Disease gene prioritization by combining network information and functional knowledge

Tim Kacprowski Nadezhda T. Doncheva David Buezas
Mario Albrecht

Max Planck Institute for Informatics, Campus E1.4, 66123 Saarbrücken, Germany

Introduction

The identification and functional characterization of disease genes are important problems in medical bioinformatics. The results of experimental high-throughput work, such as genome-wide association studies and RNA interference screens, yield long lists of candidate disease genes. However, the analysis and validation of these genes are time-consuming and expensive tasks. Therefore, computational prioritization methods are required to narrow down the set of potential disease genes.

To this end, a number of prioritization approaches has been developed recently [1, 2]. Some are based on a single data source, but many take multiple data sources into account. In particular, protein interactions and functional annotations are two major data sources [1, 2]. Recently, we introduced the MedSim method, which utilizes the functional similarity of Gene Ontology (GO) terms [3, 4]. This showed that excellent performance can already be achieved using functional annotations alone. However, if sufficient annotations of good quality are not available for specific genes, additional data sources have to be included into the prioritization procedure. Therefore, we present a novel combination approach that addresses this issue.

Materials and Methods

In contrast to other prioritization methods that consider only identical annotations or are based on GO enrichment computations, MedSim quantitatively assesses the functional similarity of GO terms using sophisticated similarity measures [5]. MedSim automatically derives a functional profile for a certain disease phenotype from the GO term annotations of the associated disease genes. The functional similarity between the disease profile and the corresponding profiles of candidate genes is then used to create the final rank list of candidate disease genes.

Based on MedSim and additional network information, our new approach prioritizes candidate genes in three main steps: (1) construction of molecular networks for candidate genes, (2) quantification of candidate gene relevance using network measures, and (3) final aggregation of multiple rank lists of potential disease genes.

To construct a functional similarity network of genes, one form of molecular network, we retrieved the functional gene similarities from the online database FunSimMat, a comprehensive web resource of functional similarity values [3]. These were transformed into weighted gene-gene relationships. Each network comprises known disease-associated genes in addition to the candidate genes. Alternatively, it is possible to use other forms of molecular networks such as protein interaction networks.

Afterwards, we quantify the disease relevance of each candidate gene to the known disease genes in a molecular network. To this end, we compute several network centrality measures based on shortest paths and random walks. These measures can be applied on both weighted and unweighted networks and indicate the closeness and betweenness of the candidate genes to the disease genes [6].

Since each computed centrality measure implicates a ranking of the candidate genes, we obtain multiple rank lists. Each list ranks the candidate genes w.r.t. their relevance to the known

disease genes. These rank lists can then be combined using rank aggregation algorithms. Additionally, different confidence values can be assigned to the individual rank lists. Finally, our prioritization approach results in a single overall rank list.

Results

We applied leave-one-out cross-validation to benchmark our combination approach on a dataset of 99 disease phenotypes and their associated genes/proteins from OMIM and UniProtKB [4]. Putative candidate genes were derived from artificial quantitative trait loci of size 10 Mbp around every disease gene. For each disease, we constructed a functional similarity network that contains all genes from the disease loci and computed the centrality measures to rank the candidate genes. We then evaluated the top-ranked genes in the finally aggregated list.

Furthermore, we compared MedSim and our new combination approach in a comprehensive case study for Crohn's disease on recently published data of 71 susceptibility loci [7]. This resulted in a number of potential disease genes in the loci for follow-up experiments. We also developed a new user interface to MedSim based on the web service FunSimMat [3]. This allows the upload of own candidate lists and the interactive usage of the prioritization procedure for many diseases. Moreover, the combination approach is implemented as Cytoscape plugin.

References

[1] Tranchevent, L.C., Capdevila, F.B., Nitsch, D., De Moor, B., et al.: A guide to web tools to prioritize candidate genes. Briefings in Bioinformatics **12** (2011) 22–32

[2] Kann, M.G.: Advances in translational bioinformatics: computational approaches for the hunting of disease genes. Briefings in Bioinformatics **11** (2010) 96–110

[3] Schlicker, A., Albrecht, M.: FunSimMat update: new features for exploring functional similarity. Nucleic Acids Research **38** (2010) D244–D248

[4] Schlicker, A., Lengauer, T., Albrecht, M.: Improving disease gene prioritization using the semantic similarity of Gene Ontology terms. Bioinformatics **26** (2010) i561–i567

[5] Schlicker, A., Domingues, F.S., Rahnenführer, J., Lengauer, T.: A new measure for functional similarity of gene products based on Gene Ontology. BMC Bioinformatics **7** (2006) 302

[6] Opsahl, T., Agneessens, F., Skvoretz, J.: Node centrality in weighted networks: Generalizing degree and shortest paths. Social Networks **32** (2010) 245–251

[7] Franke, A., McGovern, D.P.B., Barrett, J.C., Wang, K., et al.: Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nature Genetics **42** (2010) 1118–1125