Network integration boosts disease gene prioritization

Giorgio Valentini[1], Alberto Paccanaro[2], Horacio Caniza Vierci[2], Alfonso E. Romero[2], Matteo Re [1]
[1] Dipartimento di Informatica, Università degli Studi di Milano, Italy
[2] Dept. of Computer Science and Centre for Systems and Synthetic Biology, Royal Holloway, London, UK

Summary: We propose and systematically compare different gene network combination algorithms to experimentally assess the effect of network integration in the context of gene prioritization. An extensive application of network-based gene prioritization methods to 725 MeSH diseases shows how to apply network integration to find novel candidate disease genes predicted with high accuracy and reliability.

Network medicine is emerging as a systemic approach to unravel the molecular mechanisms underlying diseases. In this context, gene prioritization methods have progressed rapidly with the aim of discovering candidate "disease" genes by exploiting the large amount of available "omics" data covering different typesof relationships between genes [1]. However, despite the availability of works describing specific combinations of datasets to develop tools suitable for disease genes prioritization, "our understanding of how to perform useful predictions using multiple data sources or across biological networks is still rudimentary" [1], and in particular, to our knowledge, no systematic studies focused on the comparison of different network integration methods.
To contribute to fill this gap, we propose, compare and analyze different network integration strategies to combine multiple gene networks constructed with different sources of single or heterogeneous data. In particular, we propose: (a) simple unweighted integration methods that combine gene networks solely on the basis of the structural characteristics of the nets; (b) weighted integration methods that combine networks according to the "predictiveness strength" of each type of network, estimated through the assessment of the accuracy of the learning algorithm trained on the networks themselves. We also investigated the issue of the choice of the "seed genes" to characterize the diseases involved in the gene prioritization analysis. In order to extend the analysis to a larger set of diseases, not limited to genetic disorders, we used associations between "seed genes" and MeSH (Medical Subject Headings) diseases downloaded from the Comparative Toxicogenomic Database.

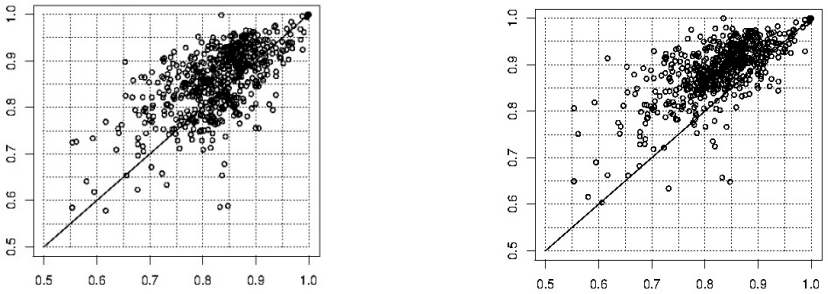

Fig.1: Each point represents the AUC score obtained through network integration methods (ordinate) and with the best single network (abscissa) on each of the 725 MeSH diseases. Left: unweighted integration; right: weighted integration.

The proposed network integration methods are general enough to be applied with both classical and recently proposed [2] gene prioritization methods. We compared 6 different network integration algorithms combining 9 different gene networks, including also semantic similarity-based nets constructed on the basis of the GO annotations of genes [3] (Fig.1). Cross-validated results with 725 MeSH diseases show that to boost gene prioritization we need: a) network integration methods, able to learn from the data how to combine different gene interaction networks ; b) gene prioritization algorithms able to exploit the overall topology of the network.
In particular, for 25 MeSH disorders for which we obtained a cross-validated AUC > 0.975 (p-value < 0.01), we found about 70 novel associations between genes and disease MeSH descriptors.

References

[1] Y. Moreau, L. Tranchevent, Computational tools for prioritizing candidate genes: boosting disease gene discovery, Nature Rev. Genet. 13 (2012) 523–536.
[2] M. Re, G. Valentini, Cancer module genes ranking using kernelized score functions, BMC Bioinformatics 13 (2012) S3
[3] H. Yang, T. Nepusz, A. Paccanaro, Improving GO semantic similarity measures by exploring the ontology beneath the terms and modelling uncertainty, Bioinformatics 28 (2012) 1383–1389