NETWORKS AND HETEROGENEOUS MEDIA
©American Institute of Mathematical Sciences
Volume **7**, Number **3**, September **2012**

doi:10.3934/nhm.2012.7.473

pp. **473–481**

# PREPROCESSING AND ANALYZING GENETIC DATA WITH COMPLEX NETWORKS: AN APPLICATION TO OBSTRUCTIVE NEPHROPATHY

Massimiliano Zanin

Faculdade de Ciências e Tecnologia, Departamento de Engenharia Electrotécnica
Universidade Nova de Lisboa

Centre for Biomedical Technology
Technical University of Madrid
Pozuelo de Alarcón, 28223 Madrid, Spain

Innaxis Foundation & Research Institute
José Ortega y Gasset 20, 28006, Madrid, Spain

Ernestina Menasalvas

Faculty of Computer Science
Technical University of Madrid
Pozuelo de Alarcón, 28223 Madrid, Spain

Pedro A. C. Sousa

Departamento de Engenharia Electrotcnica, Faculdade de Ciencias e
Tecnologia Universidade Nova de Lisboa
Quinta da Torre
2825 - 182 Caparica, Portugal

Stefano Boccaletti

Centre for Biomedical Technology
Technical University of Madrid
Pozuelo de Alarcón, 28223 Madrid, Spain

Abstract. Many diseases have a genetic origin, and a great effort is being made to detect the genes that are responsible for their insurgence. One of the most promising techniques is the analysis of genetic information through the use of complex networks theory. Yet, a practical problem of this approach is its computational cost, which scales as the square of the number of features included in the initial dataset. In this paper, we propose the use of an *iterative feature selection* strategy to identify reduced subsets of relevant features, and show an application to the analysis of congenital Obstructive Nephropathy. Results demonstrate that, besides achieving a drastic reduction of the computational cost, the topologies of the obtained networks still hold all the relevant information, and are thus able to fully characterize the severity of the disease.

2000 *Mathematics Subject Classification.* Primary: 62-07, 92D10; Secondary: 92B05.

*Key words and phrases.* Genetic data, complex networks, iterative feature selection, Obstructive Nephropathy.

## 1. Introduction.

The analysis of genetic information, and specifically of levels of genetic expressions [1], has been the center of a large number of studies in the last decades, as it allows a better understanding of the root causes beyond many diseases [2, 3, 4], as well of the cellular mechanisms responsible for the response of living systems to internal and external stimuli [5]. In the last years, a new paradigm has been proposed to address this problem, that is, the use of a network representation [6]. Specifically, the complex network approach [7, 8], has been successfully applied to a great variety of problems, where the importance of the interactions between the different elements composing the system is the same, or even greater, than that of each single element [9]; the interested reader can find a complete review in Refs. [7, 10].

Clearly, the study of how different genes interact can unveil new relevant knowledge, which usually cannot be gathered from the analysis of the behavior of individual and isolated genetic expressions. Following this principle, several works have analyzed co-expression networks, where nodes represent individual genes, and pairs of nodes are connected whenever there is a correlation in their expression [11]. Recently, a complementary strategy has been proposed, consisting in reconstructing connections between pairs of nodes when their expressions are outside the range observed in a reference condition. This generates networks that carry on information on the abnormal behaviors of genes' expressions [12]. Indeed, by applying this methodology to a set of control data (for instance, healthy subjects) and a set of subjects likely affected by a given disease, such a strategy allows to extract information on: i) whether a given subject is suffering from the studied disease, and ii) which elements (genes) of the graph are the main responsible for the pathological condition.

Still, this approach presents some practical drawbacks, the most relevant of which being the computational cost associated with the case of a large amount of genetic data, which results in the need of reconstructing very large networks. In this paper, we propose the application of a well-known data-mining tool, i.e. the iterative feature selection [13, 14], to select those genes that are genuinely relevant for the analysis. Besides reducing the quantity of information to be processed (and thus the computational cost), the elimination of *noisy* genes results in an overall improvement of the accuracy of the method. Moreover, reducing the number of genes also reduces the dimensionality of the problem, thus improving the statistical significance of the results. The proposed methodology is validated by performing the iterative feature selection on a dataset of genetic information, related to the problem of early diagnostic of the Obstructive Nephropathy, a congenital kidney disease and one of the most-important causes of renal insufficiency in children.

The remainder of the paper is structured as follows. Section 2 introduces how the networks are constructed from the data. Section 3 describes how the methodology is applied to the dataset under study, including control subjects and patients affected by Obstructive Nephropathy, and what is the relationship between genetic data and the severity of the disease. Section 4 describes the different techniques used for performing the iterative feature selection, whose results are presented in Section 5. Finally, conclusions and future perspectives are discussed in Section 6.

## 2. Construction of the networks.

As already introduced, the aim of this contribution is to improve the analysis of genetic information through complex networks, by applying a data mining approach for features selection. Therefore, the first step is the preprocessing of available data aimed at eventually reconstructing a network for each *target* subject, i.e., for each person whose state (healthy or ill) is not known in advance. For that purpose, it is previously necessary to process information about some control subjects, that is, persons that are known to be healthy, in order to extract the *normal* relations between each pair of genes.

The information corresponding to this first step is represented by $n$ different *features*, that is, genetic expression levels, measured for $m$ different control subjects. This information is organized in a matrix $D$ of size $n \times m$; the $i$-th feature of the $s$-th subject is represented as $d_i^s$. For each pair of features $i$ and $j$, a linear fit is calculated; therefore, for each subject $s$, the expression of gene $j$ (i.e., $d_j^s$) can be seen as a linear function of that of gene $i$, plus an error term:

$$d_j^s = a_{ij} + b_{ij} d_i^s + \varepsilon_{ij}^s. \tag{1}$$

In this equation, $a_{ij}$ and $b_{ij}$ are the two coefficients resulting from a linear fit of the values of $d_i$ against $d_j$, for all $m$ control subjects. $\varepsilon$ is a vector containing all the $m$ errors of the fit. Notice that, the mean values of this error vectors $(\overline{\varepsilon_{ij}})$ and their standard deviations $(\sigma_{ij})$ are relevant quantities that will be used for the reconstruction of the network. For the sake of exemplification, Fig. 1 schematically illustrates the process. Black circles are the fictitious expression levels for healthy persons, and the dashed blue line represents the linear fit - following the above convention, $a_{ij}$ is the intersection of the fit line with the vertical axis, and $b_{ij}$ its slope.

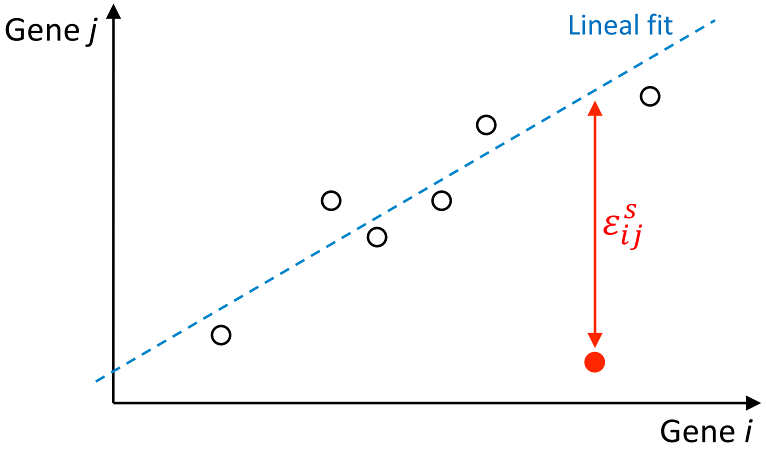

FIGURE 1. Example of the construction of a fictitious network. As a first step, the data corresponding to healthy subjects (black circles) are fitted to a linear function (blue dashed line); afterward, anomalous data of patients are identified by the distance from the linear fit (i.e., $\varepsilon_{ij}^s$ of Eq. 1, represented by the red vertical arrow).

The next step is the creation of a network for each target subject. We consider a dataset $T$, where each element $t_i^v$ corresponds to the $i$-th feature of subject $v$. For simplicity, we also consider that the features encoded in this second dataset $T$ are the same as those in $D$, and that all values are available (therefore, no invalid value is expected).

Now, let us consider the expression values $t_i^v$ and $t_j^v$ for a given subject $v$. In principle, it should be expected that these expressions are well described by the previously calculated corresponding linear fit. The error associated to such an approximation (from now on called $e_{ij}^v$) can be calculated as:

$$e_{ij}^v = a_{ij} + b_{ij} t_i^v - t_j^v. \tag{2}$$

$e$ represents how separated the value for the subject $v$ is with respect to the linear fit, as calculated for the healthy subjects. Notice that, from now on, $\varepsilon$ represents the fit error for control subjects (initially used for extracting the ground truth about the system), while $e$ is the error fit for target subjects (i.e., those under analysis). Going back to Fig. 1, this is represented by the vertical red arrow. We are interested in networks representing the abnormal relation between pairs of genetic expressions, e.g., those situations where the linear fit does not significantly represent the observation for the target subject, or, in other words, where $e_{ij}^t$ is greater than $\sigma_{ij}$. This abnormality is better represented by the absolute value of the Z-Score, defined as:

$$Z_{ij}^v = \frac{\left| e_{ij}^v - \overline{\varepsilon_{ij}} \right|}{\sigma_{ij}} \tag{3}$$

Using the above expression, it is possible to construct a network for each target person $v$, where nodes represent the $n$ different features, and the weight of the link connecting each pairs of nodes $i$ and $j$ is given by $Z_{ij}^v$; in other words, this network is a clique (a graph in which all pairs of nodes are connected by a link), where the weight of each possible link is codified in a weight matrix $W = Z$.

To further simplify, each clique can be transformed in a unweighted graph by applying a threshold $\tau$; consequently, the associated adjacency matrix $A$ is defined as:

$$\begin{cases} 1 & if \quad Z_{ij}^v > \tau \\ 0 & if \quad Z_{ij}^v \leq \tau \end{cases} \tag{4}$$

In summary, our method is based on the creation of networks, one for each subject, where nodes represent genes, and two nodes are connected when the relation between them is outside the range expected in control subjects (more precisely, when the separation from the expected value is at least $\tau$ times larger than the standard deviation found in control subjects).

The analysis of the structure (or topology) of the obtained networks may furnish relevant information. On the one side, if the data corresponds to a healthy person, we expect all pairs of values to be as close to the linear fit as the data from the control set: therefore, the network corresponding to this subject will have very few links, mostly due to noise in the measurement, and consequently a random topology. On the other side, persons suffering from a genetic-related disease will have abnormal values in some of these relations; the resulting topologies will be then easily identified by an abnormally high number of links, and by star-like structures, whose centers will point to the genes responsible for the disease [12].

## 3. Application to Obstructive Nephropathy.

We use the same dataset studied in a previous work [12], containing genetic information associated to 10 control (healthy) subject and 10 persons suffering from Obstructive Nephropathy (ON) [15, 16]. ON is one of the most complex renal diseases, with devastating consequences for the health of many new-borns. Childs affected by this disease suffer from a partial or complete blockage of the urinary tract, preventing a normal urinary flow; this, in turn, results in an accumulation of urine in the kidney. As a consequence, important lesions can appear in this organ that, in many cases, requires surgical intervention and even transplant. The genetic information available includes the expression levels of 834 microRNA, small RNA chains that block the transcription of other genes, and therefore regulate the metabolism [17, 18]. Furthermore, the dataset also includes, for each subject, the pelvic diameter, that is, a measure considered as a good proxy of the severity of the disease.

In the original work [12], it was shown how the topology of the networks created using the proposed method can unveil if a target subject is suffering, or not, from the disease; furthermore, some topological metrics of the networks were associated to two different measures of the severity of the illness, namely the differential renal function and the pelvic diameter.

Here, we focus on the relation between the structure of the networks and the pelvic diameter. The global structure of the network will be quantified by its *efficiency* [19], originally introduced to quantify how effectively information can cross a given graph. This metric is defined as:

$$E = \frac{1}{n(n-1)} \sum_{i,j \in G} \frac{1}{d(i,j)}, \tag{5}$$

where $n$ is the number of nodes (in this case, the number of features), $i$ and $j$ are two nodes of graph $G$, and $d(i,j)$ is the length of the shortest path (geodesic distance) between nodes $i$ and $j$. Notice that, if the network is composed of a small number of links, the distance $d(i,j)$ will be infinite in most of the cases, thus resulting in a very small efficiency.

## 4. **Feature selection methods.**

While it is possible to work directly with all the 834 features, i.e., microRNA expression levels, we are interested in the problem of *feature selection*, that is, the initial selection of a set of relevant features to be included in the analysis. Reducing the size of the initial dataset has three important advantages. Firstly, the computational cost, which approximately scales as the square of the number of features, is drastically reduced. Secondly, the elimination of features not relevant for the final result may improve the outcome of the algorithm, by reducing the quantity of noise it has to cope with. Finally, the reduction of the number of features implies that the number of dimensions of the space of the possible solutions is also reduced: this, in turn, improves the significance of results, thus leading to a more statistically relevant analysis.

In what follows, two different strategies for feature selection are presented: a first one based on the goodness of the linear fit of Eq. 1, and an alternative one based on a measure of mutual information between different features. In both cases, the process of feature selection starts by creating a *ranking* of features. We start by considering an initial network composed of 20 nodes, corresponding to the 20 features that display the highest scores in the ranking. Afterwards, new features are added, one at a time, following the ranking created at the beginning of the process - what is called a *greedy* algorithm [20].

### 4.1. **Goodness of linear fit.**

The first method is based on the first step of the proposed approach, i.e., the linear fit performed between each pair of features. From

Eq. 1, we can derive the goodness of the fit, that is, the corresponding Pearson's coefficient of determination $R^2$ [21]:

$$R^2_{ij} = 1 - \frac{\sum\limits_s \left( d_j^s - \tilde{d}_j^s \right)^2}{\sum\limits_s \left( d_j^s - \overline{d_j} \right)^2}, \tag{6}$$

where $i$ and $j$ are two different features, $\overline{d_j}$ is the mean value of feature $j$ for all subjects, and $\tilde{d}_j^s$ is the value of feature $j$ and subject $s$ calculated with the parameters of the linear fit, that is:

$$\tilde{d}_j^s = a_{ij} + b_{ij} d_i^s. \tag{7}$$

As is well known, $R^2$ usually lays between zero and one, where $R^2 = 1$ means that the values are perfectly described by a line.

Using this metric, a value $S$ is assigned to each feature, defined as:

$$S_i = \frac{1}{n} \sum_k R^2_{ik}. \tag{8}$$

While intuitively the opposite solution might appear more logical, the features that should be selected are the ones with a higher $S$, which means that the expression levels of these microRNAs fit well, on average, with the expression levels of other genes. When two features are completely uncorrelated, and thus the $R^2$ of their fit is close to zero, even in healthy subjects, no significant information can be extracted when patients are included in the analysis, as it would be masked by the general noise. On the other hand, perfectly correlated features (in healthy persons) can easily identify abnormal values corresponding to ON patients. Therefore, the ranking is created according to the value of $S$, and networks are constructed by including features with the highest $S_i$.

4.2. **Mutual information.** Mutual information is a well-known measure of mutual dependance between random variables [22]. If one considers the expression levels of two microRNA as two random variables ($d_i$ and $d_j$), the two marginal probabilities distribution functions [$p(d_i)$ and $p(d_j)$] and the joint probability distribution function $p(d_i, d_j)$, the mutual information $I$ for features $i$ and $j$ is defined as:

$$I_{ij} = \sum_{l=1}^{m} \sum_{k=1}^{m} p(d_i^l, d_j^k) \log_2 \left( \frac{p(d_i^l, d_j^k)}{p(d_i^l) p(d_j^k)} \right). \tag{9}$$

$I$ measures, in bits, how much information is shared by two features, i.e., how much the knowledge of one of them reduces the uncertainty about the other. As in the previous case, we are interested in selecting those features that, in the case of healthy persons, share much information with the others, thus not just codifying noise. Therefore, in a way similar to the goodness of fit, we create a metric defined as:

$$S'_i = \frac{1}{n} \sum_k I_{ik}. \tag{10}$$

The initial network is created with the 20 nodes with higher $S'_i$; afterwards, the network is expanded by including the following features, one at a time.

5. **Results.** Fig. 2 presents the evolution of the results as a funcion of the number of features included in the analysis (the number of nodes in each network). The performance of the methodology is represented by the goodness of fit between the efficiency of the constructed networks and the pelvic diameter, a proxy for the severity of the disease. Therefore, the complete evaluation of both algorithms is performed as follows. First, for a given set of features, a network is created for each person (both healthy and ON); afterwards, the efficiency (see Eq. 5) of each network is computed, and these latter values are fitted against the pelvic diameter by means of a second-order polynomial. Finally, the goodness of all the process (and, thus, the relevance of the selected features) is estimated through the coefficient of determination $R^2$ of the fit.

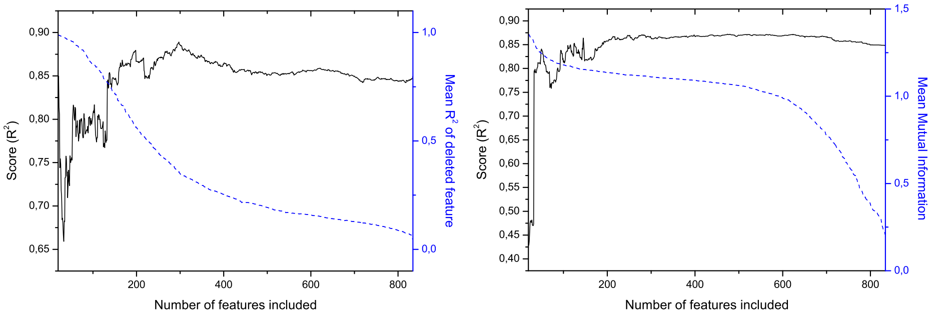

FIGURE 2. Performance of the two feature selection algorithms: (Left) goodness of linear fit and (Right) mutual information. The black solid lines represent the score (i.e., the goodness of the polinomial fit - see the text for further details) as a function of the number of features included in the analysis; blue dashed lines indicates the value of the metric ($R^2$ and Mutual Information $I$) associated with the feature included in each step.

As can be noticed from Fig. 2, both algorithms perform well for the scope of selecting the relevant features under which the results of the analysis are significant. An optimal result is obtained with a smaller number of features - the maximum of the $R^2$ corresponds to 300 features for the goodness of linear fit, and 280 for the method based on mutual information; therefore, two thirds of the initial features have been eliminated, thus reducing the computational cost by a factor of 10. Furthermore, and not surprisingly, the highest score achieved by both methods is higher than the score obtained by analyzing the whole dataset. This is due to the nature of the feature selection, as the least important features (which are, in the end, not codifying relevant information) are excluded from the analysis.

In Fig. 3 we represent the second-order polynomial fits corresponding to networks created with five different numbers of nodes, selected according to the mutual information criterion. It can be noticed that the last three plots, corresponding to 145, 280 and 834 nodes respectively, are depicting a clear relation between the characteristics of the networks and the pelvic diameter.

6. **Conclusions.** We report the results of the application of iterative feature selection in the construction of complex networks, with a specific application to the biomedical problem of relating levels of expression of microRNAs with the degree of

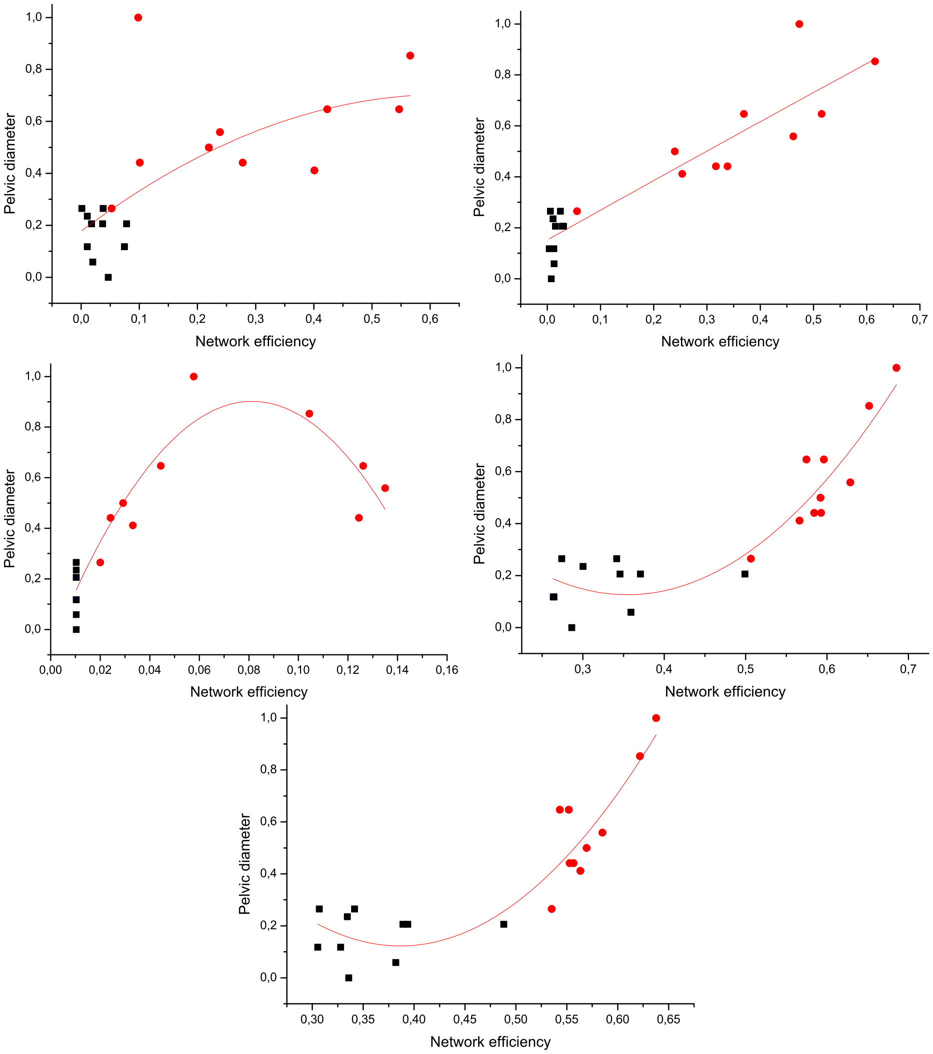

FIGURE 3. Comparison of the results obtained with different numbers of features: from left to right, top to bottom, 28 ($R^2 = 0.482$), 50 ($R^2 = 0.841$), 145 ($R^2 = 0.864$), 280 ($R^2 = 0.871$), and 834 (the full dataset, $R^2 = 0.850$). Black squares (red diamonds) represent values of control (ON) subjects.

severity of Obstructive Nephropathy. While this family of data-mining algorithms has been extensively studied in the last decades, less attention has been devoted to them from the community of researchers working with complex networks. Yet, the reduction of the number of features, and thus the reduction of the number of nodes, can drastically minimize the computational cost of the analysis, while, at the same time, improving the significance of the obtained network. The proposed approach, therefore, would be extremely useful in future genetic studies based on complex networks, where the quantity of available information (up to 20.000 expressions levels per subject) represent a computational challenge.

**Acknowledgments.** Computer resources granted by ENEA CRESCO (www.cresco.enea.it) are gratefully acknowledged.

## REFERENCES

[1] D. J. Lockhart and E. A. Winzeler, *Genomics, gene expression and DNA array*, Nature, **405** (2000), 827–836.

[2] K. I. Goh, et al., *The human disease network*, Proc. Natl. Acad. Sci. USA, **104** (2007), 8685–8690.

[3] T. R. Golub, et al., *Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring*, Science, **286** (1999), 531–537.

[4] L. J. van 't Veer, et al., *Gene expression profiling predicts clinical outcome of breast cancer*, Nature, **415** (2002), 530–536.

[5] R. Jaenisch and A. Bird, *Epigenetic regulation of gene expression: How the genome integrates intrinsic and environmental signals*, Nature Genetics, **33** (2003), 245–254.

[6] A. L. Barabási and Z. N. Oltvai, *Network biology: Understanding the cell's functional organization*, Nature Reviews Genetics, **5** (2004), 101–113.

[7] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez and D. U. Hwang, *Complex networks: Structure and dynamics*, Physics Reports, **424** (2006), 175–308.

[8] M. E. J. Newman, *The structure and function of complex networks*, SIAM Review, **45** (2003), 167–256.

[9] P. W. Anderson, *More is different*, Science, **177** (1972), 393–397.

[10] L. da F. Costa, O. N. Oliveira Jr., G. Travieso, F. A. Rodrigues, P. R. Villas Boas, L. Antiqueira, M. P. Viana and L. E. C. da Rocha, *Analyzing and modeling real-world phenomena with complex networks: A survey of applications*, preprint, arXiv:0711.3199.

[11] B. Zhang and S. Horvath, *A general framework for weighted gene co-expression network analysis*, Statistical Applications in Genetics and Molecular Biology, **4** (2005) 45 pp..

[12] M. Zanin and S. Boccaletti, *Complex networks analysis of Obstructive Nephropathy data*, Chaos, **21** (2011), 033103.

[13] I. Guyon and A. Elisseeff, *An introduction to variable and feature selection*, The Journal of Machine Learning Research, **3** (2003), 1–48.

[14] I. Guyon, S. Gunn, M. Nikravesh and L. A. Zadeh, "Feature Extraction-Foundations and Applications," $1^{st}$ edition, Springer-Verlag, Berlin, 2006.

[15] R. L. Chevalier, *Molecular and cellular pathophysiology of Obstructive Nephropathy*, Pediatric Nephrology, **13** (1999), 612–619.

[16] J. G. Wen, J. Frokiaer, T. M. Jorgensen and J. C. Djurhuus, *Obstructive Nephropathy: An update of the experimental research*, Urology Research, **27** (1999), 29–39.

[17] D. P. Bartel, *MicroRNAs: Genomics, biogenesis, mechanism, and function*, Cell, **116** (2009), 281–297.

[18] D. P. Bartel, *MicroRNAs: Target recognition and regulatory functions*, Cell, **136** (2009), 215–233.

[19] V. Latora and M. Marchiori, *Is the Boston subway a small-world network?*, Physica A, **314** (2002), 109–113.

[20] T. H. Cormen, C. E. Leiserson, R. L. Rivest and C. Stein, "Introduction to Algorithms," $3^{rd}$ edition, MIT Press, New York, 2009.

[21] R. G. D. Steel and J. H. Torrie, "Principles and Procedures of Statistics," $1^{st}$ edition, McGraw-Hill, New York, 1960.

[22] Karmeshu, "Entropy Measures, Maximum Entropy Principle and Emerging Applications," $1^{st}$ edition, Springer, Berlin, 2003.

Received December 2011; revised July 2012.

*E-mail address*: massimiliano.zanin@ctb.upm.es
*E-mail address*: emenasalvas@fi.upm.es
*E-mail address*: pas@fct.unl.pt
*E-mail address*: stefano.boccaletti@gmail.com