Send Orders for Reprints to reprints@benthamscience.net

88 *Current Genomics*, **2021**, *22*, **88-97**

RESEARCH ARTICLE

On the Use of Topological Features of Metabolic Networks for the Classification of Cancer Samples

Jeaneth Machicao[1,2,+,*], Francesco Craighero[3,+], Davide Maspero[3,4], Fabrizio Angaroni[3], Chiara Damiani[5,6,†], Alex Graudenzi[4,7,†,*], Marco Antoniotti[3,7,†] and Odemir M. Bruno[1,†,*]

[1]São Carlos Institute of Physics, University of São Paulo, São Carlos, Brazil; [2]School of Engineering, University of São Paulo, São Paulo, Brazil; [3]Department of Informatics, Systems and Communication, University of Milan-Bicocca, Milan, Italy; [4]Institute of Molecular Bioimaging and Physiology, Consiglio Nazionale delle Ricerche (IBFM-CNR), Segrate, Milan, Italy; [5]Department of Biotechnology and Biosciences, University of Milan-Bicocca, Milan, Italy; [6]Sysbio Centre for Systems Biology, Milan, Italy; [7]Bicocca Bioinformatics, Biostatistics and Bioimaging Center (B4), University of Milan-Bicocca, Milan, Italy

ARTICLE HISTORY

Received: July 07, 2020
Revised: December 16, 2020
Accepted: December 18, 2020

DOI:
10.2174/1389202922666210301084151

Abstract: ***Background*:** The increasing availability of omics data collected from patients affected by severe pathologies, such as cancer, is fostering the development of data science methods for their analysis.

***Introduction*:** The combination of data integration and machine learning approaches can provide new powerful instruments to tackle the complexity of cancer development and deliver effective diagnostic and prognostic strategies.

***Methods*:** We explore the possibility of exploiting the topological properties of sample-specific metabolic networks as features in a supervised classification task. Such networks are obtained by projecting transcriptomic data from RNA-seq experiments on genome-wide metabolic models to define weighted networks modeling the overall metabolic activity of a given sample.

***Results*:** We show the classification results on a labeled breast cancer dataset from the TCGA database, including 210 samples (cancer *vs*. normal). In particular, we investigate how the performance is affected by a threshold-based pruning of the networks by comparing Artificial Neural Networks, Support Vector Machines and Random Forests. Interestingly, the best classification performance is achieved within a small threshold range for all methods, suggesting that it might represent an effective choice to recover useful information while filtering out noise from data. Overall, the best accuracy is achieved with SVMs, which exhibit performances similar to those obtained when gene expression profiles are used as features.

***Conclusion*:** These findings demonstrate that the topological properties of sample-specific metabolic networks are effective in classifying cancer and normal samples, suggesting that useful information can be extracted from a relatively limited number of features.

Keywords: Metabolic networks, cancer sample classification, machine learning, RNA-seq data, topological properties, network pruning.

1. INTRODUCTION

The development of automated strategies for the classification of cancer samples in distinct categories (*e.g.*, subtypes, risk groups, *etc*.) is one of the key challenges in current biosciences [1]. On the one hand, this might lead to the discovery of efficient, personalized diagnostic, prognostic, and therapeutic strategies for cancer patients. On the other hand, it could allow unraveling some of the still undeciphered mechanisms and processes underlying cancer development, leading to a data-driven understanding of the disease.

It is known that effective classification and clustering of cancer samples can be achieved by employing the information on expression data [2-7], genomic alteration profiles [8, 9], interaction networks [10], and even signaling pathways [11, 12]. In this work, however, we specifically focus on the metabolic properties that may distinguish cancer from normal samples. In fact, metabolic deregulation is one of the key hallmarks of cancer [13-15], even if its underlying mechanisms are still partially unknown. In this respect, in recent years, an increasing number of computational strate-

*Address correspondence to these authors at the São Carlos Institute of Physics, University of São Paulo, São Carlos, Brazil; Institute of Molecular Bioimaging and Physiology, Consiglio Nazionale delle Ricerche (IBFM-CNR), Segrate, Milan, Italy
E-mails: machicao@usp.br, alex.graudenzi@ibfm.cnr.it, bruno@ifsc.usp.br
[+]*Co-first authors*; [†]*Co-senior authors*.

gies have been devised, in order to take advantage of the growing availability and reliability of -omics data to investigate the alterations of metabolism in cancer [16-19]. Very often, such data have been employed in constraint-based models, such as Flux Balance Analysis (FBA), in which metabolic fluxes are simulated to compare different experimental scenarios [20-24].

Moreover, more recently, approaches coupling constraint-based metabolic modeling with supervised machine learning algorithms have been proposed [25]. In our case, we explore for the first time the possibility of employing the topological properties of metabolic networks as input features of classification algorithms. To this end, we rely on an approach firstly introduced in [26,27] in which transcriptomic data, such as RNA-seq, are employed to determine the approximate activity value of the reactions included in a given metabolic network.

More in detail, by introducing a relevance threshold on the metabolic activity level, we pruned the original metabolic network to define individual-specific networks in which only the significantly active reactions are preserved. The topological properties of such individual-specific networks are then used as features to perform a supervised classification task via various algorithmic strategies and, in particular, Multi-Layer Perceptrons (MLPs), Support Vector Machines (SVMs) and Random Forests (RFs).

To investigate our hypothesis, this work presents the classification results in a simple scenario in which the sample categories are known a priori - cancer *vs*. normal - concerning the TCGA-BRCA breast cancer dataset [28], which includes 210 total samples.

We show that noteworthy classification performance can be achieved by using a few key topological properties of metabolic networks, *i.e.*, average degree, average hierarchical degree, average geodesic path length and assortativity. Interestingly, a similar pruning threshold (in the range $0.01 – 0.1$) is identified as optimal for all tested machine learning strategies, suggesting that it could be an effective choice to extract useful information from the "relevant" activity of metabolic networks, while discarding possible artifacts due to noisy observations. Overall, the best classification performance is obtained with SVMs and threshold 0.1, which exhibit 0.866 of (average) accuracy, 0.86 precision and 0.879 recall on the test set, after k-fold cross-validation and hyper-parameter estimation. Furthermore, we show that the best performing SVM classifier (with the optimal threshold) delivers similar classification performance with respect to an analogous classifier processing a reduced gene expression feature vector, as computed by selecting the 5 principal components on the list of 1673 metabolic genes from Recon2.2 [29].

These results prove that the projection of transcriptomic activity on metabolic networks provides useful information to efficiently classify cancer samples and might pave the way for the development of strategies for experimental hypothesis generation.

2. MATERIALS AND METHODS

2.1. Integration of RNA-seq and Metabolic Networks

As proposed earlier [26, 27], it is possible to project transcriptomic data onto human metabolic networks [30], to derive an approximate activity value for each metabolic reaction in any given sample.

We first employ an input metabolic network M such as the Human Metabolic Reaction (HMR) [31] or Recon [29, 32]. M is a bipartite-directed graph that includes two kinds of nodes: (*i*) metabolites (*i.e.*, substrates or products), and (*ii*) metabolic reactions. The edges in M connect either: (*i*) the substrates and the relative reaction, or (*ii*) a reaction and the relative products. The total number of nodes of M is N, whereas the total number of edges is E. Reaction nodes are associated with Gene-Protein-Reaction (GPR) rules, *i.e.*, logical formulas that describe the related catalyses *via* AND and OR logical operators. In particular, AND rules are employed when distinct genes encode different *subunits* of the same enzyme, whereas OR rules are used when distinct genes encode *isoforms* of the same enzyme.

RNA-seq data are then used to provide an approximate activity value to each reaction in the input network. In particular, our method takes as input a n (*genes*) × m (*samples*) matrix T in which each element $T_{g,s},\ g = 1, \dots, n,\ s = 1, \dots, m$, includes the transcript level of gene g in sample s (the *Reads per Kilobase per Million* mapped reads – RPKM).

For each reaction in the input network $r \in G$ and for each sample $s = 1, \dots, m$, we define a *Reaction Activity Score* (RAS), by distinguishing two cases.

Reactions with GPR including an AND operator,

$$RAS_{r,s} = \min\left(T_{g,s}: g \in \mathcal{A}_r\right), \quad (1)$$

where $\mathcal{A}_r$ is the set of genes that encode the subunits of the enzyme catalyzing reaction r.

Reactions with GPR including an OR operator,

$$RAS_{r,s} = \sum_{g \in \mathcal{O}_r} T_{g,s}, \quad (2)$$

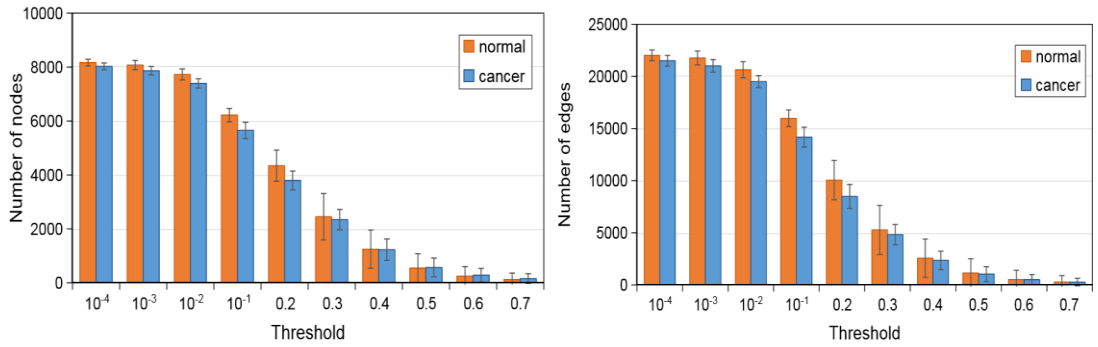
where $\mathcal{O}_r$ is the set of genes that encode isoforms of the enzyme that catalyzes reaction r.

In case of composite reactions, we respect the standard precedence of the two operators. The rationale underlying the definition of the RAS is that enzyme isoforms (OR) contribute *additively* to the overall activity of a certain reaction, whereas enzyme subunits (AND) *limit* its activity. RASs are finally normalized to obtain values in the range [0, 1] (with 0 meaning *no activity* and 1 meaning *maximum activity observed in the dataset*).

Even though this simplified approach neglects the heterogeneity of reaction kinetic constants, protein binding affinities and translation rates, it was proven effective in the investigation of cancer metabolic deregulation and in cancer sample stratification [26, 27].

2.2. Cancer Sample Classification *via* Metabolic Network Pruning

We define the sample-specific metabolic network of a given sample s as the weighted adjacency matrix W^s, which contains $N \times N$ elements, such that each element w_{ij}^s is equal to: (*i*) $RAS_{j,s}$ if i is a substrate of reaction j, (*ii*) $RAS_{i,s}$ if i is a reaction and j one of its products, (*iii*) 0 otherwise.



(a) (b)

Fig. (1). Number of nodes $\langle N^{T_l,S} \rangle$ (averaged on all samples) (**a**) and number of edges $\langle E^{T_l,S} \rangle$ (averaged on all samples) (**b**) of the giant component $G^{T_l,S}$ of the sample-specific metabolic network (computed from the Recon2.2 network [29]), in addition to their standard deviation (error bar), defined by different threshold T_l values either on normal and cancer samples. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

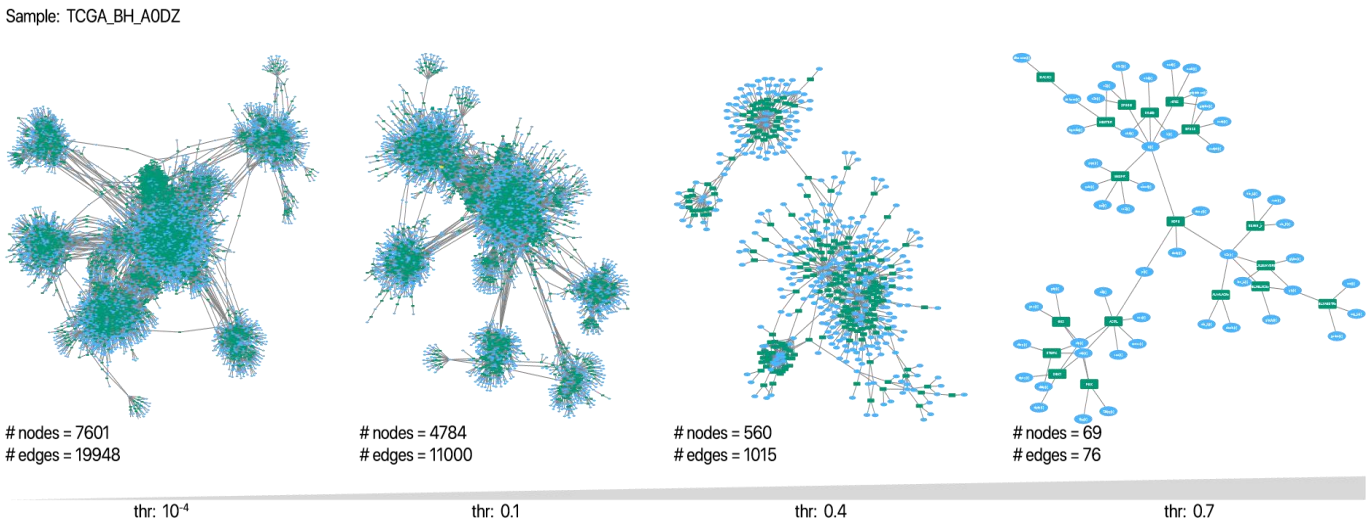


Fig. (2). The giant components of the metabolic network of the cancer sample of patient TCGA BH A0DZ obtained by projecting RNA-seq data on Recon2.2 metabolic network [29], are shown. 4 distinct giant components are shown, obtained with the following relevance thresholds: 10^{-4}, 0.1,0.4,0.7. Networks were drawn *via* Cytoscape [37]. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

Since we are interested in exploiting the topological properties of the "*giant component*" of the sample-specific metabolic network (as proposed, *e.g.*, in [33]), we employ a network pruning procedure to select the relevant metabolic reactions. This threshold criterion was employed earlier [34-36]. In detail, a threshold parameter $T_l \in [0,1]$ is used to obtain an unweighted and thresholded adjacency matrix $A^{T_l,S}$, the elements of which are defined as follows:

$$A_{ij}^{T_l,S} = \begin{cases} 1, \ if \ w_{ij}^S \geq T_l \\ 0, \ if \ w_{ij}^S < T_l \end{cases} \forall i,j = 1,\dots,N. \qquad (3)$$

It must be noted that we have focused on the *larger than* option, because we can hypothesize that only significantly active reactions (above the threshold) are responsible for the phenotypic/functional properties of cells. By scanning different values of the threshold, we can then evaluate the impact on the performance of classifiers that take as input certain topological measurements of the resulting giant component (see below), thus identifying an optimal threshold value.

Clearly the threshold parameter determines the size of the giant component, *i.e.*, the largest connected subgraph of the sample-specific metabolic network, which we define as $G^{T_l,S}$ and which includes $N^{T_l,S}$ nodes and $E^{T_l,S}$ edges.

For instance, in Fig. (**1**), one can see how the number of nodes and edges of the giant component of the sample-specific metabolic network (computed from the Recon2.2 network [29, 32]) is generally affected by the choice of distinct thresholds, regarding both cancer and normal samples. In greater detail, on the left side of Fig. (**1a**), smaller thresholds, such as $T_l \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$, lead to a larger size of the giant component (on average), while on the right side, larger thresholds, such as $T_l \in \{0.2, 0.3, \dots, 0.7\}$, lead to a radical network reduction, with a threshold $T_l = 0.7$ retaining 127 nodes on average, which represents approximately 1.45% of the total number of nodes of the original metabolic network. As a representative example, the shrinking of the giant component for a specific sample is visually represented in Fig. (**2**).

We also note that this behavior occurs similarly on both cancer and normal samples, even if the size of the giant component of the former ones tends to be slightly smaller. One may speculate that cancer subpopulations engage in a relatively lower number of metabolic functions with respect to normal cells, given that their main objective is "selfish" proliferation. Further investigations are needed to validate this interesting hypothesis [37].

2.3. Algorithmic Methods for Classification

In general, the choice of adequate network descriptors is crucial for pattern recognition purposes. Typically, the feature extraction is based on well-established network structural measures (see details in Section 2.3.1). The concurrent use of well-known measures such as *degree*, *mean degree*, *clustering coefficient*, *mean hierarchical degree*, *centrality*, and even *spectral measurements*, can identify global properties shared by a large majority of empirical and synthetic networks such as random, small-world, scale-free networks, and geographic networks models [38, 39].

2.3.1. Features Based on Network Structural Measures

Networks measurements falling in various categories (*e.g.*, connectivity-related, distance-related, spectral, degree correlation measures) can be effectively used to characterize the topological properties of real-world networks [38, 40]. In our case, we are interested in determining whether certain topological measurements of the giant component of the sample-specific metabolic network obtained from RNA-seq data projection, and after opportune threshold-based pruning, can be effectively employed as features to classify cancer samples. In particular, we selected the following measures.

Average Degree: Among the connectivity-related measurements, we here consider the degree (or connectivity) $k_i^{T_l,s}$ of node *i* of the giant component of sample *s*, given threshold T_l, as the number of neighbors of a node $i^{T_l,s}$ defined by:

$$k_i^{T_l,s} = \sum_{j=1}^{N^{T_l,s}} A_{ij}^{T_l,s}.$$

Accordingly, the average degree of the giant component is defined by Eq. (4), as follows:

$$\langle k^{T_l,s} \rangle = \frac{1}{N^{T_l,s}} \sum_{i=1}^{N^{T_l,s}} k_i^{T_l,s}. \quad (4)$$

Average Hierarchical Degree: The hierarchical degree $k_i^{T_l,s^h}$ of node i can also be measured considering the connectivity of the neighboring nodes constrained to a hierarchical level h. As an example, in social networks, the hierarchical degree of level 2 of given node i, k_i^2, is the sum of the degrees of the neighbors of its neighbors. Therefore, the mean hierarchical degree of the giant component of a sample-specific metabolic network is given by Eq. (5), as follows:

$$\langle k^{T_l,s^h} \rangle = \frac{1}{N^{T_l,s}} \sum_{i=1}^{N^{T_l,s}} k_i^{T_l,s^h}. \quad (5)$$

Average Geodesic Path Length: A path is defined as the sequence of nodes visited to go from node *i* to *j*. The distance between them is the number of edges within the path, and d_{ij} is defined as the geodesic path, *i.e.*, the smallest path length. When there is no path between *i* and *j*, $d_{ij} = 0$. The average geodesic path length of the giant component of the sample-specific metabolic network is given by:

$$\langle l^{T_l,s} \rangle = \frac{1}{N^{T_l,s}(N^{T_l,s}-1)} \sum_{i \neq j} d_{ij}, \quad (6)$$

where *i* and *j* are two nodes of the giant component and $\frac{1}{N^{T_l,s}(N^{T_l,s}-1)}$ corresponds to a normalization factor, considering a fully connected network [40].

Assortativity: The assortativity $\Gamma^{T_l,s}$ [41], *i.e.*, the Pearson correlation coefficient of degree among all pairs of linked nodes *i* and *j* of the giant component, quantifies the tendency of the nodes of a given degree *k* to connect to nodes with a similar degree and, in our case, it is defined as follows:

$$\Gamma^{T_l,s} = \frac{\left(\frac{1}{N^{T_l,s}}\right)\sum_{j>i}\left(k_i^{T_l,s}k_j^{T_l,s}A_{ij}^{T_l,s}\right) - \left[\left(1/N^{T_l,s}\right)\sum_{j>i}(1/2)\left(k_i^{T_l,s}+k_j^{T_l,s}\right)A_{ij}^{T_l,s}\right]^2}{\left(\frac{1}{N^{T_l,s}}\right)\sum_{j>i}(1/2)\left(k_i^{T_l,s^2}+k_j^{T_l,s^2}\right)A_{ij}^{T_l,s} - \left[\left(1/N^{T_l,s}\right)\sum_{j>i}(1/2)\left(k_i^{T_l,s}+k_j^{T_l,s}\right)A_{ij}^{T_l,s}\right]^2}, \quad (7)$$

$\Gamma^{T_l,s}$ is a value within the range [−1, 1]. Values closer to 1 indicate a positive correlation (nodes with high degree tend to connect to nodes with high degree), while values closer to −1, indicate a negative correlation (nodes with a high degree tend to connect to nodes with low degree), whereas values close to 0 indicates the absence of linear dependence.

In the following, we will show how to compose a feature vector by considering a set of topological measurements [35, 36, 38]. In this respect, the giant component of a sample-specific metabolic network $G^{T_l,s}$can be characterized by a tuple containing: *(i)* the average degree $\langle k^{T_l,s} \rangle$ (Eq. 4), *(ii)* the average hierarchical degree of level 2 $\langle k^{T_l,s} \rangle$ (Eq. 5), *(iii)* the average hierarchical degree of level 3 $\langle k^{T_l,s^3} \rangle$ (Eq. 5), *(iv)* the average geodesic path length $\langle l^{T_l,s} \rangle$ (Eq. 6) and *(v)* the assortativity $\Gamma^{T_l,s}$ (Eq. 7). The vector is given by:

$$\vec{\phi}(T_l, s) = [\langle k^{T_l,s} \rangle, \langle k^{T_l,s^2} \rangle, \langle k^{T_l,s^3} \rangle, \langle l^{T_l,s} \rangle, \Gamma^{T_l,s}] \quad (8)$$

We notice that other measures such as the clustering coefficient might be employed as features. However, since in our case the input network is bipartite, there are no triangle neighborhoods and, accordingly, the clustering coefficient would always be 0. Since our framework is designed to be general, one can expect this feature to be relevant in different experimental scenarios, with distinct datasets and alternative representations of reaction graphs [42-44].

2.4. Classification Setup

Given any relevance threshold T_l, the feature vectors are extracted for the resulting giant component of each sample *s*, and the classification step can be performed. The main goal of this analysis is to evaluate the classification performance of various classifiers $\mathcal{M}$, *i.e.*, MLPs, SVMs and RFs on the feature vector $\vec{\phi}(T_l, s)$. Furthermore, we tested the same classifiers on a reduced feature vector, including the 5 first principal components of the expression profiles of the 1673 metabolic genes present in the Recon2.2 model [29], in order to provide a comparison on the same number of features employed in our approach.

Table 1. Hyperparameters grid search for the tested classifiers, *i.e.*, MLPs, SVMs and RFs, executed *via* the scikit-learn Python library. Parameter names are the sklearn arguments of the related functions (default was used for the other parameters).

Methods	Functions	Parameters	Grid Search Values
MLP	neuralnetwork.MLPClassifier	solver	[adam, lbfgs]
		hidden_layer_sizes	[(50,),(100,),(50,50)]
		batch_size	[16, 32, 64]
		learning_rate_init	[0.1, 0.01, 0.001]
		learning_rate	[constant, adaptative]
		max_iter	10000
RF	ensemble.RandomForestClassifier	max_depth	[10, 20, 40, None]
		max_features	[auto, sqrt]
		min_samples_leaf	[1, 2, 3]
		min_samples_split	[2, 3, 5]
		n_estimator	[100, 200, 500, 1000]
SVM	svm.SVC	C	[2^{-5}, 2^{-4},..., 2^{12}]
		gamma	[2^{-15} 2^{-14},..., 2^{4}]
		tol	[10^{-3}, 10^{-4}]
		kernel	[rbf, sigmoid, linear]

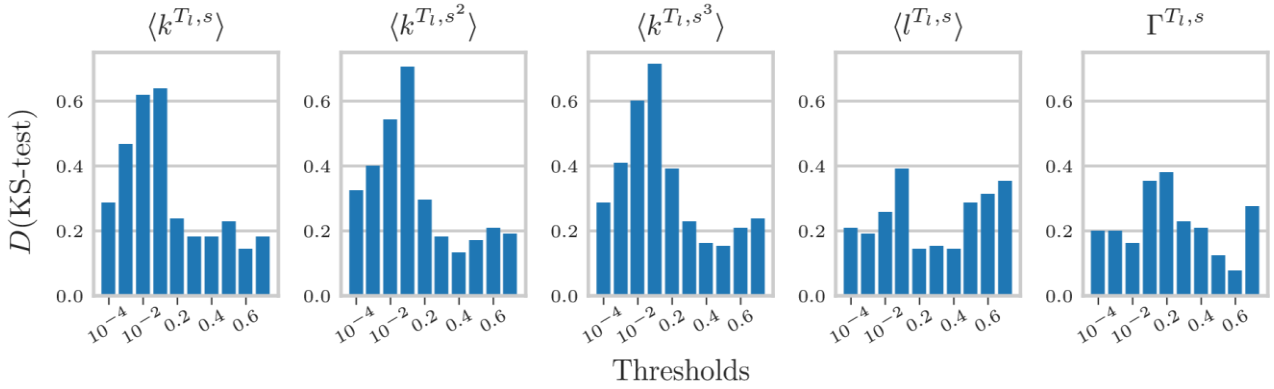


Fig. (3). Kolmogorov-Smirnov statistic (KS-test, [48]) between normal and cancer samples for each threshold and network topological measure: average degree $\langle k^{T_l,s} \rangle$, assortativity $\Gamma^{T_l,s}$ average hierarchical degree of level 2 $\langle k^{T_l,s^2} \rangle$ and 3 $\langle k^{T_l,s^3} \rangle$ and average geodesic path length $\langle l^{T_l,s} \rangle$. The higher the K-S test is, the more the distribution of the network measure is different between normal and cancer samples. The highest values are obtained with $\langle k^{T_l,s} \rangle$, $\langle k^{T_l,s^2} \rangle$, and $\langle k^{T_l,s^3} \rangle$ and thresholds equal to 10^{-2} and 0.1. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

In order to prevent over-optimistic results, we performed for each classifier a nested cross-validation as proposed earlier [45] and detailed as follows.

The original dataset, including cancer and normal samples, is split into 5 folds, ensuring the balance between classes. 5-fold outer cross-validation is executed by using: (*i*) one fold as the test set to assess the model performance and (*ii*) 4 folds in an inner 5-fold cross-validation procedure to select the optimal hyperparameters *h* of the model $\mathcal{M}(h)$ *via* grid search (Table **1**). The whole procedure is repeated 3 times to ensure robustness to the results. The performance of all classifiers is assessed on average accuracy, precision and recall with respect to ground-truth labels.

All the experiments described above were performed using the scikit-learn Python library [46].

2.5. Network Datasets

We tested our approach on the breast cancer dataset TCGA-BRCA published earlier [28]. We downloaded the dataset via the cBioPortal [47]. This dataset includes the expression profile (RNA Seq V2 RSEM) of biopsies taken from 817 patients. We selected the 105 patients for which the expression profiles of both cancer and normal tissues are provided, for a total of 210 samples used in our analysis.

RNA-seq data were projected on the Recon2.2 metabolic network [29, 32] to obtain a dataset in which a Reaction Activity Score is assigned to each metabolic reaction in each sample (see above). The RASs were then normalized by dividing each reaction score by the maximum value of all samples. Finally, normalized RAS profiles are used to weigh the metabolic network as described above.

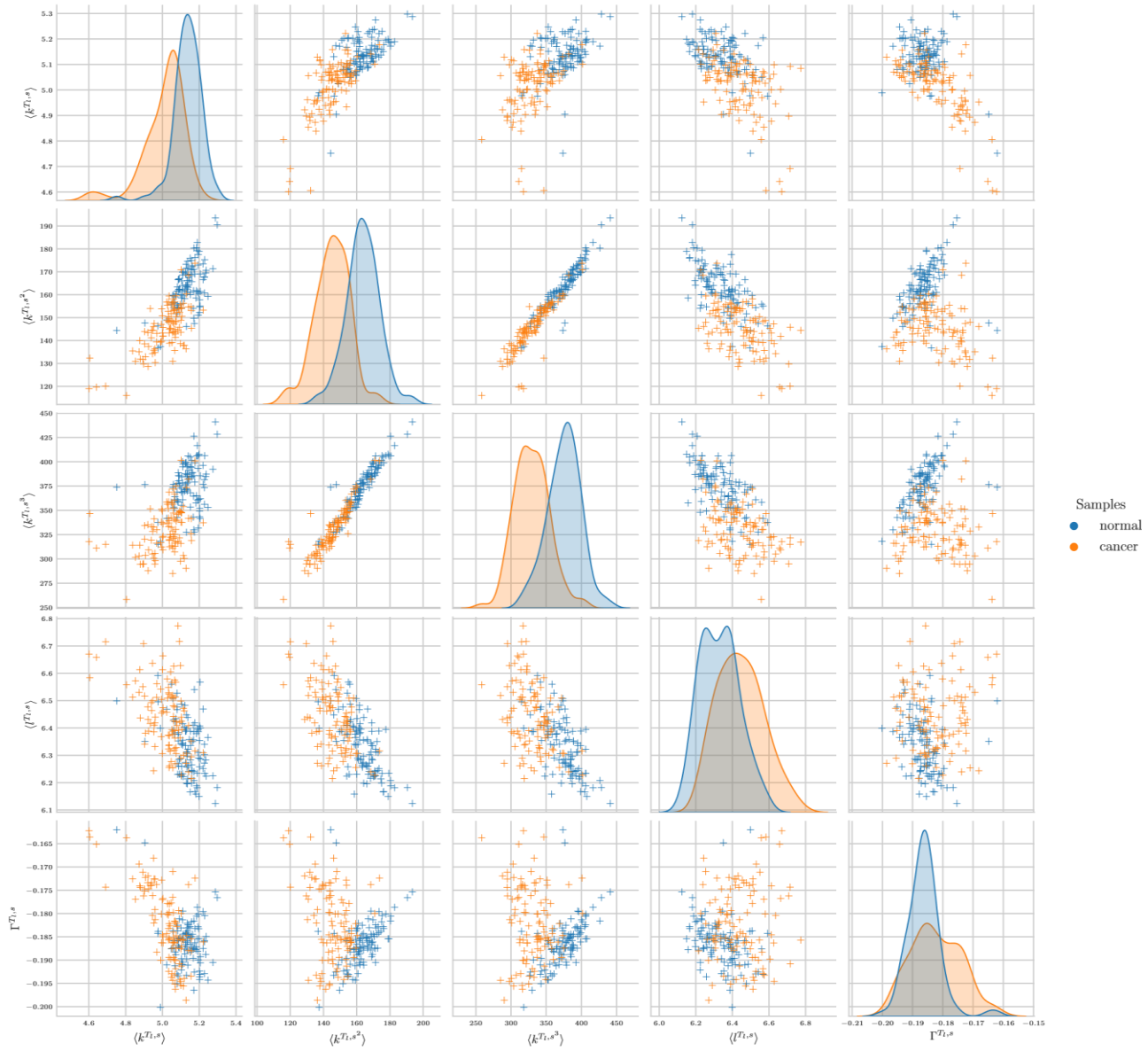


Fig. (4). Projection of cancer and normal samples on the space of topological measure pairs and (on the diagonal) the distribution for each measure and every sample category, for a selected threshold $T_l = 0.1$. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

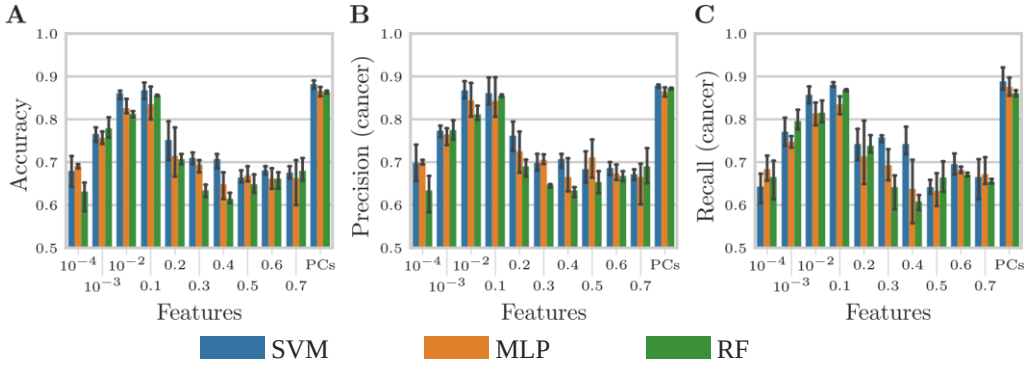


Fig. (5). From left to right: average accuracy (**A**), average precision on cancer samples (**B**) and average recall on cancer samples (**C**) with SVMs, MLPs and RFs. The average is computed on the test sets via a repeated nested cross-validation, for three different seeds, whereas the error bars represent the standard deviation (see Section 2.4 for additional details). The best thresholds are $T_l = 10^{-2}$ and $T_l = 0.1$. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

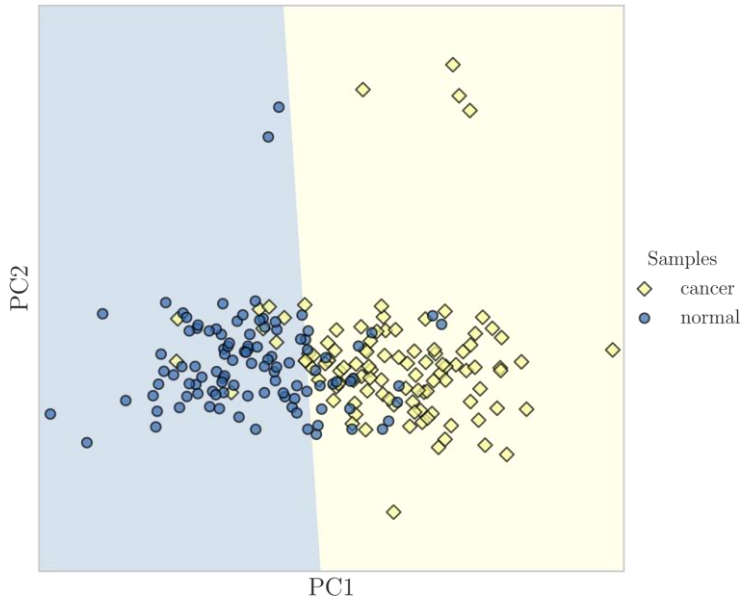


Fig. (6). Decision boundary of the SVM classifier with optimal hyperparameters and threshold $T_l = 0.1$ on the full dataset. The axes correspond to the first two principal components of the full feature vector $\vec{\phi}(T_l, s)$. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

3. RESULTS

3.1. RAS Threshold Analysis

A small T_l will result in larger giant components while, in contrast, higher values of T_l will result in smaller giant components. To choose the best classifier, we evaluated the performance obtained by the following distinct threshold values:

$$T_l \in \{10^{-4}, 10^{-3}, 10^{-2}, 0.1, 0.2, 0.3, \ldots, 0.7\}. \quad (9)$$

Thus, each feature vector $\vec{\phi}(T_l, s)$, contains the five topological measures defined above as descriptors (see Section 2.3.1).

To test the discrimination power of the feature vectors $\vec{\phi}(T_l, s)$, in Fig. (**3**), we computed the Kolmogorov-Smirnov statistic [48] between normal and cancer samples for each threshold and topological measure. The KS statistic *D* (KS-test) is the distance between the cumulative probability distributions; hence the higher is the value, the more the network measures are different between normal and cancer samples.

As a result, in our dataset, degree statistics, *i.e.*, $\langle k^{T_l,s} \rangle$, $\langle k^{T_l,s^2} \rangle$ and $\langle k^{T_l,s^3} \rangle$, achieve the highest *D* (KS-test), in particular for thresholds equal to 10^{-2} and 0.1. In Fig. (**4**), we plotted the distributions of all pairs of features in $\vec{\phi}(T_l, s)$, for $T_l = 0.1$. In accordance with the results of Fig. (**3**), the degree statistics distributions and, in particular, $\langle k^{T_l,s^2} \rangle$ and $\langle k^{T_l,s^3} \rangle$, have the sharpest difference among normal and cancer samples.

3.2. Classification Performance

The classification performance was assessed for all classifiers (*i.e.*, MLPs, SVMs and RFs) on the feature vector $\vec{\phi}(T_l, s)$, with regard to all relevance thresholds, via the nested cross-validation procedure described above (see Section 2.4). In addition, we employed as benchmark three analogous classifiers (*i.e.*, MLPs, SVMs and RFs), which were provided as input with a feature vector including the 5 first principal components (PCs) of the expression profiles of the 1673 metabolic genes.

In Fig. (**5**), we report the average accuracy, precision and recall for all tested classifiers, with respect to all relevance thresholds, as well as the benchmark classifiers on gene expression PCs, by employing the ground-truth cancer sample labels (the error bars represent the standard deviation).

Interestingly, the best performance is achieved for all classifiers with thresholds in the small range $T_l = 10^{-2}$ and $T_l = 0.1$, and points at the existence of an effective pruning strategy to maintain the "relevant" active metabolic pathways that discriminate cancer from normal samples, while limiting the confounding effects possibly due to noisy observations and biological variability.

More in detail, the best performing classifier is provided by SVMs, which reach an average accuracy of 0.86 and 0.87, a precision of 0.87 and 0.86 and a recall of 0.86 and 0.88, for $T_l = 10^{-2}$ and $T_l = 0.1$, respectively.

Interestingly, such performance is extremely similar to that obtained with SVMs on the vector of gene expression PCs (average accuracy = 0.88, precision = 0.88 and recall = 0.89) and slightly superior to that of MLPs and RFs on the same vector. This result suggests that the information extracted from the few selected topological measures on the giant component of the sample-specific metabolic network is effective in discriminating cancer from normal samples, similarly to benchmark approaches processing gene expression data (5).

Finally, in Fig. (**6**), the decision boundary of the best performing SVM classifier, *i.e.*, obtained with $T_l = 0.1$ and optimal hyperparameters is displayed on the first two PCs of the feature vector $\vec{\phi}(T_l, s)$, from which one can see that the method is able to correctly classify also the outliers of both categories.

CONCLUSION

In this work, we have introduced a new computational framework for the classification of cancer samples, which combines the integration of transcriptomic data and metabolic networks with state-of-the-art machine learning approaches. This task is of practical relevance in many biomedical contexts and might pave the way for the development of automated strategies for experimental hypothesis generation. In particular, the introduction of our framework contributes to the emerging field of approaches combining sample-specific metabolic modeling with machine learning to classify cancer samples and/or to predict drug response, as recently reviewed [49, 50].

More in detail, we here proved that the information on the metabolic activity of single samples, derived via integration of highly accessible RNA-seq data, can be effectively used to classify healthy and pathological states, a result that appears to be robust when the original networks are significantly pruned via a relevance threshold. All in all, this result would suggest that the useful information to determine possibly aberrant states in a given sample can be derived from the high-level (topological) properties of a relatively limited number of active processes. The identification and characterization of such processes deserve further investigation.

Regarding our machine learning approach, we here relied on classical topological measures, such as degree, hierarchical degrees, average geodesic path length and assortativity, to encode the structural information of the metabolic network. Additional experiments may employ recent graph representation learning techniques [51, 52], including graph kernels [53] and convolutional neural networks on graphs [54], to automatically extract a low-dimensional feature vector of the input network.

We finally remark that extensions of the framework are currently ongoing to test its applicability to more complex scenarios, involving, for instance, multiclass and multi-label classification with respect to cancer subtypes and risk categories.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

HUMAN AND ANIMAL RIGHTS

No animals/humans were used for studies that are the basis of this research.

CONSENT FOR PUBLICATION

Not applicable.

AVAILABILITY OF DATA AND MATERIALS

The datasets generated and analyzed for this study can be found at this link: https://github.com/BIMIB-DISCo/METNET-CLASSIFICATION.

FUNDING

Financial support from the Italian Ministry of University and Research (MIUR) through grant "Dipartimenti di Eccellenza 2017" to University of Milano-Bicocca, Department of Biotechnology and Biosciences is acknowledged. Partial support was also provided by the CRUK/AECC/AIRC Accelerator Award #22790, "Single-cell Cancer Evolution in the Clinic". J.M. is grateful for the support from the National Council for Scientific and Technological Development (CNPq grant #155957/2018-0) and São Paulo Research Foundation (FAPESP grant #2020/03514-9).

O.M.B. acknowledges support from CNPq (Grant #307897/2018-4) and FAPESP (grant #2014/08026-1 and 2016/18809-9).

This work was also partially supported by a Bicocca 2020 Starting Grant to F.A.

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

[1] Kourou, K.; Exarchos, T.P.; Exarchos, K.P.; Karamouzis, M.V.; Fotiadis, D.I. Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J.*, **2014**, *13*, 8-17. http://dx.doi.org/10.1016/j.csbj.2014.11.005 PMID: 25750696

[2] Furey, T.S.; Cristianini, N.; Duffy, N.; Bednarski, D.W.; Schummer, M.; Haussler, D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, **2000**, *16*(10), 906-914. http://dx.doi.org/10.1093/bioinformatics/16.10.906 PMID: 11120680

[3] Sotiriou, C.; Neo, S-Y.; McShane, L.M.; Korn, E.L.; Long, P.M.; Jazaeri, A.; Martiat, P.; Fox, S.B.; Harris, A.L.; Liu, E.T. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc. Natl. Acad. Sci. USA*, **2003**, *100*(18), 10393-10398. http://dx.doi.org/10.1073/pnas.1732912100 PMID: 12917485

[4] Lu, J.; Getz, G.; Miska, E.A.; Alvarez-Saavedra, E.; Lamb, J.; Peck, D.; Sweet-Cordero, A.; Ebert, B.L.; Mak, R.H.; Ferrando, A.A.; Downing, J.R.; Jacks, T.; Horvitz, H.R.; Golub, T.R. MicroRNA expression profiles classify human cancers. *Nature*, **2005**, *435*(7043), 834-838. http://dx.doi.org/10.1038/nature03702 PMID: 15944708

[5] CP de Souto, M.; G Costa, I.; SA de Araujo, D.; B Ludermir, T.; Schliep, A. Clustering cancer gene expression data: a comparative study. *BMC Bioinformatics*, **2008**, *9*(1), 497. http://dx.doi.org/10.1186/1471-2105-9-497

[6] Vanneschi, L.; Farinaccio, A.; Mauri, G.; Antoniotti, M.; Provero, P.; Giacobini, M. A comparison of machine learning techniques for survival prediction in breast cancer. *BioData Min.*, **2011**, *4*(1), 12. http://dx.doi.org/10.1186/1756-0381-4-12 PMID: 21569330

[7] Curtis, C.; Shah, S.P.; Chin, S.F.; Turashvili, G.; Rueda, O.M.; Dunning, M.J.; Speed, D.; Lynch, A.G.; Samarajiwa, S.; Yuan, Y.; Gräf, S.; Ha, G.; Haffari, G.; Bashashati, A.; Russell, R.; McKinney, S.; Langerød, A.; Green, A.; Provenzano, E.; Wishart, G.; Pinder, S.; Watson, P.; Markowetz, F.; Murphy, L.; Ellis, I.; Purushotham, A.; Børresen-Dale, A.L.; Brenton, J.D.; Tavaré, S.; Caldas, C.; Aparicio, S. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, **2012**, *486*(7403), 346-352. http://dx.doi.org/10.1038/nature10983 PMID: 22522925

[8] Caravagna, G.; Graudenzi, A.; Ramazzotti, D.; Sanz-Pamplona, R.; De Sano, L.; Mauri, G.; Moreno, V.; Antoniotti, M.; Mishra, B. Algorithmic methods to infer the evolutionary trajectories in cancer progression. *Proc. Natl. Acad. Sci. USA*, **2016**, *113*(28), E4025-E4034.

http://dx.doi.org/10.1073/pnas.1520213113 PMID: 27357673

[9] Caravagna, G.; Giarratano, Y.; Ramazzotti, D.; Tomlinson, I.; Graham, T.A.; Sanguinetti, G.; Sottoriva, A. Detecting repeated cancer evolution from multi-region tumor sequencing data. *Nat. Methods*, **2018**, *15*(9), 707-714. http://dx.doi.org/10.1038/s41592-018-0108-x PMID: 30171232

[10] Hofree, M.; Shen, J.P.; Carter, H.; Gross, A.; Ideker, T. Network-based stratification of tumor mutations. *Nat. Methods*, **2013**, *10*(11), 1108-1115. http://dx.doi.org/10.1038/nmeth.2651 PMID: 24037242

[11] Michael, L.G.; Joseph, E.L.; William, T.B.; Jong, W.K; Quanli, W.; Matthew, D.C; Michael, B.D; Michael, K.; Bernard Mathey, P.; Anil, P. A pathway-based classification of human breast cancer. *Proceedings of the National Academy of Sciences*, **2010**, *107*(15), 6994-6999.

[12] Graudenzi, A.; Cava, C.; Bertoli, G.; Fromm, B.; Flatmark, K.; Mauri, G.; Castiglioni, I. Pathway-based classification of breast cancer subtypes. *Front. Biosci.*, **2017**, *22*, 1697-1712. http://dx.doi.org/10.2741/4566 PMID: 28410140

[13] Hanahan, D.; Weinberg, R.A. Hallmarks of cancer: the next generation. *Cell*, **2011**, *144*(5), 646-674. http://dx.doi.org/10.1016/j.cell.2011.02.013 PMID: 21376230

[14] Cantor, J.R.; Sabatini, D.M. Cancer cell metabolism: one hallmark, many faces. *Cancer Discov.*, **2012**, *2*(10), 881-898. http://dx.doi.org/10.1158/2159-8290.CD-12-0345 PMID: 23009760

[15] Ward, P.S.; Thompson, C.B. Metabolic reprogramming: a cancer hallmark even warburg did not anticipate. *Cancer Cell*, **2012**, *21*(3), 297-308. http://dx.doi.org/10.1016/j.ccr.2012.02.014 PMID: 22439925

[16] Tomita, M.; Kami, K. Cancer. Systems biology, metabolomics, and cancer metabolism. *Science*, **2012**, *336*(6084), 990-991. http://dx.doi.org/10.1126/science.1223066 PMID: 22628644

[17] Teicher, B.A.; Linehan, W.M.; Helman, L.J. Targeting cancer metabolism. **2012**, *18*(20), 5537-5545.

[18] Hyduke, D.R.; Lewis, N.E.; Palsson, B.Ø. Analysis of omics data with genome-scale models of metabolism. *Mol. Biosyst.*, **2013**, *9*(2), 167-174. http://dx.doi.org/10.1039/C2MB25453K PMID: 23247105

[19] Lewis, N.E.; Abdel-Haleem, A.M. The evolution of genome-scale models of cancer metabolism. *Front. Physiol.*, **2013**, *4*, 237. http://dx.doi.org/10.3389/fphys.2013.00237 PMID: 24027532

[20] Orth, J.D.; Thiele, I.; Palsson, B.Ø. What is flux balance analysis? *Nat. Biotechnol.*, **2010**, *28*(3), 245-248. http://dx.doi.org/10.1038/nbt.1614 PMID: 20212490

[21] Machado, D.; Herrgård, M. Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism. *PLOS Comput. Biol.*, **2014**, *10*(4), e1003580. http://dx.doi.org/10.1371/journal.pcbi.1003580 PMID: 24762745

[22] Jamialahmadi, O.; Hashemi-Najafabadi, S.; Motamedian, E.; Romeo, S.; Bagheri, F. A benchmark-driven approach to reconstruct metabolic networks for studying cancer metabolism. *PLOS Comput. Biol.*, **2019**, *15*(4), e1006936. http://dx.doi.org/10.1371/journal.pcbi.1006936 PMID: 31009458

[23] Damiani, C.; Di Filippo, M.; Pescini, D.; Maspero, D.; Colombo, R.; Mauri, G. popFBA: tackling intratumour heterogeneity with Flux Balance Analysis. *Bioinformatics*, **2017**, *33*(14), i311-i318. http://dx.doi.org/10.1093/bioinformatics/btx251 PMID: 28881985

[24] Damiani, C.; Maspero, D.; Di Filippo, M.; Colombo, R.; Pescini, D.; Graudenzi, A.; Westerhoff, H.V.; Alberghina, L.; Vanoni, M.; Mauri, G. Integration of single-cell RNA-seq data into population models to characterize cancer metabolism. *PLOS Comput. Biol.*, **2019**, *15*(2), e1006733. http://dx.doi.org/10.1371/journal.pcbi.1006733 PMID: 30818329

[25] Damiani, C.; Gaglio, D.; Sacco, E.; Alberghina, L.; Vanoni, M. Systems metabolomics: from metabolomic snapshots to design principles. *Curr. Opin. Biotechnol.*, **2020**, *63*, 190-199. http://dx.doi.org/10.1016/j.copbio.2020.02.013 PMID: 32278263

[26] Graudenzi, A.; Maspero, D.; Di Filippo, M.; Gnugnoli, M.; Isella, C.; Mauri, G.; Medico, E.; Antoniotti, M.; Damiani, C. Integration of transcriptomic data and metabolic networks in cancer samples reveals highly significant prognostic power. *J. Biomed. Inform.*, **2018**, *87*, 37-49. http://dx.doi.org/10.1016/j.jbi.2018.09.010 PMID: 30244122

[27] Damiani, C.; Rovida, L.; Maspero, D.; Sala, I.; Rosato, L.; Di Filippo, M.; Pescini, D.; Graudenzi, A.; Antoniotti, M.; Mauri, G. MaREA4Galaxy: Metabolic reaction enrichment analysis and visualization of RNA-seq data within Galaxy. *Comput. Struct. Biotechnol. J.*, **2020**, *18*, 993-999. http://dx.doi.org/10.1016/j.csbj.2020.04.008 PMID: 32373287

[28] Ciriello, G.; Gatza, M.L.; Beck, A.H.; Wilkerson, M.D.; Rhie, S.K.; Pastore, A.; Zhang, H.; McLellan, M.; Yau, C.; Kandoth, C.; Bowlby, R.; Shen, H.; Hayat, S.; Fieldhouse, R.; Lester, S.C.; Tse, G.M.; Factor, R.E.; Collins, L.C.; Allison, K.H.; Chen, Y.Y.; Jensen, K.; Johnson, N.B.; Oesterreich, S.; Mills, G.B.; Cherniack, A.D.; Robertson, G.; Benz, C.; Sander, C.; Laird, P.W.; Hoadley, K.A.; King, T.A.; Perou, C.M. Comprehensive molecular portraits of invasive lobular breast cancer. *Cell*, **2015**, *163*(2), 506-519. http://dx.doi.org/10.1016/j.cell.2015.09.033 PMID: 26451490

[29] Swainston, N.; Smallbone, K.; Hefzi, H.; Dobson, P.D.; Brewer, J.; Hanscho, M.; Zielinski, D.C.; Ang, K.S.; Gardiner, N.J.; Gutierrez, J.M.; Kyriakopoulos, S.; Lakshmanan, M.; Li, S.; Liu, J.K.; Martínez, V.S.; Orellana, C.A.; Quek, L.E.; Thomas, A.; Zanghellini, J.; Borth, N.; Lee, D.Y.; Nielsen, L.K.; Kell, D.B.; Lewis, N.E.; Mendes, P. Recon 2.2: from reconstruction to model of human metabolism. *Metabolomics*, **2016**, *12*(7), 109. http://dx.doi.org/10.1007/s11306-016-1051-4 PMID: 27358602

[30] Cazzaniga, P.; Damiani, C.; Besozzi, D.; Colombo, R.; Nobile, M.S.; Gaglio, D.; Pescini, D.; Molinari, S.; Mauri, G.; Alberghina, L.; Vanoni, M. Computational strategies for a system-level understanding of metabolism. *Metabolites*, **2014**, *4*(4), 1034-1087. http://dx.doi.org/10.3390/metabo4041034 PMID: 25427076

[31] Mardinoglu, A.; Agren, R.; Kampf, C.; Asplund, A.; Uhlen, M.; Nielsen, J. Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease. *Nat. Commun.*, **2014**, *5*, 3083. http://dx.doi.org/10.1038/ncomms4083 PMID: 24419221

[32] Thiele, I.; Swainston, N.; Fleming, R.M.; Hoppe, A.; Sahoo, S.; Aurich, M.K.; Haraldsdottir, H.; Mo, M.L.; Rolfsson, O.; Stobbe, M.D.; Thorleifsson, S.G.; Agren, R.; Bölling, C.; Bordel, S.; Chavali, A.K.; Dobson, P.; Dunn, W.B.; Endler, L.; Hala, D.; Hucka, M.; Hull, D.; Jameson, D.; Jamshidi, N.; Jonsson, J.J.; Juty, N.; Keating, S.; Nookaew, I.; Le Novère, N.; Malys, N.; Mazein, A.; Papin, J.A.; Price, N.D.; Selkov, E., Sr; Sigurdsson, M.I.; Simeonidis, E.; Sonnenschein, N.; Smallbone, K.; Sorokin, A.; van Beek, J.H.; Weichart, D.; Goryanin, I.; Nielsen, J.; Westerhoff, H.V.; Kell, D.B.; Mendes, P.; Palsson, B.Ø. A community-driven global reconstruction of human metabolism. *Nat. Biotechnol.*, **2013**, *31*(5), 419-425. http://dx.doi.org/10.1038/nbt.2488 PMID: 23455439

[33] Ma, H-W.; Zeng, A-P. The connectivity structure, giant strong component and centrality of metabolic networks. *Bioinformatics*, **2003**, *19*(11), 1423-1430. http://dx.doi.org/10.1093/bioinformatics/btg177 PMID: 12874056

[34] Backes, A.R.; Casanova, D.; Bruno, O.M. A complex network-based approach for boundary shape analysis. *Pattern Recognit.*, **2009**, *42*(1), 54-67. http://dx.doi.org/10.1016/j.patcog.2008.07.006

[35] Miranda, G.H.B.; Machicao, J.; Bruno, O.M. An optimized shape descriptor based on structural properties of networks. *Digit. Signal Process.*, **2018**, *82*, 216-229. http://dx.doi.org/10.1016/j.dsp.2018.06.010

[36] Machicao, J.; Filho, H.A.; Lahr, D.J.G.; Buckeridge, M.; Bruno, O.M. Topological assessment of metabolic networks reveals evolutionary information. *Sci. Rep.*, **2018**, *8*(1), 15918. http://dx.doi.org/10.1038/s41598-018-34163-7 PMID: 30374088

[37] Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N.S.; Wang, J.T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **2003**, *13*(11), 2498-2504. http://dx.doi.org/10.1101/gr.1239303 PMID: 14597658

[38] Costa, L.D.F.; Boas, P.R.V.; Silva, F.N.; Rodrigues, F.A. A pattern recognition approach to complex networks. *J. Stat. Mech.*, **2010**, *2010*(11), P11015. http://dx.doi.org/10.1088/1742-5468/2010/11/P11015

[39] Banerjee, A.; Jost, J. Spectral plot properties: Towards a qualitative classification of networks. *NHM*, **2008**, *3*(2), 395-411. http://dx.doi.org/10.3934/nhm.2008.3.395

[40] Costa, L da F.; Francisco, A.; Rodrigues, G.T.; Villas Boas, P.R. Characterization of complex networks: A survey of measurements. *Adv. Phys.*, **2007**, *56*(1), 167-242. http://dx.doi.org/10.1080/00018730601170527

[41] Newman, M.E. Assortative mixing in networks. *Phys. Rev. Lett.*, **2002**, *89*(20), 208701. http://dx.doi.org/10.1103/PhysRevLett.89.208701 PMID: 12443515

[42] Filisetti, A.; Graudenzi, A.; Serra, R.; Villani, M.; De Lucrezia, D.; Rudolf, M. Füchslin, Stuart A Kauffman, Norman Packard, and Irene Poli. A stochastic model of the emergence of autocatalytic cycles. *J. Syst. Chem.*, **2011**, *2*(1), 2. http://dx.doi.org/10.1186/1759-2208-2-2

[43] Filisetti, A.; Graudenzi, A.; Serra, R.; Villani, M.; Füchslin, R.M.; Packard, N.; Kauffman, S.A.; Poli, I. A stochastic model of autocatalytic reaction networks. *Theory Biosci.*, **2012**, *131*(2), 85-93. http://dx.doi.org/10.1007/s12064-011-0136-x PMID: 21979857

[44] Serra, R.; Filisetti, A.; Villani, M.; Graudenzi, A.; Damiani, C.; Panini, T. A stochastic model of catalytic reaction networks in protocells. *Nat. Comput.*, **2014**, *13*(3), 367-377. http://dx.doi.org/10.1007/s11047-014-9445-6

[45] Cawley, G.C.; Talbot, N.L.C. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J. Mach. Learn. Res.*, **2010**, *11*, 2079-2107.

[46] Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, **2011**, *12*, 2825-2830.

[47] Cerami, E.; Gao, J.; Dogrusoz, U.; Gross, B.E.; Sumer, S.O.; Aksoy, B.A.; Jacobsen, A.; Byrne, C.J.; Heuer, M.L.; Larsson, E.; Antipin, Y.; Reva, B.; Goldberg, A.P.; Sander, C.; Schultz, N. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.*, **2012**, *2*(5), 401-404.

[48] Hodges, J.L. The significance probability of the smirnov two-sample test. *Ark. Mat.*, **1958**, *3*, 469-486. http://dx.doi.org/10.1007/BF02589501

[49] Pacheco, M.P.; Bintener, T.; Sauter, T. Towards the network-based prediction of repurposed drugs using patient-specific metabolic models. *EBioMedicine*, **2019**, *43*, 26-27. http://dx.doi.org/10.1016/j.ebiom.2019.04.017 PMID: 30979684

[50] Zampieri, G.; Vijayakumar, S.; Yaneske, E.; Angione, C. Machine and deep learning meet genome-scale metabolic modeling. *PLOS Comput. Biol.*, **2019**, *15*(7), e1007084. http://dx.doi.org/10.1371/journal.pcbi.1007084 PMID: 31295267

[51] Cai, H.Y.; Zheng, V.W.; Chang, K.C.-C. A comprehensive survey of graph embedding: Problems, techniques, and applications. *IEEE Trans. Knowl. Data Eng.*, **2018**, *30*(9), 1616-1637. http://dx.doi.org/10.1109/TKDE.2018.2807452

[52] Hamilton, W.L.; Ying, R.; Leskovec, J. Representation learning on graphs: Methods and applications. *IEEE Data Eng. Bull.*, **2017**, *40*(3), 52-74.

[53] Kriege, N.M.; Johansson, F.D.; Morris, C. A survey on graph kernels. *Appl. Network Sci.*, **2020**, *5*(1), 6. http://dx.doi.org/10.1007/s41109-019-0195-3

[54] Niepert, M.; Ahmed, M.; Kutzkov, K. Learning convolutional neural networks for graphs. *Proceedings of the 33rd International Conference on Machine Learning, ICML 2016*, New York City, NY, USA. June 19-24, **2016**, *Volume 48*, pp. 2014-2023.