Predicting clinical outcome of neuroblastoma patients using an integrative network-based approach

Léon-Charles Tranchevent[1], Petr Nazarov[1], Tony Kaoma[1], Arnaud Muller[1], Sang-Yoon Kim[1], Jagath C. Rajapakse[2], and Francisco Azuaje[1]

[1]Proteome and Genome Research Unit,Department of Oncology, Luxembourg Institute of Health, Luxembourg.
[2]Bioinformatics Research Center, School of Computer Engineering, Nanyang Technological University, Singapore

Abstract

One of the main current challenge in computational biology is to make the best of the huge amount of experimental data that is being produced. For instance, large cohorts of patients are often screened using different high-throughput technologies, effectively producing multiple molecular profiles per patients for hundreds or thousands of patients. We propose and implement a network-based method that integrates such patient omics data and use them to predict various clinical features. Using a neuroblastoma dataset, we then demonstrate that the networks inferred from omics data contain clinically relevant information and that patient clinical outcomes can therefore be predicted using only network topological data.

Introduction

In the last decade, high-throughput technologies have been massively used to study various diseases to decipher the underlying biological mechanisms and to propose novel therapeutic strategies. Initiatives such as The Cancer Genome Atlas have produced and made publicly available a huge amount of omics data from thousands of human samples. These data often correspond to measurements of different biological entities (*e.g.*, transcripts, proteins), represent different views on the same entity (*e.g.*, genetic, epigenetic), and are obtained through different technologies (*e.g.*, microarray, RNA-sequencing). This diversity has motivated the use of integrative strategies that can make sense of these complementary, and sometimes contradictory data. Such integrative strategies have, for instance, been used to define distinct molecular classes of lower-grade gliomas, which exhibit similar pathway perturbations [1].

Another popular research strategy is to represent the data as biological networks, where nodes represent biologically relevant entities (typically genes or proteins) and edges represent relationships between these entities (*e.g.*, regulation, interaction). Network-based methods can then be used, for instance, to define smaller modules within a larger network, or to understand how a biological signal is processed by a network, or to identify key nodes with respect to a biological process of interest. As an example, such network-based approaches have been used to build brain region specific networks from patient expression profiles, and to prioritize genes and gene sets with respect to Alzheimer's disease traits [2]. It is also possible to obtain relevant predictive models by relying on the network topological information, instead of the raw data. An example of such method is Mashup, an approach that summarizes topological information from protein-protein networks to predict functional annotations or genetic interactions, yielding comparable or often even better performance than state of the art methods [3].

Although most biological networks represent gene or protein networks, it is often relevant to represent the data as Patient Similarity Networks (PSN). In these networks, nodes represent patients and edges represent similarities between the patients' profiles. These networks can be used to group patients and to associate these groups with distinct clinical features. It was observed for instance that, within a network obtained by integrating multiple omics data, cancer patient clusters had different clinical outcomes, including different overall survival [4]. Similarly, a network topology based analysis of diabetes patient genotypes has revealed that patients can be clustered in three groups, and that these groups have distinct clinical features, including different comorbidities [5].

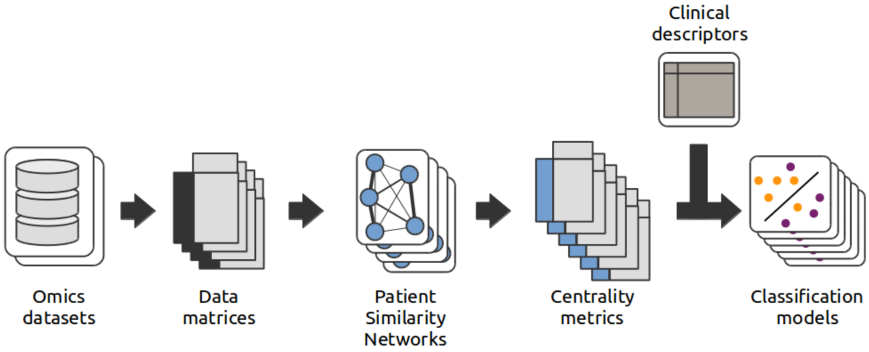

Figure 1: Overall workflow of the proposed strategy. First, the omics datasets are pre-processed and transformed into Patient Similarity Networks. Next, centrality metrics are computed and used to build classifiers (using clinical descriptors as classes to be predicted).

For the current study, we hypothesize that clinically relevant information is encoded within PSN built from omics data. To investigate whether we can use this topological information to predict patient clinical outcome, we analyze a neuroblastoma dataset that contains gene expression data, genotype data, and clinical descriptors. Our approach is similar to a previous analysis in which classifiers built from these gene expression data were used to predict clinical outcome [6]. However, our approach is still different since we transform the omics data into networks, and then train the classifiers with network topological data, instead of training the classifiers directly with omics data. Several classifiers are used to predict distinct clinical descriptors such as 'Disease progression' and 'Death from disease'. Our results indicate that the performance of classifiers trained with topological data is at least comparable to the performance of the models built on the omics data directly, and in some cases better. Altogether, our network-based approach represents therefore a novel and complementary strategy to analyze and integrate large collection of omics data.

Results and discussion

We propose a network-based method to analyze and integrate omics data that relies on the topological properties of the networks derived from these data (see Figure 1). More precisely, the omics data are first transformed into networks, then centrality metrics are computed for all network nodes (*i.e.*, representing the patients) and used as input for classification models (using clinical features to define classes). To validate our strategy, we have defined different settings (see below) and evaluated the classification performance using the Matthews Correlation Coefficients (MCC, see Material and Methods).

We have compared the performance of the classification models when inputed with omics data (*hereinafter* classical) or with network centrality values (*hereinafter* network-based), regardless of the other parameters. Our results indicate that the performance of both strategies lies within the same range. However, when looking at the shape of the distributions, we can observe that the performance of network-based models is consistently higher than the performance of the classical models. The strongest difference is observed for the 'Death from disease' clinical feature (median MCC of 0.30 and 0.09 for network-based and classical models respectively). The same observation can be made for the other clinical features 'Disease progression' and 'Risk status', although the differences between the median MCC are smaller (see Figure 2A). As a control, we have performed the same comparison for the 'gender' feature that is not correlated to the other clinical features (maximum Pearson's correlation coefficient is 0.04 with 'Death from disease'). The results are inversed (*e.g.*, median MCC of 0.38 and -0.03 for classical and network-based models respectively using a linear discriminant analysis), indicating that the network-based approach is less biased by the gender of the patients (see Figure 2B). Conversely, the classical models are able to better predict the gender of the patients because they rely on the full omics data, which include expression data from loci that are located on the sex chromosomes.

We have then investigated whether the parameters of the network-based approach can influence its classification performance. We first compare the different dimension reduction strategies and network inference methods (see Material and Methods). In both cases, we do not observe significant variations in the classification performance, for any of the clinical endpoints, and any of

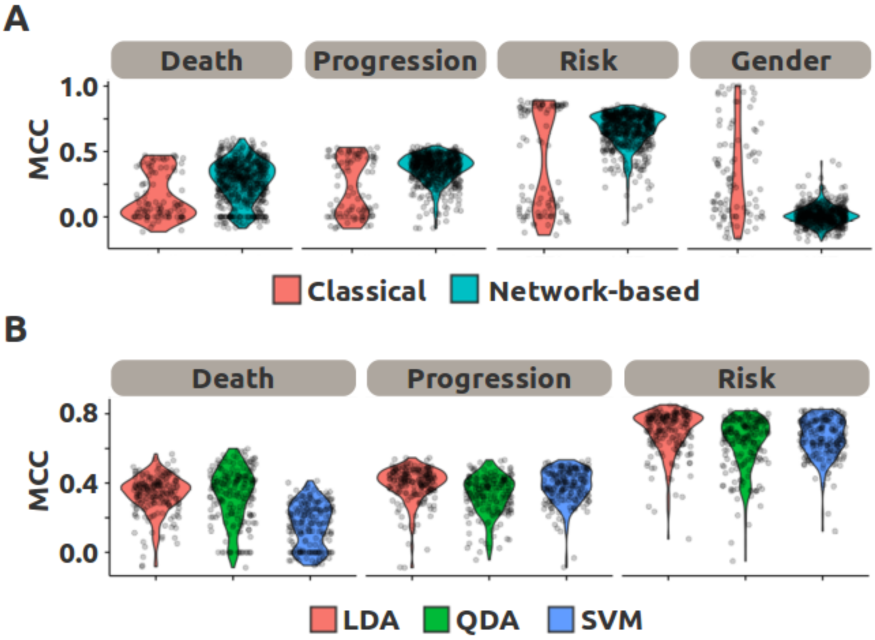

Figure 2: Performance of the classification models in different settings. (A) Violin plots of the estimated classification performance (MCC values) for the classical and network-based approaches, and for the different clinical endpoints (including the patient gender). (B) Violin plots of the estimated classification performance (MCC values) for the different algorithms and clinical endpoints.

the algorithms, indicating that our approach is stable to data transformation and normalization. We have then performed a comparison of different centrality metrics. Once again, we observed that the performance of the classification models remains similar for all clinical endpoints and algorithms when different centrality metrics are used. Altogether, these results seem to indicate that the different centrality metrics are essentially able to capture the same signal from the networks, regardless of the other parameters. However, it is important to notice that most centrality metrics are positively correlated, and it was therefore expected that they would produce similar classification results.

We have then computed the correlation between the MCC values of the different clinical endpoints (regardless of the other parameters). We observe that in general the correlation at the performance level is rather high (Pearson's correlation coefficients between 0.58 and 0.85), indicating that models that perform well for one clinical endpoint, are more likely to also perform well for the other clinical endpoints. However, the correlation between the clinical features themselves is in general lower (Pearson's correlation coefficients between 0.48 and 0.67). This indicates that the rather high correlation at the performance level cannot only be explained by the underlying correlation at the feature level. However, it should be noted that the best models for each clinical endpoint are different, indicating that no model is always superior to all other models. In particular, there is not a single classification algorithm that outperforms the other two on all clinical endpoints (Figure 2B).

To conclude, we have designed a method that uses the topological information encoded within patient similarity networks to predict clinical outcome. The results indicate that our network-based method can be complementary to existing methods. We are currently extending the validation on the neuroblastoma dataset by selecting the omics features that correlate with the clinical endpoints of interest (similarly to a previous analysis of the same dataset [6]), and by making better use of the available genomic data.

Material and Methods

Data collection

The twelve datasets were collected on the 28th of February 2017 from GEO[1] (eleven expression datasets obtained from the series GSE49710, GSE49711 and GSE62564), and from the BOKU website[2] as specified in the CAMDA guidelines[3] (single aCGH dataset). The clinical descriptors have been extracted from the above mentioned datasets and uniformized manually to keep only six clinical descriptors. Multi-class descriptors (*e.g.*, 'Stage') have been split into multiple binary descriptors for binary classification.

Data preparation

The eleven expression datasets contain pre-processed profiles for 498 samples, corresponding to 498 patients. For aCGH, we have extracted the 185 samples, corresponding to 145 patients for which we also have expression data. Since the aCGH data were produced by different laboratories and using different arrays, the data have first been filtered to keep only the genomic features that are shared by all platforms, and further normalized by correcting for the potential lab, platform, and batch effects. For all datasets, features with at least one missing point are dropped prior to the network inference step. For each dataset, we then derive three data matrices by either keeping all features (no dimension reduction) or by applying two different dimension reduction strategies: (i) keeping only the 20% most varying features, (ii) keeping the PCA based pseudo-features that explain more than 90% of the variance.

Network inference

Each data matrix is then used to infer two Patient Similarity Networks (PSN) by using two slightly different inference methods. In both cases, the Pearson correlation coefficients between all patient pairs are computed. Then, these correlation coefficients are rescaled to represent positive edge weights using either (i) a simple rescaling strategy, or (ii) a Weighted Correlation Network Analysis (WGCNA) based strategy that enforces scale-freeness of the associated network. Both approaches are summarized by Equation 1).

$$w_{a,b} = \left(\frac{c_{a,b} - \min(C)}{\max(C) - \min(C)} \right)^{\beta} \tag{1}$$

with $w_{a,b}$ the edge weight between the nodes representing the patients a and b, $c_{a,b}$ the correlation between the molecular profiles of patients a and b, C the set of all correlations, and β the parameter that controls the scale-freeness of the network. For a simple rescaling, we have set β to one. Alternatively, and as recommended previously, for the WGCNA-based strategy, we have used the smallest β that gives a truncated scale-free index of at least 90% (for our networks, $\beta \in \{2, 4, 6, 8\}$).

Centrality metrics

For each network, we then compute six centrality metrics: weighted degree, closeness centrality, current flow closeness centrality, current flow betweenness centrality, eigen vector centrality, and Katz centrality. All centrality metrics are then individually standardized to a zero mean and a unit standard deviation. These values are then considered as features that can be used for classification.

Classification algorithms

Class definitions have been extracted from the clinical descriptors provided with the omics data. Several classification algorithms have been considered, including Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and Support Vector Machine (SVM), using dedicated R and Python libraries. In addition, we have considered several classification scenarios by varying the number of data sources, networks and centrality metrics used (see Table 1). As a control, we have also built classifiers using the original omics data (without any network inference). For convenience, we have used the same train and test stratified split than a previous study of the same data [6]. The performance of the classifiers on the test data is estimated using the

[1] https://www.ncbi.nlm.nih.gov/geo/
[2] http://ala.boku.ac.at/camda2017/NB/
[3] http://camda2017.bioinf.jku.at/doku.php/contest_dataset

Data sources
Microarray
RNA-seq (genes, MAV)
RNA-seq (transcripts, MAV)
RNA-seq (junctions, MAV)
RNA-seq (genes, TAV)
RNA-seq (transcripts, TAV)
RNA-seq (junctions, TAV)
RNA-seq (genes, TUC)
RNA-seq (transcripts, TUC)
RNA-seq (junctions, TUC)
RNA-seq (transcripts, TUC-RE)

Network inference
Simple rescaling
WGCNA

Dimension reduction
None
Variance-based
PCA-based

Centralities
Weighted degree
Closeness
Current flow closeness
Current flow betweenness
Eigen vector
Katz

Table 1: Lists of the possible values for the four parameters of the network-based method. There are 11 data sources covering two technologies, distinct biological entities (genes, transcripts, junctions), and several mapping strategies (MAV, TAV, TUC, TUC-RE, see [6] for details). In addition, we have defined 2 network inference methods, 3 dimension reduction policies, and 6 centrality metrics. In total, there are therefore 396 features available for classification.

classification accuracy and the Matthews Correlation Coefficient (MCC), similarly to a previous analysis of these data [6].

Acknowledgments

Project supported by the Fonds National de la Recherche (FNR), Luxembourg (SINGALUN). We thank Ayasdi for access to its Topological Data Analysis platform[4], and the Fisher lab for sharing the neuroblastoma data.

References

[1] The Cancer Genome Atlas Research Network. Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas. *New England Journal of Medicine*, 372(26):2481–2498, June 2015.

[2] Minghui Wang, Panos Roussos, Andrew McKenzie, Xianxiao Zhou, Yuji Kajiwara, Kristen J. Brennand, Gabriele C. De Luca, John F. Crary, Patrizia Casaccia, Joseph D. Buxbaum, Michelle Ehrlich, Sam Gandy, Alison Goate, Pavel Katsel, Eric Schadt, Vahram Haroutunian, and Bin Zhang. Integrative network analysis of nineteen brain regions identifies molecular signatures and networks underlying selective regional vulnerability to Alzheimer's disease. *Genome Medicine*, 8(1), December 2016.

[3] Hyunghoon Cho, Bonnie Berger, and Jian Peng. Compact Integration of Multi-Network Topology for Functional Analysis of Genes. *Cell Systems*, 3(6):540–548.e5, December 2016.

[4] Bo Wang, Aziz M Mezlini, Feyyaz Demir, Marc Fiume, Zhuowen Tu, Michael Brudno, Benjamin Haibe-Kains, and Anna Goldenberg. Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods*, 11(3):333–337, January 2014.

[5] L. Li, W.-Y. Cheng, B. S. Glicksberg, O. Gottesman, R. Tamler, R. Chen, E. P. Bottinger, and J. T. Dudley. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Science Translational Medicine*, 7(311):311ra174–311ra174, October 2015.

[6] Wenqian Zhang, Ying Yu, Falk Hertwig, Jean Thierry-Mieg, Wenwei Zhang, Danielle Thierry-Mieg, Jian Wang, Cesare Furlanello, Viswanath Devanarayan, Jie Cheng, Youping Deng, Barbara Hero, Huixiao Hong, Meiwen Jia, Li Li, Simon M Lin, Yuri Nikolsky, André Oberthuer, Tao Qing, and Zhenqiang Su *et al.* Comparison of RNA-seq and microarray-based models for clinical endpoint prediction. *Genome Biology*, 16(1), December 2015.

[4] www.ayasdi.com