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

Background

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*.
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.

Aims

Introduction of 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 goal 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

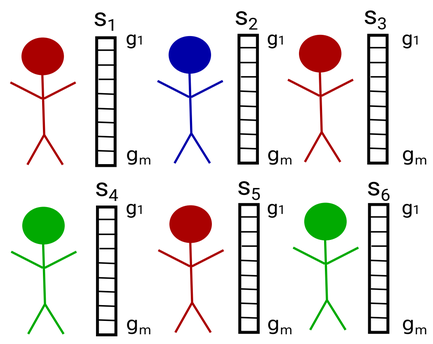
From a machine learning standpoint this problem can be modeled as a semi-supervised node label ranking prediction problem in a graph, where samples are nodes, edges functional relationships between molecular profiles and the labels represent the clinical outcome/ phenotypic variable to be predicted.

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.

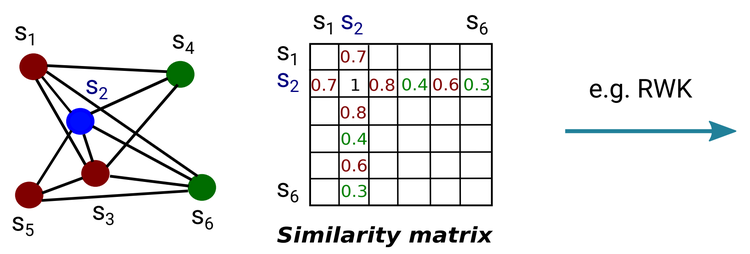
The main advantages of this proposed approach with respect to classical inductive supervised methods are:
a) the construction of a network in the "sample space" allows to exploit relational information between the molecular profiles of patients;
b) unlabeled examples can be used in the learning process;
c) both local and global semi-supervised network-based learning strategies are applied to rank the samples.

Steps of S-Net

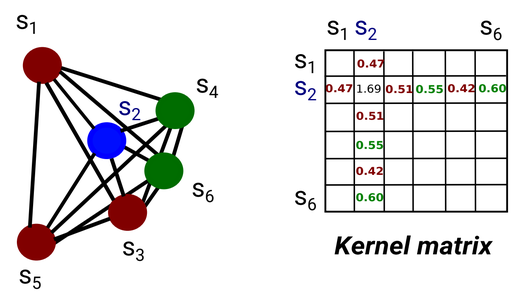
1. Each sample has an associated biomolecular profile



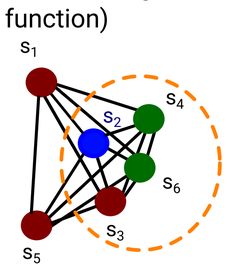
2. *S-Net* computes the functional similarity matrix S between samples by using the correlation between the biomolecular profiles of each sample



3. A graph kernel (e.g. a random walk kernel) is applied to enrich the original graph with new edges according to the topological characteristics of the graph itself



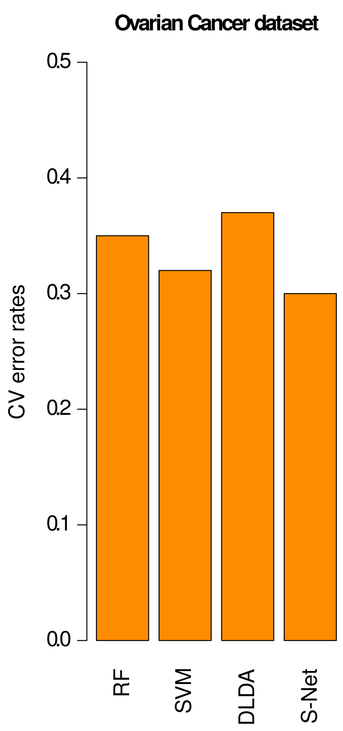
4. The "kernelized" graph is used to infer the phenotypic variable of interest associated with each sample (node) by adopting simple local learning strategies (e.g. k- Nearest Neighbour score function)

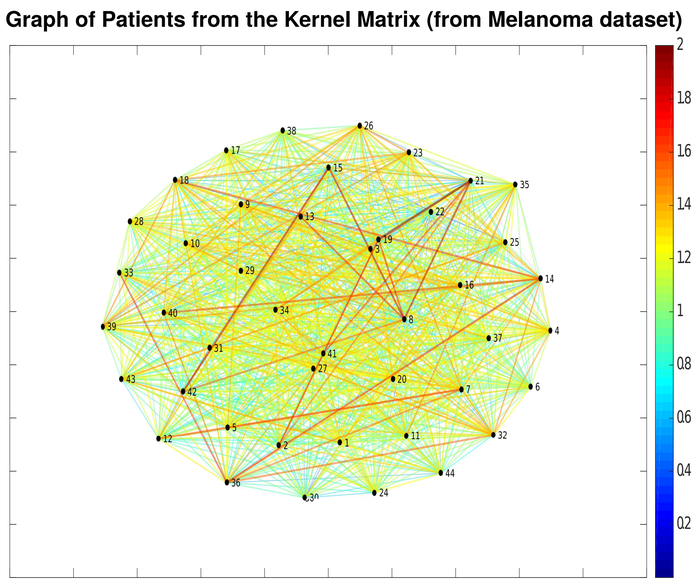


Rank	Score	Predicted output
S_1	0.902	Red
S_3	0.856	Red
S_5	0,807	Red
S_6	0.521	Green
S_4	0.471	Green
S_2	**0.348**	**Green**

Results

We applied S-Net to two different publicly available datasets of patients afflicted with a specific type of tumor (melanoma and ovarian cancer) and we compared S-Net to classical supervised methods (Left figure). Moreover, we can easily visualize the graph representation of the samples (Right figure), where 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.





Conclusions

a) S-Net achieves **results competitive** with classical supervised inductive systems.
b) The **graph representation of the samples** can be easily visualized, and can be used to gain visual clues about the relationships between samples, taking into account the phenotype associated or predicted for each sample.
c) 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.

For further information

Please, feel free to contact us:
valentini@di.unimi.it
jessica.gliozzo@gmail.com

Anacleto Lab website (scan the **QR code** if you wish): **https://sites.google.com/site/anacletolaboratory/**