptional regulatory networks are structures where the nodes are the genes and the edges are the interactions. A gene can be considered the source of a direct regulatory edge if it encodes a molecule with known regulatory function. Algorithms to infer the structure of gene-gene relationships take as primary input the data from a set of microarrays measuring the mRNA expression levels in different physiological states and use either classical statistics (e.g., Pearson correlation), concepts from the information theory (i.e., the mutual information as in ARACNe (Algorithm for the Reconstruction of Accurate Cellular Networks [12]) and CLR algorithms [13]) or probabilistic models (as in the Bayesian networks [14]) to reconstruct the network of transcriptional interactions. Once reconstructed, a gene regulatory network can be inspected and analyzed at different levels, ranging from the single gene to groups of genes (sub-networks) and to the whole network. Putative regulatory targets of a gene of interest can be searched at the single gene level with the goal of identifying previously unknown targets that can be the focus of subsequent experimental validation. At a higher level, the presence of groups of genes organized in sub-networks may suggest novel interactions and shed light on regulatory modules involving these genes or their common targets. Nevertheless, both strategies rely on prior knowledge to select the genes of interest, thus hampering the capacity of extracting *de-novo* knowledge from the network. To overcome this limitation, the inference algorithms used to reconstruct gene regulatory networks are often supplemented with methods derived from the network theory, i.e. borrowing analytical schemas typically used to explore communication or infrastructure networks. For instance, network theory recently focused on the resilience of complex networks to the malfunctioning of its component and to external disturbances. A key aspect of this analysis is the identification of the most *critical components* of the networks, i.e. those nodes/edges that are really crucial for the functioning of the system [15]. Specifically, the importance of an element is assessed considering the drop in the network performance caused by its deactivation and is quantified calculating the performance of the perturbed network as compared with the original one. The very same approach can be adopted to dissect gene regulatory networks for the identification of critical genes and interactions.

Here, the analysis of the network *critical components* was applied on the nodes of transcriptional and post-transcriptional regulatory networks reconstructed from mRNA and miRNA expression data of multiple myeloma samples with the aim of i) identifying genes that are critical for the structure of the networks irrespectively of the number of their ingoing or outgoing links, i.e. of the fact that they are hubs or not, and ii) inferring novel regulatory relationships.

## 2. Materials and Methods

### 2.1. *mRNA and miRNA expression data*

Gene (mRNA) and miRNA expression data were obtained from multiple myeloma (MM) specimens. In details, the gene expression dataset (hereafter denoted as MM158GE) comprises 5 normal, 11 monoclonal gammopathies of unknown significance (MGUS), 133 MM, and 9 plasma cell leukemia (PCL) for a total of 158 samples [16]. Matched mRNA and miRNA expression data (hereafter denoted as MM40MGE) have been obtained for a subset of samples included in the MM158GE dataset. In particular, 40 MM samples representative of five translocation groups (9 TC1, 10 TC2, 9 TC3, 7 TC4 and 5 TC5) have been analyzed using both Affymetrix HG-U133A and Agilent Human miRNA Microarray V2 arrays. The Agilent Human miRNA Microarray V2 consists of 60-mer DNA probes synthesized in situ, which represent 723 human and 76 human viral miRNAs derived from the Sanger database v10.1. Gene expression signals have been quantified using RMA (*affy* Bioconductor package) and the

GeneAnnot custom Chip Definition Files [17]. Genes with low signal variability across samples were eliminated using an entropy-based filter. Briefly, given the expression levels $w_{g,}$ of gene $g$ in sample $t$ ($t=1,...,N$), the entropy of the expression distribution was defined as:

$$H_g = \sum_{1 \le t \le N} -p_{g|t} \log_2\left(p_{g|t}\right) \quad (1)$$

where

$$p_{t|g} = w_{g,t} \Big/ \sum_{1 \le t \le N} w_{g,t} \quad (2)$$

is the relative expression of gene $g$ in sample $t$. The entropy threshold was selected in order to eliminate the 20% of less variable data [18]. Expression signals of the miRNAs arrays have been normalized using *aroma.light* Bioconductor package.

### 2.2. *Network critical components analysis*

The topological structure of a network can be used to identify the components (nodes or links) that are critical for the functioning of the system (*critical components*). *Network critical components analysis* has been successfully applied in different fields as communication or transportation. For instance, critical analysis is used to identify nodes that must be protected from terrorist attacks in communication networks, in social networks finding critical nodes can be fundamental to reduce the spreading of viruses, and in biological systems, this analysis can be extremely helpful to understand complex phenomena and to find more powerful ways to defend the system from a disease. Nodes and links can be removed using various techniques and different networks exhibit different levels of resilience to such disturbances. Networks can be perturbed simulating the deletion of node/links chosen at random (*error* removal or *failure*) or targeting a specific class of nodes/links (removal through intentional *attacks*). Attacks can be addressed sorting and removing progressively the nodes in descending order of degree or betweenness or the links in descending order of betweenness or range [19-21]. The network robustness is usually measured by the size of the largest connected component and by the average node-node distance as a function of the percentage of nodes/links removed.

The method used here to identify the critical components of gene regulatory networks is based on an *ad-hoc* definition of network performance, rather than on local node information such as the number of ingoing or outgoing links. Specifically, the importance of a node is measured by the drop in the network efficiency caused by the removal of that node, where the network efficiency *E(G)* quantifies how efficiently the nodes of the network exchange information [22]. The definition of *E(G)* requires recalling some formalism from the graph theory.

A network can be modeled by a graph $G$ of nodes that are tied by one or more specific type of interdependency. Formally, an *undirected* graph $G=(N, L)$ consists of two sets $N$ and $L$ such that $N \neq \varnothing$ and $L$ is a set of unordered pairs of element of $N$. The elements of $N=\{n_1,n_2,...n_M\}$ are the nodes of the graph G while the elements of $L=\{l_1,l_2,...,l_K\}$ are the edges. Two nodes joined by an edge are referred to as *adjacent* or *neighboring*. A graph is *weighted* when there exists a function $w : L \to \Re$ from edges to real numbers, such that each edge has associated a number that represents the strength of the connection. A graph is called *m-partite* if $N$ admits a partition into $m$ classes such that every edge has its ends in different classes: vertices in the same partition class must not be adjacent. When $m$=2, the graph is called *bipartite*. A *walk* from node $i$ to node $j$ is an alternating sequence of nodes and edges that begins with $i$ and ends with $j$. If no node is visited more than once, the walk is called a *path*. A graph $G$ is said to be *connected* if, for every pair of distinct nodes $i$ and $j$, there is a path from $i$ to $j$ in $G$. The degree or connectivity $k_i$ of a node $i$ is the number of edges incident with the node, i.e. the number of neighbors of that node. One of the most relevant topological characterizations of a graph $G$ can be obtained from the degree distribution $P(k)$, which is normally represented plotting the number of nodes having degree of connectivity $k$ against $k$ in a log-log scale. A decreasing linear dependency in this plot indicates that the network has a *scale-free* structure, associated with a corresponding power-law $n(k) \propto k^{-\gamma}$ (Figure 1). Graphs can be further classified as *assortative* if $k_{nn}(k)$, i.e., the average degree of the neighbors of degree $k$, is an increasing function of $k$; otherwise they are referred to as

*disassortative*. In *assortative* networks the nodes tend to connect to their connectivity peers, while in *disassortative* networks nodes with low degree are more likely connected with highly connected ones.

The efficiency of G relies on the calculation of the shortest path lengths $d_{ij}$ between two generic nodes *i* and *j*. In a weighted graph $d_{ij}$ is defined as the smallest sum of the physical distances throughout all the possible paths in the graph from *i* to *j*, while in an un-weighted graph $d_{ij}$ reduces to the minimum number of edges traversed to get from *i* to *j*. The maximum value of $d_{ij}$ is called the *diameter* of the graph and the average shortest path length *L* is quantified as follows:

$$L = \frac{1}{M(M-1)} \sum_{i,j \in N, i \neq j} d_{ij} \tag{3}$$

Supposing that every node sends information along the network, through its links, the efficiency $\varepsilon_{ij}$ in the communication between node *i* and node *j* is assumed to be inversely proportional to their shortest distance, i.e. $\varepsilon_{ij} = 1/d_{ij} \ \forall i, j$. It's worthwhile noting that the assumption that efficiency and distance are inversely proportional is a reasonable approximation although sometimes other relationships might be used, especially if justified by a more specific knowledge of the system. By assuming $\varepsilon_{ij} = 1/d_{ij}$, when there is no path in the graph between *i* and *j*, $d_{ij} = +\infty$ and consistently $\varepsilon_{ij} = 0$. Consequently, the average efficiency *E(G)* of the graph *G* can be defined as:

$$E(G) = \frac{1}{M(M-1)} \sum_{i,j \in N, i \neq j} \varepsilon_{ij} = \frac{1}{M(M-1)} \sum_{i,j \in N, i \neq j} \frac{1}{d_{ij}} \tag{4}$$

The definition of *E(G)* according to Eq.(4) avoids the divergence of *L* in case of disconnected components thus allowing the analysis of the entire network and not only of the biggest connected sub-graph. Since *E(G)* varies in the range $[0,\infty]$, it would be more practical to normalize *E(G)* in the interval [0,1]. The most natural way to normalize *E(G)* is with respect to the efficiency of a network $G^{ideal}$ composed of all the $M(M-1)/2$ possible edges:

$$E_{glob} = \frac{E(G)}{E(G^{ideal})} \tag{5}$$

Though the maximum value *E(G)*=1 is reached only when there is a link between each pair of nodes, real networks can nevertheless assume high values of E. This definition is valid for both un-weighted and weighted graphs and can also be applied to disconnected graphs.

The efficiency can be evaluated on any *sub-graph G'=(N',L')* of *G=(N,L)*, where *G'* of *G* is a graph such that $N' \subseteq N$ and $L' \subseteq L$. The sub-graph of the neighbors of a given node *i*, denoted as $G_i$, is the sub-graph induced by $N_i$, i.e., the set of nodes adjacent to *i*. Given $c_i$ the node cardinality of $G_i$, the local efficiency $E_{loc}$ is defined as the average of the sub-graph efficiencies $E(G_i)$ normalized with respect to the ideal sub-graphs in which all the $c_i(c_i-1)/2$ edges are present:

$$E_{loc} = \frac{1}{M} \sum_{i \in G} \frac{E(G_i)}{E(G_i^{ideal})} \tag{6}$$

Since $i \notin G_i$, the local efficiency $E_{loc}$ quantifies the efficiency of the system in tolerating faults, i.e., how efficient is the communication between the first neighbors of *i* when *i* is removed. Graphs that have high value of $E_{glob}$ and $E_{loc}$, i.e., that are very efficient both in their global and local communication, are defined as *small-words networks*.

Given the definition of *E(G)* and assuming that the efficiency is an appropriate quantity to characterize the average properties of a network, critical components can be identified considering the efficiency drop, caused by the deactivation of a component, as a measure of the centrality of that component. Therefore, the topological importance of a node $\alpha$ in a graph is quantified by the *network relevance* $r_\alpha$:

$$r_\alpha = \frac{E(G) - E(G_\alpha)}{E(G)} = \frac{\Delta E_\alpha}{E} \tag{7}$$

where $G_\alpha$ is the graph obtained by removing node $\alpha$ from G, for each $\alpha = 1,...,M$. The most critical nodes are those whose removal causes the largest drop in efficiency, i.e., those with the highest $r_\alpha$ (Figure 1). Although here the focus is on the determination of the critical nodes, the method is of general applicability to any subset (nodes, links and combination of nodes and links) of $G$ [23].

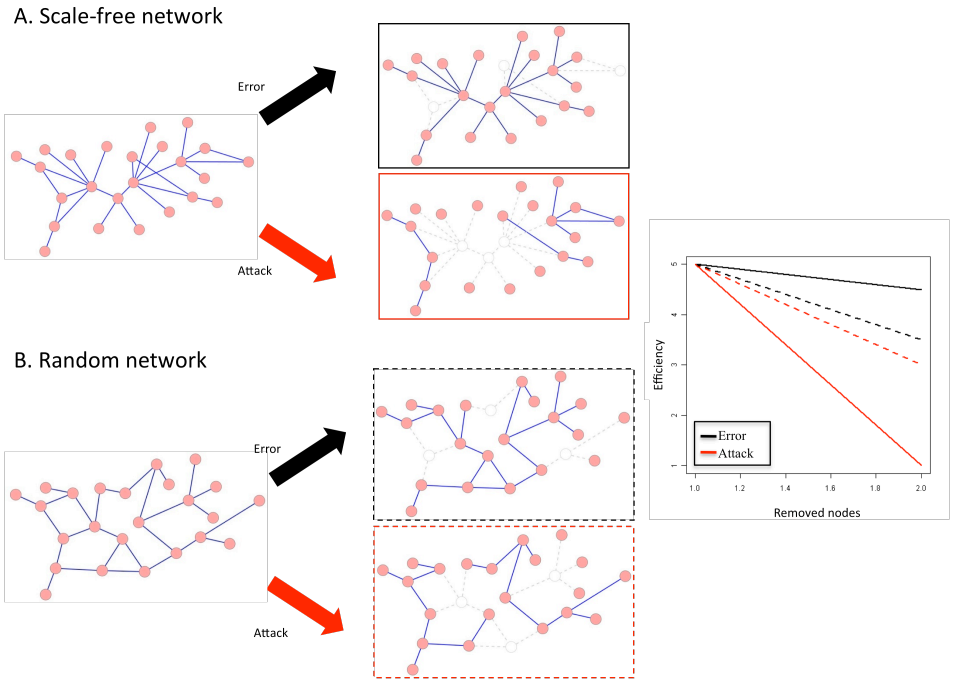


Figure 1. Effect of node removal through *error* and *attack* in scale free (A., solid lines) and in random (B., dotted lines) networks. In a scale-free network (solid lines), the random removal (*error*) of even a large fraction of vertices impacts the overall connectedness of the network very little (black line), while targeted *attack* (red line) destroys the connectedness very quickly, causing a rapid drop in efficiency. On the contrary, in random graphs, removal of nodes through either *error* or *attack* has the same effect on the network performance.

### 2.3. *Transcriptional regulatory network*

The transcriptional regulatory network was reconstructed using ARACNe and the MM158GE dataset of gene expression signals. ARACNe utilizes information and data transmission concepts (i.e., mutual information and data processing inequality) to identify statistically significant co-regulations among genes from microarray expression profiles. Mutual information and data processing inequality allow reconstructing gene-gene relationships which most likely represent either direct regulatory interactions or interactions mediated by post-transcriptional modifiers. Briefly, the algorithm first uses the expression data to calculate pair wise Mutual Information (MI) through a computationally efficient Gaussian kernel estimator. ARACNe calculates the kernel width depending on the size and statistics of the dataset. The second step is the elimination of the interactions that are not statistically significant according to a p-value or a MI threshold and returns a series of irreducible statistical dependencies. The post-processing step eliminates interactions that are likely to be indirect. The Data Processing Inequality (DPI) theorem removes indirect regulatory influences that appear as direct because of a high MI score due to the presence of a common neighbor. An additional parameter, called DPI tolerance, can be used to compensate for errors in the MI estimate that might affect DPI application [12].

The parameters of the kernel width and the Mutual Information threshold were calculated using MATLAB scripts. The p-value to determine the MI threshold was set at $1e^{-7}$, while the DPI tolerance was set equal to 10%. A list of Transcription Factors (TF) for the platform HG-U133A was also imputed as a parameter to prevent the DPI from removing transcriptional interactions in favor of non-transcriptional ones (interactions between two non-TFs).

### 2.4. *Post-transcriptional regulatory network*

The post-transcriptional regulatory network was reconstructed calculating the Pearson correlation coefficient of the expression vectors of miRNA target genes in the MM40MGE dataset. In details, the procedure required to i) identify the miRNA target genes basing on computational predictions; ii) select the target relationships supported by miRNA and mRNA expression data; iii) compute the Pearson correlation coefficient of the expression levels of target genes sharing at least one supported miRNA-target relationship; iv) reconstruct the post-transcriptional regulatory network from the adjacency matrix **S** of regulatory relations supported by miRNA and mRNA expression levels. Computational prediction of miRNA targets presents significant challenges due to the lack of a sufficiently large group of known miRNA targets to be used as training set. As such, most computational algorithms for target prediction (miRanda, TargetScan, PicTar, PITA, RNAhybrid) result in a significant proportion of false positives, i.e. in the prediction of not-functional miRNA-mRNA interactions. Given the increasing experimental evidences supporting the miRNA mechanism of target degradation, the integration of *in-silico* predictions with miRNA and target gene expression profiles has been proposed as a method to select functional miRNA-mRNA relationships. Since miRNAs tend to down-regulate target mRNAs, the expression profiles of genuinely interacting pairs are expected to be anti-correlated. This integrative analysis can be performed using a variational Bayesian model [24] or, as in this case, through a non-heuristic methodology based on the anti-correlation between miRNA and mRNA matched expression profiles [25-26].

Specifically, miRanda algorithm [27] was applied to predict miRNA targets from the human miRNA sequences and transcripts of miRBase Release 12.0 and ENSEMBL Release 52, respectively. Targeting predictions were retained if the miRanda score was higher than 160. The Pearson correlation coefficient of expression vectors was calculated for each miRNA-gene pair scored as potentially interacting according to the prediction of miRanda and used as an estimator of the functional activity of miRNAs on predicted target genes. Genes were considered genuine miRNA targets only if included within the top 3% of all anti-correlated pairs [25]. This selection gave rise to a final adjacency matrix **S** of regulatory relations supported by expression levels. The adjacency matrix **S** defined a bipartite directed network with two types of nodes (miRNAs and mRNAs) connected by directed edges, each representing a probably functional regulatory effect of a miRNA on a target gene. The same matrix **S** was used to derive a gene-only network in which genes (nodes) are connected by undirected weighted links and the edge weight quantifies the number of shared miRNAs regulating each gene pair.

## 3. Results

ARACNe inferred a transcriptional network with 9666 nodes (i.e. genes) and 86846 edges (i.e. interactions) from the MM158GE dataset. The topological characteristics of the network are reported in Table 1, in terms of number of nodes, number of edges, maximum $k_{max}$ and average $k_{mean}$ connectivity ($k$ being the degree of a node, i.e. the number of its interactions), diameter (representing the maximum value of $d_{ij}$) and global and local efficiencies.

Table 1. Metrics of transcriptional and post-transcriptional networks.

| Network type | Nodes (M) | Edges (K) | $k_{max}$ | $k_{mean}$ | Diameter | $E_{glob}$ | $E_{loc}$ |
|---|---|---|---|---|---|---|---|
| Transcriptional | 9666 | 86846 | 219 | 17.96 | 8 | 0.279 | 0.150 |
| Post-transcriptional | 6435 | 909324 | 1811 | 282.62 | 8 | 0.611 | 0.866 |

The connectivity distribution shows a power-law tail suggesting that the underlying structure of the network is scale-free (Figure 2A). At low connectivity values (k<11), the degree distribution loses its linear progression probably as a consequence of the limited number of genes. The relationship between the average connectivity $k_{nn}$ of the neighbors of a node and the node connectivity suggests an assortative behavior of the network, i.e., the nodes tend to connect with nodes with a similar connectivity thus partly implying a hierarchical structure of the network (Figures 2B). Ranking the nodes according to their connectivity allowed indentifying 27 *hubs*, i.e. genes with more than 100 interactions (data not shown).

The miRNA-mRNA integrated analysis resulted in a post-transcriptional gene network with 6435 nodes (genes) and 909324 weighted edges. The network was reconstructed first refining the predicted targeting relationships of

MiRanda through the selection of those predictions more supported by miRNA-mRNA expression data (the 3% most highly anti-correlated miRNA-gene pairs). This corresponded to 23729 regulatory relations involving 692 miRNAs and 6,435 target genes. It's worth noting that about 48% of genes associated to an expression profile resulted not to be real target of any considered miRNA and 9 miRNAs were not detected as sufficiently active on any target gene. Then, the remaining 692 miRNAs and 6,435 target genes were employed to reconstruct a bipartite directed miRNAs-mRNAs regulatory network, representing the probably functional regulatory effects of all these miRNA to their targets in MM. The number of target genes per miRNA ranges from 1 to 440 (average 33.3 with a mean value of 3.7 miRNAs per gene). Finally, a weighted post-transcriptional network of 6435 genes was extracted from the bipartite miRNA-mRNA regulatory network with the weight of an edge representing the number of functional interactions with microRNAs shared by the couple of connected genes. The topological characteristics of the network are reported in Table 1. Similarly to the transcriptional network, the connectivity distribution and the relationship between average connectivity $k_{nn}$ of the neighbors of a node and the node connectivity suggest a scale-free, assortative structure (Figures 2C and 2D).

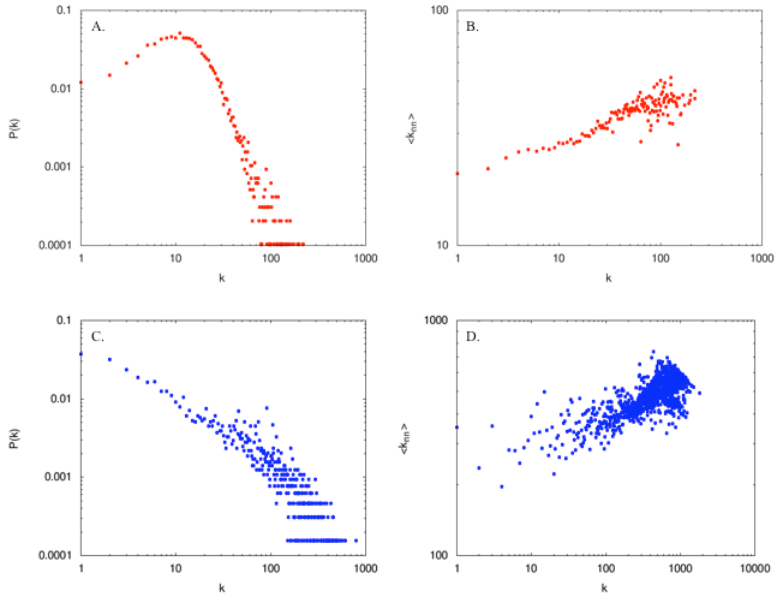


Figure 2. Connectivity properties of transcriptional (A. and B.) and post-transcriptional (C. and D.) networks. A. Connectivity distribution *P(k)* of nodes with a specific number of incident edges (degree of connectivity *k*) in the transcriptional network. The connectivity distribution shows a power-law tail suggesting a scale-free structure of the network. B. Relationship between the average connectivity $k_{nn}$ of the neighbors of a node and the node connectivity *k*. The linear trend suggests an assortative behavior of the transcriptional network. C. Same as A. for the post-transcriptional network. D. Same as B. for the post-transcriptional network.

The critical nodes of both the transcriptional and post-transcriptional regulatory networks have been determined by the static analysis of *error* and *attack* tolerance. The drop in the network efficiency caused by the node removal (i.e., the node relevance $r_\alpha$ as defined in Eq. (7)) has been used as the criteria to determine the importance of a node. The critical analysis has been applied to the transcriptional network, to the post-transcriptional one, and, for testing the robustness of results, to a random graph [28]. The random network has been constructed starting from an initial condition of *M* nodes and no edges and then adding *K* edges between pairs of randomly selected nodes, where *M* and *K* were the same as in the transcriptional network. Figures 3A and 2B show the global efficiency for the transcriptional scale-free network and for the random graph (both with *M=9666* nodes and *K=86846* edges) as functions of the number of removed nodes through efficiency-based attacks (i.e., attacks performed removing nodes with the highest efficiency; red line) and random removals (errors; black line). The true transcriptional network shows a different behaviour with respect to attacks and errors (Figure 3A). The removal of ~30% of nodes in a

targeted way (*attack*) reduces the network efficiency to about half the initial value and removing ~60% of the nodes destroys completely the system. Instead, when removing nodes randomly (*error*), the drop of the network global efficiency shows a linear dependency with the number of removed nodes and even for high value of removals (>60%) the system maintains a considerable efficiency (Figure 3C). The fact that removing specific nodes causes a rapid drop in the capability of the system to communicate further supports the scale-free structure of the regulatory graphs and proves the existence of a discrete number of *critical components*, i.e. of nodes responsible for the specific structure of the network. As far as the random graph is concerned (Figures 3B and 3C), differences of tolerance to *attacks* and to *errors* are much less pronounced. In this case, in fact, there is no substantial variability in the efficiency and the removal of a node in a targeted or in a random way produces similar, though not equal, behaviours.

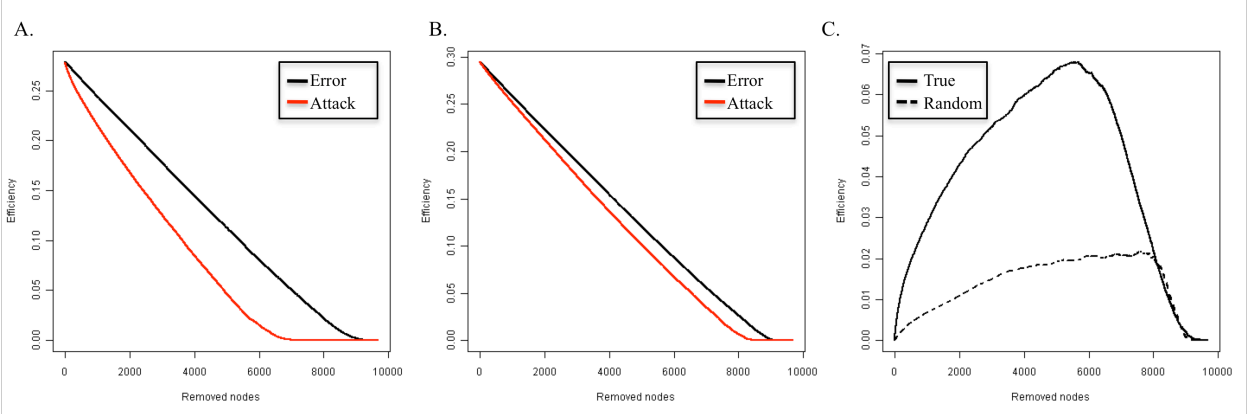


Figure 3. Global efficiency E(G) as a function of the number of removed nodes. In both the cases, the graphs were composed of M=9666 nodes and K=86846 edges and the node removal simulated by errors (black line) and efficiency-based attacks (red line). A. Transcriptional network generated by ARACNe using gene expression data from the MM158GE dataset. B, Random graph. C. Difference of tolerance to attacks and to errors (i.e., difference between the drop in efficiency caused by efficiency-based attacks and error node removals) for true transcriptional (unbroken line) and the randomly generated networks (broken line).

The analysis of critical components revealed that, in the transcriptional and post-transcriptional networks, critical nodes are not limited to hub genes and that also genes with a limited number of connections can be *critical* for the structure of the network. Figures 4A, 4B and 4C report the comparison between the node rankings calculated according to node degree ($k$) and node *criticality* ($r_\alpha$) in the random, transcriptional, and post-transcriptional networks.

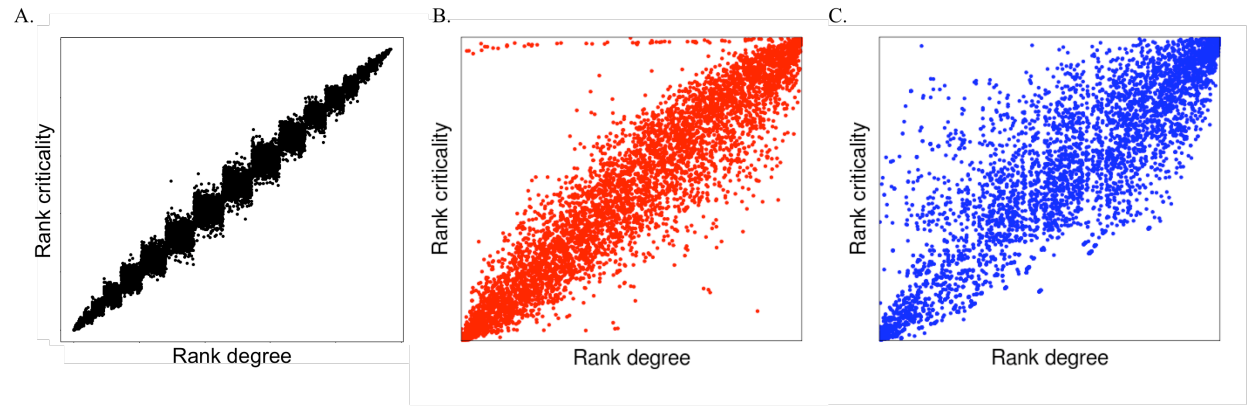


Figure 4. Comparison between the rank according to node degree k (rank degree) and the rank given by criticality $r_\alpha$ (rank criticality) for the nodes of A. random, B. transcriptional, and C. post-transcriptional networks.

As expected, this comparison indicates that i) in random network there is no significant difference between node degree and node criticality (Figure 4A) and ii) the vast majority of hub genes are critical nodes, i.e., the nodes whose removal causes a large drop in global efficiency correspond to the most connected genes (Figures 4B and 4C). Nevertheless, there exist a fraction of nodes that are characterized by a higher criticality value than expectable according to their degree. These nodes, even if characterized by a low node degree, are indeed critical and would have been disregarded as putative regulatory targets due to their limited number of connections. For instance, in the MM transcriptional network, the B cell linker gene (*BLNK*) emerged as one of the most critical genes although being connected to only 34 other nodes (as compared to hubs characterized by >100 links). *BLNK* is known to be involved in normal B-cell development and deficiency in this protein has been shown in some cases of pre-B acute lymphoblastic leukemia, suggesting its putative role as a tumor suppressor gene. A search in the network for putative targets of *BLNK* allowed identifying 8 genes that, once ranked according to the mutual information, indicated *CDKN1B* (cyclin dependent kinase inhibitor, p27 kip1) as the most strongly connected target. This interaction was confirmed by the same analysis conducted on another MM dataset ([29]; data not shown) and by recent experimental evidences, which reported p27 kip1 induction by *BLNK* trough *JAK3* [30]. Unfortunately, *JAK3* was not present in the datasets used for the network reconstruction and thus this evidence could not be further confirmed.

To integrate the results of the critical analyses, a list of 5145 non redundant genes/nodes represented in both transcriptional and post-transcriptional networks was compiled and used to select, in each network, the top 1% critical nodes and the top 1% most connected nodes (hubs). The intersections of such lists could help clarifying the role of nodes, which are critical in terms of network efficiency, although being not highly connected (Figure 5).

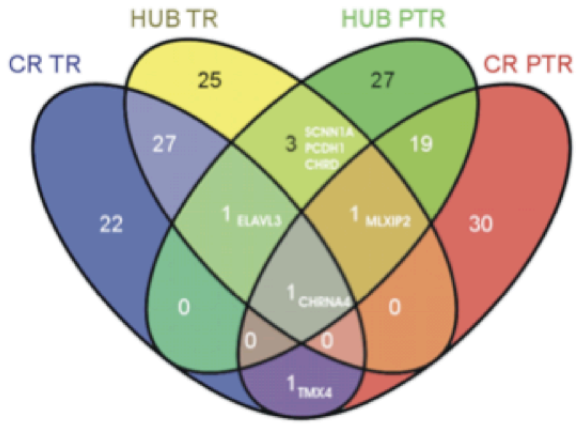


Figure 5. Venn diagram showing the intersections of four sets: CR TR (top 1% critical nodes in the transcriptional regulatory network), HUB TR (1% most connected nodes, hubs, in the transcriptional regulatory network), CR PTR (top 1% critical nodes in the post-transcriptional regulatory network) and HUB PTR (1% most connected nodes, in the post-transcriptional regulatory network).

In particular, when comparing the lists of top 1% critical genes for the two networks, 2 genes emerged as critical for both transcriptional and post-transcriptional regulation, i.e. *CHRNA4* (cholinergic receptor, nicotinic, alpha 4) and *TMX4* (thioredoxin-related transmembrane protein 4), both known to be expressed in B cells and the latter involved in B cells activation and in cancer. Interestingly, *CHRNA4* shares, as post-transcriptional regulators supported by expression data, two miRNAs (i.e., hsa-miR-15a* and hsa-miR-30c-1*) that regulate also *ELAVL3* and *MLXIPL*, two genes which are hubs and critical in the transcriptional and post-transcriptional network, respectively. Both miR-15a and miR-30c are differentially expressed in peripheral blood cells [31] and the presence of a *MYB*-miR-15a auto-regulatory feedback loop is of potential importance in human hematopoiesis [32]. In particular, miR-15a expression inversely correlates with *MYB* expression in cells undergoing erythroid differentiation and the over-expression of miR-15a blocks both erythroid and myeloid colony formation *in-vitro*.

## 4. Discussion

Reconstructing regulatory network from expression data is a crucial step to understand the mechanisms underlying biological systems. However, the high number of genes and interactions still represents a challenging issue for the extraction of relevant targets and relationships from such large systems. A standard approach is searching targets among the most connected genes (hubs) or among sub groups of genes known to be relevant in the analyzed phenotype. The goal of this type of analysis is to identify previously unknown relationships that can be the object of a subsequent experimental validation. An alternative approach is studying the network characteristics to identify groups of genes organized in sub-networks, which may suggest novel interactions and shed light on regulatory modules involving these genes or their common targets. Although effective, both strategies rely on prior knowledge to select the genes of interest, thus hampering the capacity of extracting *de-novo* knowledge from the network. A way to overcome this limitation could be adapting techniques commonly used in the analysis of communication and infrastructure networks. In these fields, a key analysis is the resilience of the network to external disturbances and to malfunctioning. Network robustness strongly relies on the network structure and, in particular, on the existence of paths between the nodes. When nodes or links are removed, the lengths of these paths can increase and some nodes will become disconnected. It is therefore interesting to find the critical component of the network, i.e. the nodes or edges that are really important for the functioning of the network. In Latora and Marchiori, the authors proposed a method to evaluate the importance of a network element (that can be a node or an edge) by considering the drop in the network performance caused by its deactivation [15]. The performance of the perturbed network is compared with the original one. Iturria-Medina and colleagues applied a similar approach to investigate the human brain anatomical network [33].

Different criteria can be used to measure the performance, such as efficiency or mean flow rate of information. Efficiency measures how efficiently the nodes of the network exchange information. Applying this concept to regulatory network, critical nodes and edges are critical genes and critical regulatory interaction, respectively. Usually the most important nodes are considered the most connected ones (hubs), but this is not always the case [23]. In genetic networks a gene can be connected to many genes simply because is a transcription factor that normally controls many targets or a gene that is controlled by many other genes. For instance, in the analysis of the B cell networks, the largest hub with more than 300 interactions was a poorly characterized gene, *BYSL*. Instead, a much more interesting gene was *MYC* that, with only 56 neighbours, ranked 410$^{th}$ in terms of connectivity. *MYC*, a well-known proto-oncogene, had neighbours that were themselves genetic hubs (including *BYSL*), such that *MYC* could modulate a substantial percentage of all genes in the cell trough a relatively small number of neighbours [12].

Recently, some approaches exploited the topological features of large gene regulatory networks to identify individual components that are biologically relevant or to elucidate the role of each particular element in regulation. Patapov and co-workers introduced the pair-wise disconnectivity index to quantitatively evaluate the topological significance of each element (i.e., nodes and edges) in the context of all other elements of the regulatory network [34]. The application of this approach to the analysis of the TLR4 signal transduction network allowed identifying a number of key signalling and transcription regulators among the nodes top-ranking in terms of disconnectivity index. Differently, Emmert-Streib and Dehmer used the concept of functional robustness, originally introduced by Li et al. [35], to study the functional robustness of the transcriptional regulatory network in yeast [36]. The definition of an information theoretic measure to estimate the influence of single node perturbations on the global network topology allowed identifying nodes which are fragile with respect to single node knockouts and revealed significant differences between fragile nodes and hubs. Interestingly, the set of fragile nodes was statistically enriched in essential genes, i.e. in genes required to sustain vital yeast.

Here, the critical analysis of network components has been applied to inspect the transcriptional and post-transcriptional regulatory networks reconstructed from mRNA and miRNA expression data of multiple myeloma samples. The transcriptional and post-transcriptional networks were reconstructed using ARACNe and the Pearson correlation coefficient of the expression vectors of miRNA target genes, respectively. Both networks showed a scale free structure, i.e. a type of structure reported with evidence in lower organisms, but still argument of debate in eukaryotes. The connectivity plots of Figure 2 strengthen the hypothesis that the structure of human interaction

networks has a scale free nature with a saturation effect also reported for other scale-free networks, when the maximum connectivity range is below 1000 [12, 37-39]. Both networks are also slightly assortative, meaning that they tend to have an aristocratic behaviour where nodes with high degree tend to connect with nodes with similar degree. This suggests a hierarchical control mechanism, as also reported in [12]. The analysis of critical components revealed that genes with a limited number of connections could be critical for the structure of the network and that hubs are not necessarily critical nodes. Indeed, about one half of most connected nodes in each considered network were not included in the corresponding list of most critical nodes and genes like *BLNK*, characterized by a low node degree, were instead critical. These *non-hub* critical nodes would have been disregarded as putative regulatory targets due to their limited number of connections although they may provide clues to the detection of key regulatory circuits. Finally, the integration of the transcriptional and post-transcriptional levels allowed identifying critical genes for both types of regulatory interactions and dissecting direct critical relationships at transcriptional level from interaction that are instead indirect since mediated by post-transcriptional regulation.

## 5. Acknowledgments

This work was supported by grants from Fondazione CARIPARO (Progetti Eccellenza 2006); MIUR (PRIN 2007Y84HTJ and PRIN 2007CHSMEB); University of Padova (CPDA065788/06 and CPDR074285/07); University of Modena (Finanziamento Linee Strategiche di Sviluppo dell'Ateneo, Medicina Molecolare e Rigenerativa, 2008); Fondazione Cassa di Risparmio di Modena (Bando ricerca 2007), and Associazione Italiana Ricerca sul Cancro (AIRC).

## 6. References

1. T. Ideker, *Nat Biotechnol*. **22**(4), 473 (2004)
2. S. H. Strogatz, *Nature.* **410**, 268 (2001)
3. S. Wasserman and K. Faust, *Social Networks Analysis*, Cambridge University Press, Cambridge (1994)
4. S.N. Dorogovtesev and J. F. F. Mendes, *Evolution of networks*, Oxford University Press, (2003)
5. H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai ZN and A. L. Barabasi, *Nature*. **407**(6804), 651 (2000)
6. H. Jeong, S. P. Mason, A. L. Barabasi and Z. N. Oltvai, *Nature*. **411**(6833), 41 (2001)
7. S. A. Wagner and D. A. Fell, *Proc. R. Soc. London*. **B268**, 1803 (2001)
8. S. Maslov and K. Sneppen, *Science*. **296**(5569), 910 (2002)
9. R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii and U. Alon, *Science*. **298**(5594), 824 (2002)
10. A. Vazquez, R. Dobrin, D. Sergi, J. P. Eckmann, Z. N. Oltvai, A. L. Barabási, *Proc Natl Acad Sci U S A*. **101**(52), 17940 (2004)
11. R. Sharan and T. Ideker, *Nat Biotechnol*. **24**(4), 427 (2006)
12. K. Basso, A. A. Margolin, G. Stolovitzky, U. Klein, R. Dalla-Favera and A. Califano, *Nat Genet*. **37**(4), 382 (2005)
13. J. J. Faith, B. Hayete, J. T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J. J. Collins and T. S. Gardner, *PLoS Biology*. **5**(1) (2007).
14. J. Pearl, *Probabilistic reasoning in intelligent system: networks of plausible inference*, San Francisco, CA, Morgan Kaufmann Publishers, Inc. (1988)
15. V. Latora and M. Marchiori, *Physical Review E Statistical, Nonlinear and Soft Matter Physics*. **71**(1Pt2), 015103 (2005)
16. L. Agnelli, S. Bicciato, M. Mattioli, S. Fabris, D. Intini, D. Verdelli, L. Baldini, F. Morabito, V. Callea, L. Lombardi and A. Neri, *Journal of Clinical Oncology*. **23**(29), 7296 (2005)
17. F. Ferrari, S. Bortoluzzi, A. Coppe, A. Sirota, M. Safran, M. Shmoish, S. Ferrari, D. Lancet, G. A. Danieli and S. Bicciato, *BMC Bioinformatics*. **8**, 446 (2007)
18. J. Schug, W. P. Schuller, C. Kappen, J. M. Salbaum, M. Bucan and C. J. Stoeckert Jr, *Genome Biol*. **6**(4), R33 (2005)
19. P. Holme, B. J. Kim, C. N. Yoon and S. K. Han, *Phys Rev E*. **65**, 056109 (2002)
20. R. Albert, I. Albert, and G. L. Nakarado, *Phys. Rev. E*. **69**, 025103 (2004)
21. A.E. Motter and Y. Lai, *Phys. Rev. E*. **66**, 065102 (2002)
22. V. Latora and M. Marchiori, *Phys. Rev. Lett.* **87**, 198701 (2001)
23. V. Latora and M. Marchiori, *Chaos, Solitons and Fractals*. **20**, 69 (2004)

24. J. C. Huang, Q. D. Morris and B. J. Frey, *J Comput Biol*. **14**(5), 550 (2007)
25. V. A. Gennarino, M. Sardiello, R. Avellino, N. Meola, V. Maselli, S. Anand, L. Cutillo, A. Ballabio and S. Banfi, *Genome Res*. **19**(3), 481 (2009)
26. F. Xin, M. Li, C. Balch, M. Thomson, M. Fan, Y. Liu, S. M. Hammond, S. Kim and K. P. Nephew, *Bioinformatics*. **25**(4), 430 (2009)
27. B. John, A. J. Enright, A. Aravin, T. Tuschl, C. Sander and D. S. Marks, *PLoS Biol*. **2**(11), e363 (2004)
28. P. Crucitti, V. Latora, M. Marchiori and A. Rapisarda, *Physica A*. **340**, 388 (2004)
29. W. J. Chng, S. Kumar, S. Vanwier, G. Ahmann, T. Price-Troska, K. Henderson, T. H. Chung, S. Kim, G. Mulligan, B. Bryant, J. Carpten, M. Gertz, S. V. Rajkumar, M. Lacy, A. Dispenzieri, R. Kyle, P. Greipp, P. L. Bergsagel and R. Fonseca, *Cancer Research*. **67**(7), 2982 (2007)
30. J. Nakayama, M. Yamamoto, K. Hayashi, H. Satoh, K. Bundo, M. Kubo, R. Goitsuka, M. A. Farrar and D. Kitamura, *Blood*. **113**(7), 1483 (2009)
31. M. Merkerova, M. Belickova and H. Bruchova, *H. Eur J Haematol*. **81**(4), 304 (2008)
32. H. Zhao, A. Kalota, S. Jin and A. M. Gewirtz, *Blood*. **113**(3), 505 (2009)
33. Y. Iturria-Medina, R. C. Sotero, E. J. Canales-Rodríguez, Y. Alemán-Gómez and L. Melie-García, *Neuroimage*. **40**(3), 1064 (2008)
34. A. P. Potapov, B. Goemann, E. Wingender, *BMC Bioinformatics*. **9**:227 (2008)
35. F. Li, T. Long, Y. Lu, Q. Ouyang, C. Tang, *Proc Natl Acad Sci U S A*. **101**(14), 4781 (2004)
36. F. Emmert-Streib, M. Dehmer, *BMC Systems Biology*. **3**, 35 (2009)
37. A. L. Barabasi and R. Albert, *Science*. **286**(5439), 509 (1999)
38. R. Albert, *J Cell Sci*. **118**, 4947 (2005)
39. E. Almaas, *J Exp Biol*. **210**(Pt 9), 1548 (2007)