Genes prioritization with respect to Cancer Gene Modules using functional interaction network data.

M. Re and G. Valentini
DSI – Dip. Scienze dell'Informazione, Università degli Studi di Milano

Abstract

The classification of genes as belonging or not to Cancer Gene modules (CGMs) can help in shedding light on bio-molecular mechanisms involved in the onset and progression of many types of tumors and is also able to open novel research directions for diagnostic, prognostic and therapeutic studies. In this contribution we propose a novel method suitable for CGMs membership prioritization in Functional Interaction networks. The proposed method was evaluated on previously published datasets and compares favorably with other state-of-the-art methods.

Introduction

In the last decade advances in high-throughput biotechnologies allowed researchers to investigate at system level the bio molecular alterations leading to or supporting the progression of important pathologies such as tumors. Aberrations in proliferation and survival pathways are common characteristics shared by all tumors but alterations interfering with other classes of pathways are specific to certain cancer types. Popular methods for the characterization of the variations in the expression patterns of single genes in cancers w.r.t. normal tissues are based on DNA microarray technology, but this approach suffers of important limitations because the comparison of the expression patterns detectable in two conditions only (i.e. a specific type of tumor and a generic control) is able to provide useful information about specific sets of genes in the cancer type under investigation but is unable to address the variations and commonalities of expression patterns in different types of tumor.

In order to overcome this important limitation Segal and colleagues [1] analyzed a 'cancer compendium' of expression profiles collected in 26 studies reporting the expression of 14,145 genes in 1,975 arrays spanning 22 tumor types. The analysis of this compendium led to the definition of 454 sets of genes (hereafter referred to as Cancer Gene Modules, CGM) in which the genes act in concert to carry out specific functions. The results reported in [1] highlighted that while the activation or inactivation of some CGMs is observed across all the investigated tumor types and clinical conditions (suggesting the existence of common tumor progression mechanisms) the alteration of other CGMs expression patterns are signatures specific of particular tumors. The availability of this collection of CGMs opened novel research directions for diagnostic, prognostic and therapeutic studies.

About seven years after the publication of the aforementioned CGM collection other types of information, besides expression data, have proven to be effective in cancer data analysis. A crucial problem to face in order to allow the most effective use of the vast amount of bio molecular data at today available in the public domain is to find a biologically meaningful way to integrate structurally different types of data. In [2] Wu and colleagues constructed a Functional Interaction network (FI network, hereafter) by integrating several protein-protein interactions networks,

proteins domains interactions data, functional co-annotations data (in terms of Gene Ontology BP terms sharing) and normal tissue co-expression profiles. In order to lower the noise in this large high-coverage FI network the authors filtered the functional interactions by means of a Naïve Bayes classifiers trained on the functional interactions stored in the Reactome database. This make the FI networks informative w.r.t. the task of ranking the genes according to their likelihood to belong to a particular CGM, because curated pathways datasets are also involved in the definition of the CGMs reported in [1]. The approach adopted in [2] is also appealing because it provides a principled way to integrate diverse types of data while keeping under control the often high levels of noise characterizing network-structured representation of bio molecular data.

In this contribution we propose a method for node ranking in an undirected network according to their likelihood to belong to a collection of predefined gene sets. This method is based on scores obtained by using kernelized similarity score functions. The performance of the proposed approach are compared with the ones achieved by a state-of-the-art node label prediction algorithm, GeneMANIA [3] which was reported to be among the best performing methods in a public large scale computational biology challenge [4].

Methods

The FI network is represented as a graph G = (V;E), where the nodes V represent genes and the edges E represent some notion of strength of the link between a pair of genes (this could be realized by mean of weights associated to each edge and expressing the strength of the functional interaction between the genes). Once constructed the graph we can simply label a subset of nodes as positives" (i.e. all the genes indicated belonging to a specific CGM) and then propagate the labeling to the "negative" nodes in the graph that are closer (in a topological sense) to the positives.

Our algorithm is based on the notion of gene functional similarity in a graph representing the relationships between genes and on the hypothesis that the genes that are most similar w.r.t. a positive set of genes could be involved in the realization of the same biological process carried out by the members of the positive set. It is easy to understand that crucial points in this process are the choice of the entities (vertexes) composing the graph and also the appropriate definition of the notion of similarity (functional interaction strength, in this case) adopted to compute the weights associated to the edges. We designed score functions to rank genes according to their likelihood to belong to a specific Cancer Gene Module by exploiting kernelized similarities between genes in the underlying FI network.

The proposed gene prioritization algorithm is fast (approximately linear with respect to the number of genes/nodes of the network) and scales nicely with the dimension of data.

Results and discussion

The 9,448 genes in the FI network proposed in [2] have been labeled according to their membership to the CGMs identified in [1]. We then filtered out all the CGMs composed by less than 20 genes obtaining a final CGM list composed by 297 modules. From this set of CGMs we randomly selected 15 CGMs composed by a number of genes ranging from 20 to more than 300 genes. The genes were finally ranked according to their likelihood to belong to each of the considered CGMs using both GeneMANIA and the proposed method. Performances were evaluated according to a

canonical 5 folds cross validation scheme using the area under the ROC curves (AUC) and the Precision collected at fixed Recall levels. Figure 1 shows the average precision results across the 15 CGMs for different recall rates.

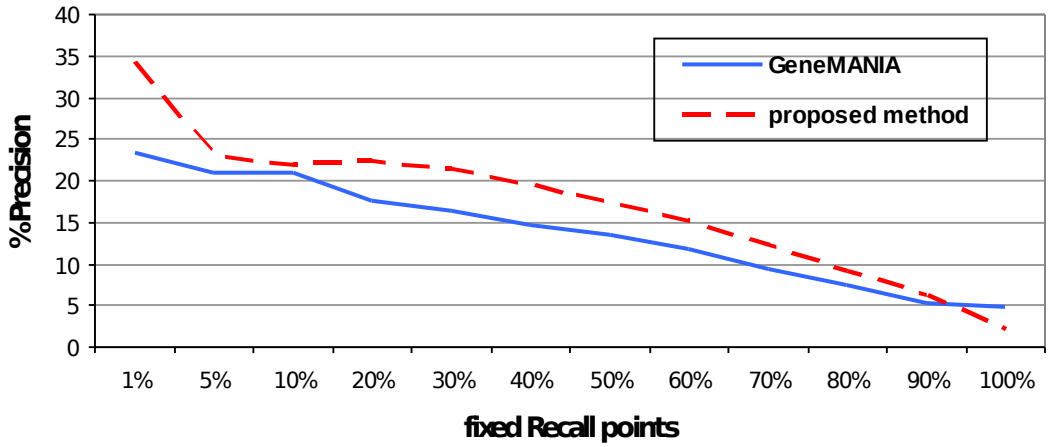

Figure 1

While the average AUC across the considered CGMs is lower when compared with the one obtained by the GeneMANIA algorithm, we observe that the proposed method compares favorably in terms of average Precisions at fixed Recall levels.

It is also worth noting that a critical issue in ranking problems involving large networks is represented by the computational efficiency. This is among the reasons leading to our choice of GeneMANIA as baseline, one of the fastest state-of-the-art node label ranking algorithms available in the literature, designed to perform gene categorization in whole genomes. The entire experiment was performed by GeneMANIA in about 53 seconds while the computational time required by our method was 12 seconds, using in both cases a 64 bit Intel quad core i7 CPU 860 processor (2.80 GHz) with 16 GB RAM.

References:

[1] Segal E., Friedman N., Koller D. and Regev A., “A module map showing conditional activity of expression modules in cancer”, Nature Genetics, 36(10), 2004

[2] Wu G., Feng X. and Stein L. “A human functional protein interaction network and its application to cancer data analysis”, Genome Biology 11:R53, 2010

[3] Mostafavi S., Debajyoti R., Warde-Farley D., Grouios C. and Morris Q., "GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function", Genome Biology, vol. 9, Supp. 1, 2008

[4] Peña-Castillo L. et al., "A critical assessment of Mus musculus gene function prediction using integrated genomic evidence", Genome Biology, Vol.9, Supp.1, 2008