Network-based gene-disease prioritization using PROPHNET

Víctor Martínez✉, Carlos Cano, Armando Blanco
Department of Computer Science and A.I., University of Granada, Granada, Spain

Motivation and Objectives

A major goal in biomedicine is to determine the underlying genetic causes of human diseases in order to better understand them and support their prevention and treatment. However, the genetic bases of many multifactorial diseases are still unclear, and high-throughput technologies typically report hundred or thousands of genes associated to a disease of interest. In this context is where gene-disease prioritization methods are of use. These computational methods make use of available data to obtain prioritized lists of genes (diseases) associated to a query set of diseases (genes). Prioritization is based on "guilt-by-association" which states that biological entities that are associated or interacting are more likely to share function. This allows to infer new relationships from already known interactions.

Many network-based prioritization methods have been proposed in the literature, performing well across different validation tests (Wang et al., 2011; Barabasi et al., 2011; Navlakha et al., 2010). We focus our study on two recent methods: rcNet (Hwang et al., 2011) and domainRBF (Zhang et al., 2011) since they outperform previous methods. Despite their good performance, these methods have clear limitations. First, they are strongly tailored to a specific domain of interest (gene-disease prioritization for rcNet and protein domain-disease prioritization for domainRBF, respectively). Hence, they cannot be applied to the prioritization of other biological entities of interest. Second, they do not allow to consider more than two types of networks for performing the prioritization (gene and disease networks in rcNet and domain and disease networks in domainRBF). However, we hypothesise that simultaneously integrating data from more than two complementary sources may improve the obtained results. For example, a gene-disease prioritization may benefit from known relationships between genes and diseases, but also from known interactions between drugs targeting certain genes to prevent or treat a specific disease.

We present ProphNet, a generic method of prioritization that achieves a better performance by integrating and propagating information in an arbitrary number of heterogeneous data networks. Our method is generic since it allows to prioritize any biological entity of any kind with respect to some biological entities of another kind. Therefore, the user can customize the goal of the prioritization task (disease-gene, domain-disease, etc.) and the networks that are being taking into account for prophNet to achieve this goal. ProphNet is available as a web application at http://genome2.ugr.es/prophnet/. MATLAB source code, datasets and detailed experiments can also be downloaded at http://genome2.ugr.es/prophnet/prophnet.zip. In this talk we present prophNet and compare its results to those obtained by rcNet and domainRBF in two cases of study associated to gene-disease and domain-disease prioritization, respectively.

Methods

To perform the prioritization task, our method measures the influence of a query set of biological entities of a certain type (e.g. genes or diseases) in a target set of entities of another type (e.g. diseases or genes, respectively). To this end, the algorithm uses a graph representation of data sources where each node corresponds to a biological entity of a type of interest (gene/protein, disease, protein domain, etc.), and the arcs between two nodes are labeled with a weight (from 0 to 1) representing the strength of the relationship between the connected entities. These weights are derived from different biological sources and their interpretation varies depending on the type of the connected entities and the final goal of the study. The nodes of the graph may also be labeled with a value, representing the degree of association of each entity to the query or target set.

Our method integrates a set of networks, each one connecting entities of one type, into a global network in which entities of different types are interconnected. In this work we focus on the prioritization of entities of different types, i.e. the query and target sets belong to two different networks. A simplified version of the method described below can be used to prioritize biological entities with respect to other entities of the same type (basically limiting the propagation to other networks).

To measure the degree of relationship between the query set and the target set, we first assign an initial value to the nodes of these sets. This initial value is set to 1/|X| by default, where |X| is the cardinal of the query set and target set, respectively. However, the initial values assigned to the entities of the query set may also be different in case we want to assign a different relative importance to the elements in this set, always satisfying that the sum of the assigned values equals one both in the target and query sets. Nodes not in the query or target sets are initially set to 0.

After the initial values have been set, these node values are propagated within each network and between networks. The propagation within a network is performed using the Flow Propagation algorithm (Vanunu et al., 2008) that iteratively propagates node values until convergence, taking into account the weight of the arc connecting two entities to perform the propagation of node values between these entities. The propagation of values from one network A to a neighbor network B is performed by assigning each node in B directly connected to nodes in A the average of the values of the neighbour nodes in A. Neighbour nodes which are connected with an arc labeled with a weight below a defined threshold are not considered in order to reduce the propagation noise.

This propagation within and between networks is performed through all the networks in the path connecting the query network to the target network. This process causes the nodes of the networks adjacent to the target network to take a value based on their degree of relationship to the query set.

Finally, to calculate the degree of relationship between the query set and the target set, we compute the correlation between the values of the target nodes and the values of the nodes from adjacent networks directly connected to the target nodes. To obtain a prioritized list of genes (target) associated to a particular disease (query), this prioritization algorithm is applied iteratively using each node (gene) from the target network as the target set and computing the degree of relationship with the query disease. The prioritized list is obtained by ordering the resultant correlation values in decreasing order. Since our method requires to iteratively compute correlation values for each query node and each target node, ProphNet is computationally expensive. However, it can be highly optimized by pre-calculating propagation scores in target networks and using parallelization techniques for fast response times. This way, a typical run for a gene-disease prioritization task takes a few seconds in our servers.

Results and Discussion

To compare the results obtained by prophNet with those obtained by state-of-the-art methods such as rcNet and domainRBF, we applied prophNet to the prioritization of genes-diseases and domain-diseases, respectively.

To perform a fair comparison of the results, we used the same data sources and methodology applied by rcNet and domainRBF to build the global network. The phenotype network was extracted from OMIM using text-mining techniques (van Driel et al., 2006) yielding a phenotype network with 5080 diseases. The phenotype-gene connections were extracted from OMIM using BioMart. The gene network was obtained from the Human Protein Reference Database (HPRD) and the protein domain network was derived from DOMINE and InterDom, with the domain-gene and domain-phenotype relationships extracted from Pfam.

We ran rcNet, domainRBF and ProphNet and tested their performance on different leave-one-out (LOO) cross-validation experiments. These LOO experiments were created by iteratively removing one gene-disease or one domain-disease relation from the available global network and using the corresponding gene or domain as query to check whether the prioritization method was able to predict the removed relationship. The accuracy of the result was measured as the rank assigned to the disease associated to the removed relation.

Apart from the LOO cross-validation experiments, we also performed experiments to test whether the different methods were able to predict new associations recently added to OMIM.

To measure the performance of the different prioritization methods, we computed Receiver Operating Characteristic (ROC) curves (data not shown due to format restriction) by plotting the

Table1: comparison of results for cross-validation experiments and new predictions for gene-disease and domain-disease prioritization tasks. Our method clearly outperforms rcNet and domainRBF in terms of AUC and mean ranking values. All ranking values are computed for a list of 5080 diseases. ROC curves associated to the AUC values were not included due to format restrictions.

Test	**Method**	**AUC**	**Mean rank (Std. Dev)**	**Norm. Mean rank**
LOO gene-disease	ProphNet	0,94	309.28 (811.51)	0.0609 (0.1597)
prioritization	rcNet	0,81	987.77 (1243.59)	0.1944 (0.2448)
New gene	ProphNet	0,81	980.37 (1329.85)	0.1930 (0.2618)
prioritization	rcNet	0,72	1441.59 (1476.84)	0.2835 (0.2907)
LOO domain-disease	ProphNet	0,93	346.87 (779.09)	0.0683 (0.1537)
prioritization	domain-RBF	0,87	671.58 (1199.2)	0.1322 (0.2361)

fraction of true positives out of the positives vs. the fraction of false positives out of the negatives, at various threshold settings. Areas under the ROC curves (AUC) and the average position in which the correct entity was ranked (see Table 1) are also computed.

The obtained results show that prophNet outperforms the other methods in all the proposed experiments, achieving the best avg. ranking position with the lowest standard deviation. A t-test has been performed on the mean ranking values obtained by the two algorithms compared in each expermient. The obtained p-values are less than 0.0001 for all the tests.

Although ProphNet results are significantly better than those obtained by other methods, a high mean ranking value was obtained due to the high variability of the results. Detailed results (not shown) reveal that although in most LOO runs the correct disease is prioritized at the top positions, for a small fraction of cases the disease is much worse ranked, increasing the mean ranking value. Further studies are needed to analyze these cases to improve the results of the algorithm.

ProphNet was also applied to obtain prioritized lists of genes associated to Alzheimer, Diabetes Mellitus Type II and Breast Cancer (results not shown). The resultant top-ranked genes were related to these diseases according to recent publications in the literature.

Acknowledgements

This project was supported by grants from the Plan Propio de Investigación of the University of Granada, Spain, and it is part of projects P08-TIC-4299 of J. A., Sevilla and TIN2009-13489 of DGICT, Madrid.

References

1. Wang X, Gulbahce N and Yu H (2011) Network-based methods for human disease gene prediction. Briefings in Functional Genomics 10(5):280-293. doi: 10.1093/bfgp/elr024
2. Barabasi A, Gulbahce N and Loscalzo J (2011) Network medicine: a network-based approach to human disease. Nature Reviews Genetics 12:56-68. doi:10.1038/nrg2918
3. Navlakha S and Kingsford C (2010) The power of protein interaction networks for associating genes with diseases. Bioinformatics 26(8):1057-1063. doi:10.1093/bioinformatics/btq076
4. Hwang T, Zhang W, et al. (2011) Inferring disease and gene set associations with rank coherence in networks. Bioinformatics 27(19): 2692-2699. doi:10.1093/bioinformatics/btr463
5. Zhang W, et al. (2011) DomainRBF: a Bayesian regression approach to the prioritization of candidate domains for complex diseases. BMC Systems Biology 5:55. doi:10.1186/1752-0509-5-55
6. Vanunu O and Sharan R (2008) A propagation based algorithm for inferring gene-disease associations. Proceedings of the German Conference on Bioinformatics. German Conference on Bioinformatics: 54-62.
7. van Driel MA, et al. (2006) A text-mining analysis of the human phenome. European Journal of Human Genetics 14: 535-542. doi:10.1038/sj.ejhg.5201585