Protein-disease association significance from candidate ranking lists
Berenstein Ariel, Ibañez Irene, Chernomoretz Ariel
Integrative Systems Biology Group, Fundación Instituto Leloir, & CONICET, Buenos Aires, Argentina

Introduction
One of the most challenging problems in biomedical research is to understand the mechanisms underlying human diseases. Great effort has been spent on determining genes associated to diseases, from early genetic mapping and molecular biology studies to recent efforts making use of knowledge from human genome sequence projects. All these studies have provided with large data sets on the genetic basis of human diseases.

In this context the network-metaphor has appeared as an appealing framework not only for the sake of data organization, but also to unveil patterns of biological relevance. In recent years there has been a lot of interest in the application of complex network theory in human health related research, in order to analyze molecular basis of human diseases, identify comorbidity patterns, perform drug prioritization tasks, and to predict new disease-gene product associations. A large number of this last type of research programs assume that protein involved in the same disease have an increased tendency to interact with each other. They involve the use of already known gene-disease associations, a complex network of interacting proteins that encode physical or functional relationships between them, and a kind of information propagation technique in order to rank candidate proteins in terms of their degree of association with the disease-related seeds.

In this communication we show that predictions based on ranking candidate lists produced by this type of algorithms can be highly biased by network topology properties and produce inaccurate estimation of protein-disease association significances, and we propose a boot-strapping technique in order to alleviate this problem.

Methods
In order to perform our analysis we considered genes and protein associated to the Alzheimer disease as reported by the DisGenet database [Bauer-Mehren 2011]. Protein-protein interactions were inferred from the Human Interaction Network (HIN), a literature-based manually curated network that includes 9700 proteins and 71538 interactions [Ceriami 2010]. We considered the FuncionalFlow (FF) algorithm, that takes advantage of both, network topology and some measure of locality [Navieba2005] as a prototypical technique to generate a ranking of novel putative Alzheimer-related proteins. We used the implementation provided by the GUILD framework [Guney2011] with default values.

Results and discussion
DisGenet reported 35 genes associated to the Alzheimer disease, from which 25 were also included in the HIN network. Using this 25 proteins as seeds for the FunctionalFlow algorithm we obtained a ranking of the 9700 HIN proteins. Usually researchers are tempted to consider the top ranked proteins as new candidates. However this procedure does not take into account the statistical significance of the gene-disease association.

In order to address this issue we produced 10000 ranking lists using randomly selected 25 proteins as seeds, and assigned a p-value to each protein looking at how many randomly generated scores resulted in larger values than the observed ones. After correcting for multiple hypothesis testing using a FDR prescription, no protein remains significant at a level lower than 0.3. This means that the observed scores generated by the original seed-set show no statistical significance at all support the protein-
disease association.
It is well established that biological networks are far from random, and present a rich topological structure that includes nodes with different connectivity patterns. With this image in mind, we reasoned that a local null hypothesis will be more appropriate to tests the gene/disease association.
In order to take into account the local topological context of each node in the network we assigned statistical significance to the observed protein’s scores, considering different null hypothesis (one per node) built upon FF scores assigned to each single node, when random seeds were considered. Using this technique (local statistical score, LSS) we found 100 proteins with corrected q-values lower than a 0.01 confidence level. Importantly, only 68 of them are included in the top-100 ranking list.

In Fig 1 we show the location of all HIN proteins in a degree-betweenness plane. Blue and green points correspond to the top-100 ranking and LSS candidate proteins, whereas red points correspond to proteins detected by both techniques.

It can be show that top-ranking proteins tend to be hub nodes with high levels of betweenness. This explain why they present no statistical significance, as random seeds aslo typically produce high scores for this proteins, reflecting topological network properties, and no significant association with the tested disease. The new presented method (green points), on the other hand, preserves some of this predictions, but allows one to also pinpoint significant associations involving more poorly connected nodes.

References


