BITS 2017

Modeling biomolecular profiles in a graph-structured sample space for clinical outcome prediction with melanoma and ovarian cancer patients.

J. Gliozzo[1], M. Notaro[2], A. Petrini[2], P. Perlasca[2], M. Mesiti[2], E. Casiraghi[2], M.Frasca[2], G. Grossi[2], M. Re[2], A. Paccanaro[3], G. Valentini[2]

[1] Fondazione IRCCS Ca' Granda - Ospedale Maggiore Policlinico, Università degli Studi di Milano
[2] AnacletoLab – Dipartimento di Informatica, Università degli Studi di Milano
[3] Centre for Systems and Synthetic Biology & Department of Computer Science, Royal Holloway, University of London.

Motivation

Phenotype and outcome prediction using a set of selected biomarkers (e.g. gene expression signatures or allelic configurations of SNPs) are well-established problems in the context of computational biology. State-of-the-art methods are largely based on inductive supervised models that use selected biomarkers to predict the phenotype or outcome of interest, and several works showed the effectiveness of these methods. Nevertheless supervised inductive models do not explicitly take into account the functional or the genetic relationships between individuals, in the sense that they usually use vectors of selected biomarkers to discriminate between patients and do not directly exploit the existing relationships between them.

Recently the emerging discipline of “Network Medicine” opened a new “systemic” approach to unravel the molecular mechanisms underlying diseases, by analyzing the functional relationships between bio-molecular entities (i.e. proteins, genes, metabolites) in the “biomarker space” with the aim, e.g., of ranking genes with respect to a given phenotype or disease.

In this work we introduce a novel “Network Medicine”-based approach in which biomolecular profiles of patients are modeled in a graph-structured “sample space” instead of the “biomarker space”.

Our aim is to transfer the systemic approach usually applied to analyze networks of biomolecules in the context of networks of samples/patients constructed relying on the similarities between the biomolecular profiles of patients.

Methods

From a machine learning standpoint this problem can be modeled as a semi-supervised node label ranking prediction problem in a graph, where samples are nodes, edges functional relationships between molecular profiles and the labels represent the clinical outcome or more in general a phenotypic variable to be predicted.

We construct networks of patients on the basis of their functional or genetic similarities, and then we apply a semi-supervised transductive method to predict the phenotype or the clinical outcome of patients.

To this end we propose a novel network-based semi-supervised learning algorithm Sample-Net (*S-Net*) that exploits the relationships between samples coded in the network and the a priori knowledge available for a subset of samples (patients) to predict the clinical outcome of patients.

At first *S-Net* computes the functional similarity matrix between samples by using the correlation between the biomolecular profiles of each sample. The resulting similarity matrix *S* represents a basic adjacency matrix for the graph of samples and then a graph-kernel (e.g. a random walk kernel) can be applied to *S* to enrich the original graph with new edges according to the topological characteristics of the graph itself (Fig.1). This enriched “kernelized” graph is then used to infer the phenotypic variable of interest associated with each sample (node) by adopting simple local learning strategies based on the guilt-by-association principle. We note that even if we use simple local learning strategies, global learning strategies are implicitly applied through the graph-kernels.

Results

We applied S-Net to two different publicly available datasets of patients afflicted with a specific type of tumor: melanoma and ovarian cancer. We show that network-based methods in the sample space achieve results competitive with classical supervised inductive systems. Moreover the graph representation of the samples can be easily visualized (Fig.1), and can be used to gain visual clues about the relationships between samples, taking into account the phenotype associated or predicted for each sample. To our knowledge this is the first work that proposes graph-based algorithms working in the kernelized sample space of the biomolecular profiles of the patients to predict their phenotype or outcome, thus contributing with a novel research line in the framework of the Network Medicine.

References

[1] G. Valentini, G. Armano, M. Frasca, J. Lin, M. Mesiti and M. Re RANKS: a flexible tool for node label ranking and classification in biological networks, *Bioinformatics*, 32(18), 2016.
[2] G. Valentini, A. Paccanaro, H. Caniza, A. Romero, M. Re, An extensive analysis of disease-gene associations using network integration and fast kernel-based gene prioritization methods, *Artificial Intelligence in Medicine*, Volume 61, Issue 2, pages 63-78, 2014.
[3] M. Re, and G. Valentini, Network-based Drug Ranking and Repositioning with respect to DrugBank Therapeutic Categories, *IEEE ACM Transactions on Computational Biology and Bioinformatics* 10(6), pp. 1359-1371, 2013
[4] M. Re, M. Mesiti and G. Valentini, A Fast Ranking Algorithm for Predicting Gene Functions in Biomolecular Networks, *IEEE ACM Transactions on Computational Biology and Bioinformatics* 9(6) pp. 1812-1818, 2012
[5] M. Re and G. Valentini Cancer module genes ranking using kernelized score functions, *BMC Bioinformatics* 13 (Suppl 14): S3, 2012

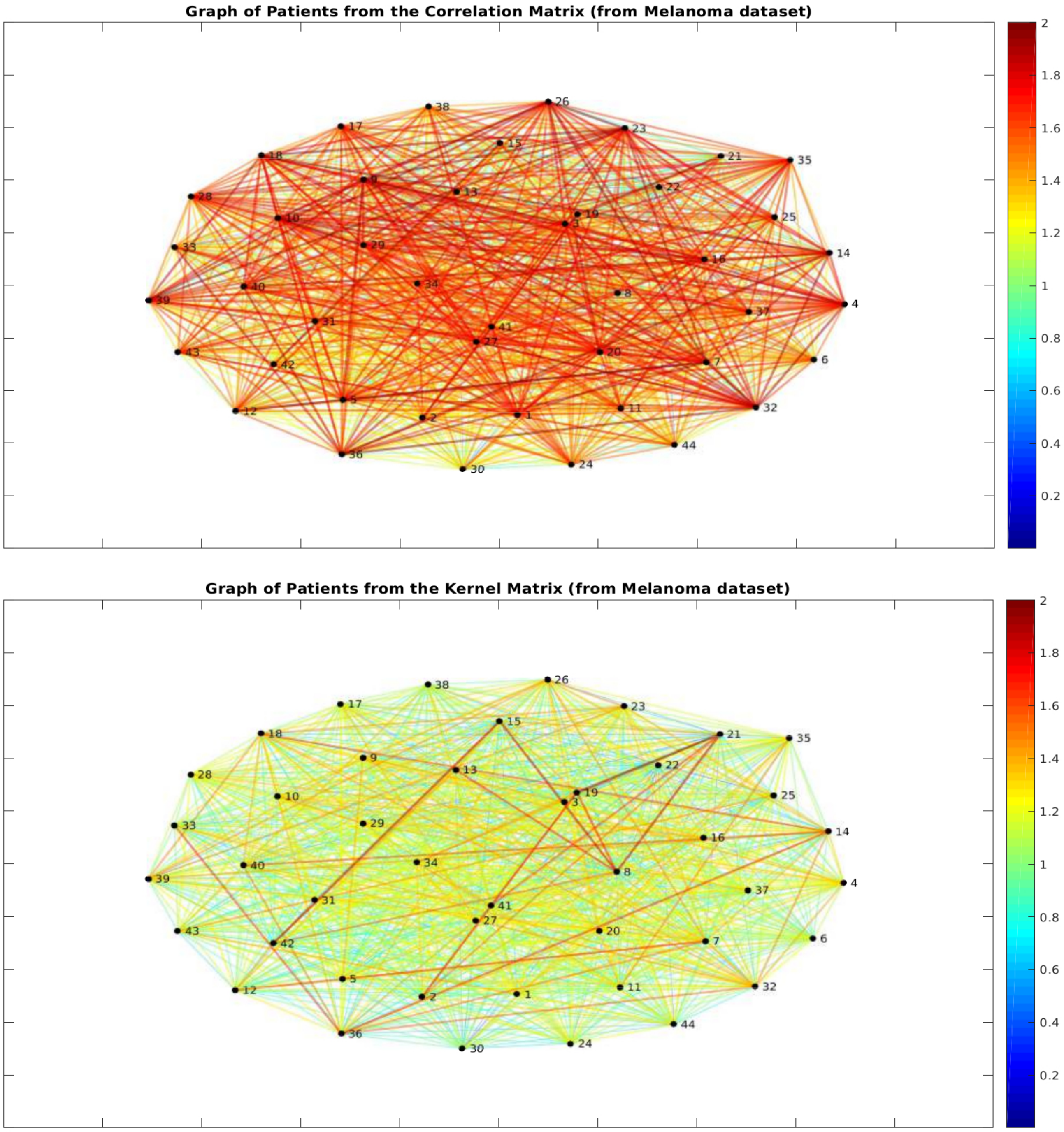

Fig.1. Graphs of the sample (patients) obtained from the correlation of their molecular profiles (top) and after the application of a random walk kernel (bottom) with the Melanoma data set. The colour and thickness of the edges represent the weight of the corresponding edge: higher is the weight higher is the thickness and the colour is closer to red. The application of the random walk kernel make more clear the functional relationships among patients: strongest (red) edges are more sparse and connect each patient (node) only to specific other patients.