COMPARISON OF EARLY AND LATE OMICS DATA INTEGRATION FOR CANCER MODULES GENE RANKING

Matteo Re[1], Marco Mesiti[1] and Giorgio Valentini[1]

[1]Bioinformatics laboratory, Department of Computer Science, University of Milan, Italy
e-mail: re@di.unimi.it, mesiti@di.unimi.it, valentini@di.unimi.it (corresponding author).

Motivation and Objectives

In a recent work we evaluated the ability of semi-supervised learning methods based on random walks to rank genes with respect to Cancer Modules (CM) using networks constructed from different sources of information (Re and Valentini, 2012). The performance of this approach was evaluated using a relatively simple data integration scheme consisting in the unweighted sum of the adjacency matrices of the biomolecular networks involved in our experiments.

Despite the achievement of good performances, our tests were all based on a network integration approach applied before the gene prioritization phase (early data integration). Recently published works demonstrated that good results can also be obtained by performing the integration step after the production of a prioritization ranking for each available dataset (late data integration), through the integration the ranking vectors (Kolde et al., 2012).

The aim of this contribution is to compare prioritization performances on CM genes using early and late data integration methods in order to highlight benefits and potential pitfalls characterizing these approaches when applied in large scale gene prioritization problems.

Methods

In our experiments we applied random walk (RW) and random walk with restart (RWR), to rank genes with respect to CMs.

All the methods described below refer to an undirected weighted graph, with nodes corresponding to genes and edges expressing similarities between them according to a specific data source. CMs were originally defined (Segal et al. 2004) only using gene expression, but it is commonly accepted that biomolecular alterations involved in onset and progression of tumors cannot be detected only in terms of gene expression alterations. This motivates the usage of gene networks derived from the integration of different (and complementary) types of evidences ranging from protein-protein interactions to evolutionary dynamics detected in the form of protein domains conservation between the protein products of genes and many others. In this work we use two gene networks: a predicted functional gene interaction network, FInet (Wu et al., 2010) and a functional interaction network, HumanNet, enriched using comparative genomics methods (Lee et al., 2011).

We adopted “early” and “late” integration strategies to combine the functional networks. For “early” integration we mean integration of data at network level, before the application of ranking algorithms to prioritize genes. “Late” integration means combination of the rankings obtained from different networks, i.e. “merging” the ordered lists of genes obtained after the application of the ranking algorithms to each separate network.

As an example of “early” integration we considered the unweighted sum of the adjacency matrices of Finet and HumanNet, while for the “late” integration we adopted the Robust Rank Aggregation algorithm (RRA), recently proposed for gene list integration problems (Kolde et al., 2012).

RRA is a statistical method based on a null model describing distributions of ranks when all the prioritization lists to be integrated are uninformative; it estimates the statistical significance of genes deviating from the null model. This allows RRA to focus on the significance of single genes instead of on the overall significance of the ranking lists to be integrated.

Results and Discussion

In our preliminary experiments we considered functional networks composed by about 8,500 human genes, and 159 CMs, selected by the entire set of CMs by choosing only those with at least 20 and no more than 100 genes.

In our experiments we produced two prioritization lists (one obtained by evaluating FInet and one obtained evaluating HumanNet) for each considered CM, using both RW and RWR algorithms, and considering different values for the restart probability parameter (Pr). We then used RW and RWR to prioritize genes w.r.t. the CMs using a network representing the unweighted sum of the adjacency matrices of Finet and HumanNet. The gene prioritization lists obtained using the integration network have been compared with the ones obtained by integrating with RRA the rankings produced for each CM using FInet and HumanNet separately.

Table 1 shows the experimental results in terms of the Area Under the ROC curve (AUC) averaged across all the considered 159 CMs. Performances for each CM have been computed using a standard stratified 5 folds cross validation scheme.

	FInet	**HumanNet**	**Early int.**	**Late int.**
RW	0.7799	0.7349	0.7539	0.8043
RWR, Pr=0.8	0.8040	0.8361	0.8650	0.8606
RWR, Pr=0.9	0.8065	0.8347	0.8645	0.8608

Table1: Average AUC: Comparison of the prioritization performance of the early and late integration methods with the ones obtained using Finet and HumanNet separately.

Results show that both network integration (Early int.) and ranking integration (Late int.) significantly improve the average AUC with respect to single networks/rankings, independently of the applied ranking algorithm, but it is unclear whether in this context should be preferable an early or late integration approach. For this reason we need to experiment with different types of early and late integration methods, as well as with different ranking algorithms and more sources of omics data to obtain an overall picture of these different approaches to data integration for network-based methods in biomedicine.

Acknowledgements

The authors gratefully acknowledge financial support by Università degli Studi di Milano.

References

- Re M and Valentini G (2012) Cancer module genes ranking using kernelized score functions, *BMC Bioinformatics* **13** (Suppl 14):S3, doi:10.1186/1471-2105-13-S14-S3
- Segal E, Friedman N, et al. (2004) A module map showing conditional activity of expression modules in cancer, *Nat Genet* **36**:1090-1098, doi:10.1038/ng1434
- Wu G et al. (2010), A human functional protein interaction network and its application to cancer data analysis. *Genome Biology*, **11**:R53, doi:10.1186/gb-2010-11-5-r53
- Lee I, Blom U et al (2011) Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res* **21**:1109-1121, doi:10.1101/gr.118992.110
- Kolde R, Laur S et al. (2012) Robust rank aggregation for gene list integration and meta-analysis *Bioinformatics* **28** (4): 573-580. doi:10.1093/bioinformatics/btr709