Introduction to spectral metrics in biological network theory

Roberto Visintainer[1,2], **Giuseppe Jurman**[1], **Marco Grimaldi**[1], **Cesare Furlanello**[1]
[1] Fondazione Bruno Kessler and [2] DISI University of Trento
I-38123 Trento, Italy
{visintainer, jurman, grimaldi, furlan}@fbk.eu

Abstract

Many functions have been recently defined to assess the similarity among networks as tools for quantitative comparison, with deeply different definitions and extents. Here we present a comparative overview of the spectral distances, highlighting their behavior in some basic cases of static and dynamic synthetic and *E. coli* networks. We show examples where spectral distances are more effective than classical methods in assessing network differences. In particular, we individuate Ipsen-Mikhailov distance as the most robust spectral measure and we show an example of application in a functional genomic task.

1 Introduction

The basic goal of network comparison is quantifying difference between two homogeneous objects in some network space. The theory of network measurements relies on the quantitative description of main properties such as degree distribution and correlation, path lenghts, diameter, clustering, presence of motives. These and other properties have been described for complex networks in [1]. Furthermore, network measurements can be encoded into a feature vector, yelding a representation convenient for classification tasks. The use of similarity measures on the topology of the underlying graphs defines a different strategy, whose roots date back to the 70's with the theory of graph distances (regarding both metrics inter- and intra-graphs [2]). Since then, a number of similarity measures have been introduced, including metrics relaxed to less stringent bounds. Cost-based functions stems from the parallel theory of graph alignment: the edit distance and its variants use the minimum cost of transformation of one graph into another by means of the usual edit operations - insertion and deletion of links. Feature-based measures are instead obtained when the similarity function is based on measurements feature vectors. One notable example in this family is the recently proposed use of ζ-functions for network volume measurements. Finally, the label "structure-based" distance groups all other measures that do not rely on cost functions or characteristic features. A typical example are those measures based on functions of the maximal common subgraphs between the two networks, or those based on the common motifs [3], i.e. patterns of interconnections occurring in complex networks significantly more often than in randomized networks. Remarkably, equivalence of some structure-based distance and the edit distance has been proven [4]). Although in most cases only network topology is considered, measures were also introduced that deal with directed or weighted links: for an example of a generic construction and an application to biological networks, see [5]. The family of spectral measures, which is investigated in this paper, is also part of the group of structure-based distances. Basically, it consists of a variety of maps of network's eigenvalues. The theory of graph spectra started in the early 50's and since then many of its aspects have been deeply mined, including a first classification of networks. The spectral theory has been applied to biological networks [6], where the properties of being scale-free (the degree distribution following a power law) and small-world (most nodes are not neighbors of one another, but most nodes can be reached from every other by a small number of hops or steps) are particularly evident. Estimates (also asymptotic)

of the eigenvalues distribution are available for complex networks [7]. The idea of using spectral measures for network comparison is instead only recent and it relies on similarity measures that are functions of the network eigenvalues. However, it is important to note that, because of the existence of isospectral networks, all these measures are indeed distances between classes of isospectral graphs: they are relatively rare (especially in real networks) and qualitatively similar. In this work, a comparative overview of the most common spectral similarity measures and of their basic properties is presented[1]: their behaviour is shown on a few synthetic benchmark datasets and on the transcriptional network of *E. coli*. On the ground of such experiments, we choose Ipsen-Mikhailov ϵ distance [8] out of the six original metrics on the ground of stability and robustness, and we conclude with an example of application on a profiling task on a Hepatocellular Carcinoma miRNA dataset.

2 Overview of spectral similarity measures

In this section, we introduce a set of similarity measures based on the graph spectra that was recently proposed in literature, following an ideal chronological timeline. The six distances described hereafter are analytically summarized in Tab. 1. They are functions of the eigenvalues of one of the connectivity matrices of the networks, namely: the adjacency matrix A, an $N \times N$ binary matrix whose nonzero entries denote the links between the graph's N nodes; the degree diagonal matrix D with the vertex degree deg, i.e. the number of edges touching the vertex itself; the Laplacian matrix $L = D - A$ and the normalized (isospectral) Laplacian matrices $\mathcal{L} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}}$ and $\Delta = D^{\frac{1}{2}} \mathcal{L} D^{-\frac{1}{2}}$, for $D^{-\frac{1}{2}}$ the diagonal matrix with entries $-\frac{\delta_{ij}}{\sqrt{\deg_i}}$. The first distance D1 (or, indeed, one-parameter family of distances) was introduced as an intra-graph measure [14] and mentioned as an inter-graph distance by Pincombe [9], for evaluating changes in time-series of graphs. The D1 measure is non-negative, separated, symmetric and it satisfies the triangle inequality, so it is a metric. The more refined spectral distance D2 was introduced as a step towards reconstructing a graph from its spectrum through a Metropolis algorithm [8]. Its definition follows the dynamical interpretation of a N-nodes network as a N-atoms molecules connected by identical elastic strings, where the pattern of connections is defined by the adjacency matrix of the corresponding network. The dynamical system is described by the set of N differential equations whose vibrational frequencies ω_i are given by squared root of the eigenvalues of the Laplacian matrix of the network. The spectral density $\rho(\omega)$ for a graph is defined as the sum of Lorentz distributions $\rho(\omega) = K \sum_{i=1}^{N-1} \frac{\gamma}{(\omega-\omega_k)^2+\gamma^2}$, where γ is the common width and K is the unit area normalization constant. A simpler measure D3 was introduced in [10] for graph matching, using the graph edit distance as the reference baseline. The authors compute the spectrum associated to the classical adjacency matrix, Laplacian matrix, signless Laplacian matrix $|L| = D + A$, and normalized Laplacian ($\mathcal{L}$) matrix. They also introduce two more functions: the path length distribution and the heat diffusion kernel h_t, related to the Laplacian by the equation A similar formula D4 is proposed in [11] as the squared Euclidean (L_2) distance between the vectors of the Laplacian matrix. The distance D5 is introduced in [12], aiming at comparing Internet networks topologies: the authors provide both an exact and an approximated version, both expressed in terms of the normalized Laplacian eigenvalues distribution f_λ, and of a suitable weighting function $\mu(\lambda) = (1-\lambda)^n$. The last spectral measure D6 in this review was presented in [13] and it employs two different divergence measures, Kullback-Leibler (KL) and Jensen-Shannon. The Kullback-Leibler divergence measure $\mathrm{KL}(p_1, p_2) = \sum_{x \in X} p_1(x) \log \frac{p_1(x)}{p_2(x)}$ for two probability distributions p_1, p_2 is not a metric, because is not symmetric and it does not satisfy the triangle inequality. To overcome this problem, the author consider the Jensen-Shannon JS measure, which is the symmetrization of KL.

3 Benchmarking experiments

In this section, we demonstrate the use of the distances in Tab. 1 in the comparison of network topologies in a controlled situation. To such aim, we constructed three synthetic benchmark datasets, detailed hereafter. All simulations have been performed within the R statistical environment. Throughout all simulations, we kept, for each distance, the parameter values as in the reference paper wherever possible, e.g., $\gamma = 0.08$ for the scale of the Lorentz distribution in D2; the

[1] An extended version of this paper can be found on arXiv:1005.0103

Table 1: Spectral graph distances: distance label, formula, involved connectivity matrix and literature reference for the six considered spectral measures. For distance D3, many different matrices can be considered; in the experiments of the present work, only the heat diffusion kernel h_t has been used.

Distance	Formula	Conn. Mat.	Ref.
D1	$d_k(G,H) = \begin{cases} \sqrt{\dfrac{\sum_{i=N-k}^{N-1}(\lambda_i-\mu_i)^2}{\sum_{i=N-k}^{N-1}\lambda_i^2}} & \text{if } \sum_{i=N-k}^{N-1}\lambda_i^2 \le \sum_{i=N-k}^{N-1}\mu_i^2 \\ \sqrt{\dfrac{\sum_{i=N-k}^{N-1}(\lambda_i-\mu_i)^2}{\sum_{i=N-k}^{N-1}\mu_i^2}} & \text{if } \sum_{i=N-k}^{N-1}\mu_i^2 < \sum_{i=N-k}^{N-1}\lambda_i^2 \end{cases}$	L	[9]
D2	$\epsilon(G,H) = \sqrt{\int_0^\infty [\rho_G(\omega)-\rho_H(\omega)]^2 \mathrm{d}\omega}$	L	[8]
D3	$d_M(G,H) = \sqrt{\sum_{i=0}^{N-1}\left(\lambda_i^{(G,M)}-\lambda_i^{(H,M)}\right)^2}$	(h_t)	[10]
D4	$d(G,H) = \sum_{i=0}^{N-1}\left(\lambda_i^{(G,L)}-\lambda_i^{(H,L)}\right)^2$	L	[11]
D5	$d(G,H) = \int_\lambda (1-\lambda)^4 \left(f_{\lambda,G}(\lambda)-f_{\lambda,H}(\lambda)\right)^2 \mathrm{d}\lambda$	$\mathcal{L}$	[12]
D6	$d(G,H) = \sqrt{\mathrm{JS}(f_G, f_H)}$	$\mathcal{L}$	[13]

heat diffusion kernel in D3; the time $t = 3.5$ for the kernel in distance D3. For D1 we chose to use the $\lfloor \frac{N}{2} \rfloor$ largest eigenvalues. The simulated topologies are generated within the R statistical environment by means of the simulator provided by the package *netsim* [15], producing networks that reproduce principal characteristics of transcriptional regulatory networks. The simulator takes into account the scale-free distribution of the connectivity and constructs networks whose clustering coefficient is independent of the number of nodes in the network. All random graphs are generated by keeping the default values of *netsim* for the structural parameters. In the first experiment we consider a random network A on N vertices and we compare it with the full connected network with the same number of nodes F, the complemental network $\overline{A}$ and a matrix A_p obtained from A by modifying (inserting/deleting) about the $p\%$ of the nodes. For smoothing purposes, the process is repeated b times to obtain the first benchmarking dataset $\mathcal{B}_1(b, N, p)$. Because of the small number of links in the original matrix, the 5% perturbation mostly reflects in links insertion. On average, the density of the original graph A can be expressed by the relation $l \simeq 1.7N - 5$, where l is the number of links and N the number of vertices. In the second experiment we simulate a time-series of T networks on N nodes starting from a randomly generated graph S_1, where each successive element S_i of the series is generated from its ancestor S_{i-1} by randomly modifying $p\%$ of the links.

Again $b = 50$ instances of the series are created and collected into the second benchmarking dataset $\mathcal{B}_2(b, T, N, p)$. With this strategy, the number of existing links is increasing with the series index, being the original adjacency matrix almost sparse. The starting matrix S_1 has on average 38.1±5.2 nodes, while the last element of the series S_{20} has 132.3±8.2. The third experiment is based on a benchmark dataset $\mathcal{B}_3(b, T, N, nd, na)$. Starting from $\mathcal{B}_2(b, T, N, p)$, different perturbations are applied: each successive element S_i of the series is generated from its ancestor S_{i-1} by randomly deleting nd links and adding na links. By construction, the number of existing links for all elements of the series is constant.

In Exp. 1 the six distances D1-D6 were applied on 4 instances of $\mathcal{B}_1(50, N, 5)$ for $N = 10, 20, 25, 100$ and distances between the original graph A and the three companion matrices F, $\overline{A}$ and A_p were computed. Distance D4 spans a considerably wider range than other measures, due to the absence of the square root in the comparison of the Laplacian spectra, while D5 is restricted into a very small interval. The same distance D4 also shows a high dependency on the dimension of the considered matrices and the number of the links. The best stability in terms of the relative standard deviation σ/μ is reached by D2 and D4. Furthermore, D2, differently from all other measures, is almost independent of the number of vertices. Finally, D6 is the only measure that, in the cases with $N > 10$, gives a lower distance for F than for $\overline{A}$. For Exp. 2, the benchmark dataset $\mathcal{B}_2(50, 20, 25, 5)$ was used. Distances between consecutive elements (S_i, S_{i+1}) of the series (defined Step i) were computed: results are averaged on the 50 replicates. For all D1-D6, distance decreases for increasing steps, although on different ranges (as already pointed out for Exp. 1) and with different widths for the confidence intervals. D3 and D5 decrease more quickly for initial steps, so they are less useful when comparing large networks. To better highlight similarities and differences among the distances regardless of their ranges of values, we also computed their mutual correlations. All correlation values are quite high, ranging from 0.8225 to 0.9970: D3 and D5 are mutually strongly correlated, but they tend to separate from the other distances, as evidenced both from the global correlation values and the scatter plot profiles distancing from the panel diagonals. Finally, Experiment 3 was performed on the benchmark dataset $\mathcal{B}_3(50, 20, 25, 5, 5)$. Since the difference between consecutive pairs of elements of the series is quite similar throughout all the steps, as expected all distances show a nearly constant trend. The obscillations around the mean value are nevertheless strongly varying among different measures. In particular, distance D3 is anticorrelated to all distances but D5; furthermore only in 4 cases out of 15 we obtain a correlation value higher than 0.7, with again $D1$, $D2$, $D4$ and $D6$ forming a group of more similar behaviour. Possible hierarchy of the six distances was explored by clustering. The clusters have average linkage and the correlation distance $\mathrm{cd}(\cdot, \cdot) = 1 - \mathrm{Corr}(\cdot, \cdot)$ is used as the dissimilarity measure. Although there is an appreciable coherence among measures on macroscopic trends, when downscaling to microscopic trends correlations get much looser. Distances $D1$, $D2$, $D4$, $D6$ seem to group together, while $D3$ has a more erratic behaviour. Finally, a wide range difference occurs in the cluster heights between the two experiments: the homogeneous macroscopic situation of Exp. 2 has a narrower height span than the microscopic case in Exp. 3. The full set of results, with figures and tables can be found in [16].

4 A regulatory network example

To conclude with, we apply D1-D6 to three different perturbations of the transcriptional interactions network[2] in *Escherichia coli*, described in [17]. The transcriptional database contains 577 interactions between 116 TFs and 419 operons. Starting from an existing database (RegulonDB[3]), the authors added 35 new TFs, including alternative sigma factors, and over a hundred new interactions from the literature. The original adjacency network (without self-interactions) consists of 420 vertices and 519 (undirected) links. To show the influence on distances, we compare the distances between the original network and the three networks obtained by silencing out (thus deleting the link involving such vertex) the activator/repressor factor *crp* and the two repressor factors *frn* and *himA*, having respectively 72, 22 and 21 links. In Tab. 2 we list the value of the distances between the original network EC and its three perturbations, denoted respectively as $EC_{\overline{crp}}$, $EC_{\overline{fnr}}$ and $EC_{\overline{himA}}$. All distances seem to be heavily dependent on the number of removed links: for all six distances,

[2]Publicly available at `http://www.weizmann.ac.il/mcb/UriAlon/Network_motifs_in_coli/ColiNet-1.1/`

[3]`http://regulondb.ccg.unam.mx/`

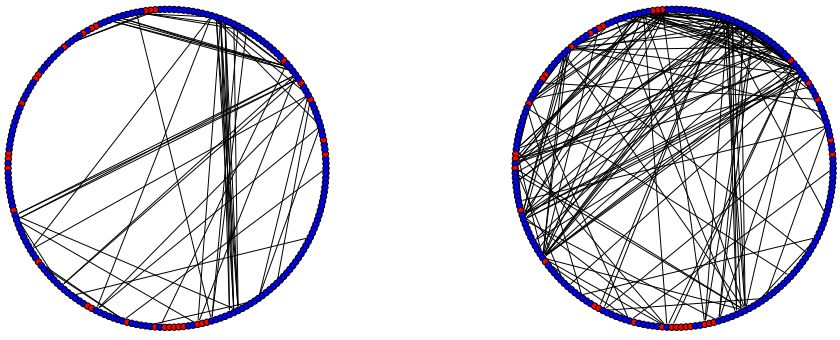
Table 2: Distances between EC and the perturbed networks $EC_{\overline{crp}}$, $EC_{\overline{fnr}}$ and $EC_{\overline{himA}}$.

Network	Links	D1	D2	D3	D4	D5	D6
(EC,$EC_{\overline{crp}}$)	519 vs 453	0.418	0.085	8.711	2191.9	$1.01178{\cdot}10^{-3}$	0.555
(EC,$EC_{\overline{fnr}}$)	519 vs 497	0.058	0.023	0.191	41.4	$0.01256{\cdot}10^{-3}$	0.083
(EC,$EC_{\overline{himA}}$)	519 vs 498	0.056	0.065	5.187	44.4	$0.43315{\cdot}10^{-3}$	0.404
($EC_{\overline{crp}}$,$EC_{\overline{fnr}}$)	453 vs 497	0.557	0.074	6.938	2140.9	$0.82079{\cdot}10^{-3}$	0.479
($EC_{\overline{crp}}$,$EC_{\overline{himA}}$)	453 vs 498	0.557	0.072	0.982	2138.1	$0.26794{\cdot}10^{-3}$	0.180
($EC_{\overline{fnr}}$,$EC_{\overline{himA}}$)	497 vs 498	0.023	0.071	3.730	10.2	$0.30303{\cdot}10^{-3}$	0.357

$D(EC, EC_{\overline{crp}}) > D(EC, EC_{\overline{fnr}}), D(EC, EC_{\overline{himA}})$. Nevertheless, when the number of removed links are almost equal, such relation is not valid anymore. The distance $D(EC, EC_{\overline{himA}})$ is comparable to $D(EC, EC_{\overline{fnr}})$ for $D = D1, D4$, while the former is much bigger than the latter for all other distances. The explanation is in the quite different structure of the two networks $EC_{\overline{fnr}}$ and $EC_{\overline{himA}}$, although being obtained silencing out almost the same number of links from the original network. The intrinsic structural difference between $EC_{\overline{fnr}}$ and $EC_{\overline{himA}}$ is indeed highlighted by the remarkable variation in the size of the respective group of automorphisms. For instance, the structure of $EC_{\overline{himA}}$ is almost $e^{347.4488-341.4692} \simeq 400$ times more symmetric than $EC_{\overline{fnr}}$. From this point of view, spectral distances can greatly help in analyzing subtle differences between networks where more classical methods are not helping much. As an example, the leading Laplacian eigenvalue is commonly used when network structure, because it is a good indicator of the stability and the local dynamics [18]. For instance here this value is of no help, since EC, $EC_{\overline{fnr}}$ and $EC_{\overline{himA}}$ have essentially the same leading eigenvalue; nevertheless the spectral distances, encoding information coming from the whole spectrum, can better separate very similar networks: in particular, $D2$ seems to achieve the msot significant results. Both the synthetic and the biological set of experiments indicate Ipsen-Mikhailow $D2$ distance as the most reliable metric, both in terms of stability and robustness in terms of being less prone to odd behaviours.

5 Association with profiling

In this section we show the use of the ϵ distance in a functional genomic task associating a profiling problem with the construction of relevance networks [19] on different classes of samples.



(a) M Tumorous (b) M Non Tumorous

Figure 1: Relevance network at correlation 0.85 for (a) the HCC tissues and (b) the healthy controls for the male patients, with top-20 ranked features marked as red nodes.

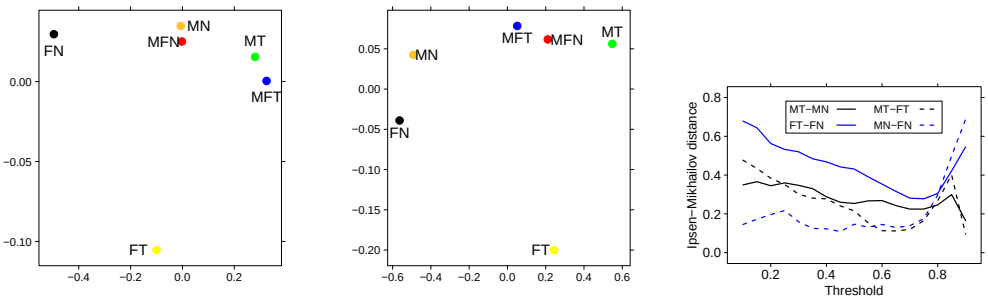
Table 3: Ipsen-Mikhailov distances are reported among all six networks at correlation threshold 0.85, on the whole set of 210 miRNA (upper triangular) or on the top-20 set of optimal features (lower triangular). M=Male patients, F=Female patients, T=Tumoral tissue, N=Non-tumoral tissue.

	$(M+F)_T$	$(M+F)_N$	M_T	M_N	F_T	F_N
$(M+F)_T$		0.338	0.068	0.346	0.441	0.822
$(M+F)_N$	0.192		0.292	0.052	0.168	0.502
M_T	0.537	0.382		0.300	0.401	0.777
M_N	0.562	0.723	1.043		0.175	0.499
F_T	0.347	0.281	0.417	0.779		0.422
F_N	0.645	0.785	1.121	0.175	0.830	

We start out from the hepatocellular carcinoma dataset introduced in the paper [20] and later used in [21], publicly available at the Gene Expression Omnibus (GEO, `http://www.ncbi.nlm.nih.gov/geo/`) at the accession number GSE6857. The dataset collects 482 tissue samples from 241 patients affected by hepatocellular carcinoma (HCC). For each patients, a sample from cancerous hepatic tissue and a sample from surrounding non-cancerous hepatic tissue are available, hybridized on the Ohio State University CCC MicroRNA Microarray Version 2.0 platform consisting of 11520 probes collecting expressions of 250 non-redundant human and 200 mouse microRNA (miRNA). After a preprocessing phase including imputation of missing values as in [22] and discarding probes corresponding to non-human (mouse and controls) miRNA, we end up with the dataset $\mathcal{HCC}$ of 240+240 paired samples described by 210 human miRNA. We perform three profiling experiments aimed at differentiating the two classes of samples (cancerous and non-cancerous) within the whole set $\mathcal{HCC}$, in the subset $\mathcal{HCC}_M$ of the 210+210 samples belonging to male patients and in the subset $\mathcal{HCC}_F$ of the 30+30 samples belonging to female patients. Mimicking the procedure in [20], we set up an experiment whose Data Analysis Protocol (DAP) consists of 1000 runs of a 10-fold Cross Validation.

The Spectral Regression Discriminant Analysis algorithm (SRDA for short, see [23]) is chosen as the classifier with parameter $\alpha = 100$. The feature ranking algorithm is the Entropy-based Recursive Feature Elimination (E-RFE, [24]), discarding at each step a bunch of features according to the shape of the distribution of the classifier weights. Classifier performance are computed as MCC averaged on the 1000 test set for models with different number of features, with confidence intervals as 95% student bootstrap. The homogeneity level of the 10000 ranked lists generated in each profiling experiment by the 100x10-CV DAP we use the stability indicator I introduced in [25] defined as the mean of mutual Canberra distances among the lists, normalized with respect to the whole set of possible permutations. The smaller the indicator value, the higher the stability level of the lists, with 0 corresponding to a set of 10000 identical lists and 1 to a set of randomly ranked lists. In all the three task the classifier can discriminate very well the two classes with a small and

Figure 2: Multidimensional scaling of the distances among all six relevance networks of Tab. 3, for the complete set of miRNA (a) and the top-20 subset (b). In panel (c), evolutions of distances between 4 couples of dynamic relevance networks as a function of the correlation threshold.



(a) (b) (c)

consistent (high stability) set of features. In particular, for all tasks, the model with 20 features is a reasonable compromise between classifier performance, list stability and small number of features, so we identify it as the optimal model for the investigated task The corresponding MCC values (with CI) are 0.845 (0.839,0.850) for $\mathcal{HCC}$, 0.931 (0.927,0.934) for $\mathcal{HCC}_M$ and 0.859 (0.846,0.871) for $\mathcal{HCC}_F$, with stability indicator values 0.166, 0.323 and 0.349, respectively. Then the optimal list for each of the three problems is computed as the top-20 sublist of the whole Borda list [25]. 16 miRNA are common to at least 2 of the top-20 models: in particular, 7 are common to all the three problems (221-prec, 222-precNo1, 26a-1No2, 021-prec-17No1, 099-prec-21, 128b-precNo1, 21No1). Most of them are already known in literature to be associated with hepatocellular carcinoma. We build now the relevance (o co-expression) networks for the two classes of samples in the three classification tasks considered. We recall that, in its simplest form, a weighted gene coexpression networks N on n genes $g_1, \ldots, g_n$ is defined in terms of the adjacency matrix as $A_N(g_i, g_j) = f(|\text{cor}(\mathbf{x}_i, \mathbf{x}_j)|^\beta)$, where $\mathbf{x}_i$ is the vector of the expression of gene g_i across the considered samples, $\beta \geq 1$ and f is a thresholding function, *e.g.*, $f(x) = \begin{cases} 1 & \text{for } x > 0.8 \\ 0 & \text{otherwise} \end{cases}$. In Fig. 1 we show an example of the relevance networks at threshold 0.85 for two of the considered cases: in general, the number of links in the healthy tissue case is always larger than in the cancerous tissue case, but such difference is less appreciable for female patients. In Tab. 3 the Ipsen-Mikhailov distances are reported among all six networks at correlation threshold 0.85, either on the whole set of 210 miRNA or on the top-20 set of optimal features. The multidimensional scaling projection of the two set of distances is displayed in Fig. 2. The distances among networks show that there are substantial differences not only between the Tumorous/NonTumorous tissue samples, but also between Male and Female patients, both on the cancerous and the surrounding healthy tissue relevance networks. In particular, it can be pointed out that the networks corresponding to the tumoral tissue for female patients has a deeply different structure with respect to all other networks, while the differences between the models on all patients and those on the sole male population are smaller: this might also be an effect of the different numerosity between male and female patients (210 versus 30). Finally, in Fig. 2(c) we show how distances between four couples of networks (Male/Female, Tumorous/Non-Tumorous) evolve with the correlation threshold travelling between 0.1 and 0.9. A few observations can be immediatly drawn from the 4 plotted curves: (i) The two closest networks along the correlation threshold are those corresponding to the Control tissue; (ii) The curves have a classwise related trend across gender: the two curves expressing respectively the distance in the Tumorous tissue case between Male and Female patients and the corresponding curve for the Control tissue have a similar shape (at least up to correlation threshold 0.8); (iii) As shown by the multidimensional scaling plot in Fig. 2(a-b), for Female patients, the Tumoral network is quite distant from the Control one, hilighting a wider biological transformation caused by the disease than in Male patients. Finally, we can conclude observing that Ipsen-Mikhailov distance can be an effective tool for helping in finding out a suitable correlation threshold for the relevance network, in order to better focus on interesting phenomena occuring between networks of different classes, from which to deduce biological insights on the underlying process.

Acknowledgments

The authors acknowledge funding by the European Union FP7 Project HiperDART, by the Italian Ministry of Health Project ISITAD (RF 2007 conv. 42), and by the PAT funded Project Envirochange.

References

[1] M.E.J. Newman. The Structure and Function of Complex Networks. *SIAM Review*, 45:167–256, 2003.

[2] R.C. Entringer, D.E. Jackson, and D.A. Snyder. Distance in graphs. *Czech. Math. J.*, 26(2):283–296, 1976.

[3] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network Motifs: Simple Building Blocks of Complex Networks. *Science*, 298(5594):824–827, 2002.

[4] H. Bunke. On a relation between graph edit distance and maximum common subgraph. *Pattern Recognition Letters*, 18:689–694, 1997.

[5] S.E. Ahnert, D. Garlaschelli, T.M.A. Fink, and G. Caldarelli. Applying weighted network measures to microarray distance matrices. *J. Phys. A: Math. Theor.*, 41:224011, 2008.

[6] A. Banerjee and J. Jost. Graph spectra as a systematic tool in computational biology. *Discrete Appl. Math.*, 157(10):2425–2431, 2009.

[7] G.J. Rodgers, K. Austin, B. Kahng, and D. Kim. Eigenvalue spectra of complex networks. *Journal of Physics A: Mathematical and General*, 38(43):9431, 2005.

[8] M. Ipsen and A.S. Mikhailov. Evolutionary reconstruction of networks. *Phys. Rev. E*, 66(4):046109, 2002.

[9] B. Pincombe. Detecting changes in time series of network graphs using minimum mean squared error and cumulative summation. In W. Read and A.J. Roberts, editors, *Proceedings of the 13th Biennial Computational Techniques and Applications Conference, CTAC-2006*, volume 48 of *ANZIAM J.*, pages C450–C473, 2007.

[10] P. Zhu and R.C. Wilson. A study of graph spectra for comparing graphs. In W. Clocksin, A. Fitzgibbon, and P. Torr, editors, *Proc. British Machine Vision Conference 16*, 2005.

[11] F. Comellas and J. Diaz-Lopez. Spectral reconstruction of complex networks. *Physica A*, 387:6436–6442, 2008.

[12] D. Fay, H. Haddadi, A.W. Moore, R. Mortier, S. Uhlig, and A. Jamakovic. A weighted spectrum metric for comparison of Internet topologies. *SIGMETRICS Perform. Eval. Rev.*, 37(3):67–72, 2009.

[13] A. Banerjee. Structural distance and evolutionary relationship of networks. arXiv:0807.3185v2 [q-bio.QM], 2009.

[14] D. Jakobson and I. Rivin. Extremal metrics on graphs, I. *Forum Math.*, 14(1):147–163, 2002.

[15] B. Di Camillo. *netsim: Gene network simulator*, 2007. R package version 1.1.

[16] G. Jurman, R. Visintainer, and C. Furlanello. An introduction to spectral distances in networks. In *Proc. of the 20th Italian Workshop on Neural Nets (WIRN2010)*. IOS Press, 2010 (In press).

[17] S.S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of Escherichia coli. *Nat. Genet.*, 31:64–68, 2002.

[18] R. Steuer and G. Zamora Lopez. Global network properties. In B.H. Junker and F. Schreiber, editors, *Analysis of biological networks*, pages 31–64. Wiley, 2008.

[19] W. Zhao, P. Langfelder, T. Fuller, J. Dong, A. Li, and S. Hovarth. Weighted Gene Coexpression Network Analysis: State of the Art. *J. Biopharm. Stat.*, 20(2):281–300, 2010.

[20] A. Budhu, H.-L. Jia, M. Forgues, C.-G. Liu, D. Goldstein, A. Lam, K. A. Zanetti, Q.-H. Ye, L.-X. Qin, C. M. Croce, Z.-Y. Tang, and X. W. Wang. Identification of Metastasis-Related MicroRNAs in Hepatocellular Carcinoma. *Hepatology*, 47(3):897–907, 2008.

[21] J. Ji, J. Shi, A. Budhu, Z. Yu, M. Forgues, S. Roessler, S. Ambs, Y. Chen, P.S. Meltzer, C.M. Croce, L.-X. Qin, K. Man, C.-M. Lo, J. Lee, I.O.L. Ng, J. Fan, Z.-Y. Tang, H.-C. Sun, and X.W. Wang. MicroRNA Expression, Survival, and Response to Interferon in Liver Cancer. *N. Engl. J. Med.*, 361:1437–1447, 2009.

[22] O.G. Troyanskaya, M. Cantor, G. Sherlock, P.O. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R.B. Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525, 2001.

[23] D. Cai, X. He, and J. Han. SRDA: An Efficient Algorithm for Large-Scale Discriminant Analysis. *IEEE Transactions on Knowledge and Data Engineering*, 20(1):1–12, 2008.

[24] C. Furlanello, M. Serafini, S. Merler, and G. Jurman. Entropy-Based Gene Ranking without Selection Bias for the Predictive Classification of Microarray Data. *BMC Bioinformatics*, 4(1):54, 2003.

[25] G. Jurman, S. Merler, A. Barla, S. Paoli, A. Galea, and C. Furlanello. Algebraic stability indicators for ranked lists in molecular profiling. *Bioinformatics*, 24(2):258–264, 2008.