DiGSNP: a web tool for Disease-Gene-SNP hierarchical prioritization

Carmen Navarro[1]✉, Carlos Cano[1], Armando Blanco[1], Fernando García-Alcalde[2]

[1]Department of Computer Science and Artificial Intelligence, University of Granada, Granada, Spain
[2]Max Planck Institute for Infection Biology, Berlin, Germany

Motivation and Objectives

Understanding the genetic causes of human diseases is the major goal towards an effective personalized medicine. High-throughput technologies such as linkage analysis, association studies and array experiments allow to obtain experimental evidence of chromosomal regions associated with phenotypes. However, these technologies typically report a large number of results (i.e. genes, variants, etc.) associated with the conditions under study. In this context, tools supporting researchers in the process of prioritizing diseases, genes, and variations are highly desired to assist the scientific research and provide guidance on the most promising hypotheses. To this end, many gene-disease prioritization methods have been proposed in the literature (Moreau and Tranchevent, 2012). These methods describe computational approaches that use information retrieved from diverse sources in order to obtain prioritized lists of candidate genes to be related with a certain target disease. However, most of these tools do not consider gene variations, although they are known to be the main cause for many diseases. Proposals like AnnTools (Makarov et al., 2012) or SNPRank (Jadamba et al., 2012), based in genome-wide association studies (GWAS), relate variations directly to diseases, but leave gene-disease information out or implicit. Moreover, most currently available tools for associating genome variations to diseases focus on coding regions, disregarding relevant information present in the promoter regions of genes, such as variations that alter the binding affinity of transcription factor binding sites (TFBS), which have been shown to play an important role in the regulatory machinery of the cell.

In this work we present DiGSNP (Disease-Gene-SNP Prioritizer), a tool which allows to relate diseases, genes and variations in regulatory regions of the genome (particularly, those affecting TFBSs), simultaneously, helping researchers to understand how these relations work and focus on the most relevant regulatory elements in the early research stage of any disease.

Methods

DiGSNP prioritizer searches for relations between diseases, genes and variations in a two-level hierarchy. The first level builds an ordered list of genes related to a query disease. The second level adds a list of single-nucleotide polymorphisms (SNP) present in TFBSs in the regulatory regions of each gene (i.e. building a Disease-gene-SNP hierarchy), prioritized by the expected level of influence in the binding affinity of the TFBS.

Disease-gene prioritization method is based on ProphNet (Martinez et al., 2012, http://genome2.ugr.es/prophnet/). This method allows to prioritize biological entities from different domains (e.g. genes, diseases, protein domains) by integrating an arbitrary amount of heterogeneous sources of data represented as networks. The resultant super-graph is then mined using a Random Walk with Restart (RWR) algorithm for obtaining prioritized lists of elements associated to the user query. We have applied the ProphNet algorithm to obtain prioritized lists of genes for a query disease. The algorithm was applied on a network composed of three different types of nodes: genes/proteins, phenotypes and protein domains. The phenotype network and the phenotype-gene connections were extracted from OMIM using text-mining techniques; the gene network was obtained from the Human Protein Reference Database (HPRD, http://www.hprd.org/) and the protein domain network was derived from DOMINE and InterDom (http://interdom.i2r.a-star.edu.sg/), with the domain-gene and domain-phenotype relationships extracted from Pfam (http://pfam.sanger.ac.uk/).

After obtaining a prioritized list of genes related to the query disease, each gene is associated to a list of SNPs present in its promoter regions. SNPs are obtained from dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/). Gene-SNP prioritization is applied to each SNP list based on two criteria. First, SNPs located in a TFBS are candidate regulatory SNPs. Second, a SNP that causes drastic changes in the binding affinity of a TFBS has a higher probability to affect the gene regulation

and therefore to be related to the query disorder. In order to assess whether a SNP is located in a known TFBS, SC_{intuit} (García-Alcalde et al., 2010), a sequence-motif similarity measure, is used. SC_{intuit} applies intuitionistic theory, an extension of fuzzy theory, to generate a similarity score. Motif information is retrieved from Jaspar (http://jaspar.cgb.ki.se/) and TRANSFAC. According to the mentioned two criteria, DiGSNP reduces the set of SNPs in regulatory regions of the gene to a set of SNPs located in a site matching a TFBS, i.e. sites showing a SC_{intuit} similarity score to a known TFBS above a provided threshold. Selected SNPs are ordered depending on the difference of similarity between the mutated and wild-type alleles. Therefore, SNPs that dramatically alter the binding affinity of a TFBS are ranked at the top by DiGSNP.

Table 1: Fragment of the information obtained with DiGSNP for Alzheimer disesase. A top rank of 5 genes is shown, and for each of them the top ranked 3 SNPs. Each SNPs is also associated with the region of the gene where it was found and the motif that generated its score.

Disease	Gene	SNP	SNP score	Gene region	Motif ID
Alzheimer	APP	rs199610454	0,32	5' UTR	Kid3
		rs201528959	0,27	5' UTR	Churchill
		rs200990709	0,25	5' UTR	HNF4
	PSEN2	rs200123803	0,34	5' near gene	ZNF354C
		rs200034334	0,32	5' UTR	Kid3
		rs150618255	0,29	5' near gene	Kid3
	PSEN1	rs202004275	0,32	5' UTR	Kid3
		rs201506908	0,29	5' UTR	Kid3
		rs200531676	0,29	5' UTR	MAFB
	TREM2	rs113167129	0,34	5' near gene	ZNF333
		rs187797067	0,34	5' near gene	C-MAF
		rs138222305	0,32	5' near gene	Kid3
	HD	rs192838728	0,32	5' near gene	Kid3
		rs28616835	0,32	5' near gene	Kid3
		rs398691	0,32	5' near gene	Kid3

Integrating these two prioritization methods, we offer an approach to disease-gene-SNP prioritization. In table 1, a summary result relating Alzheimer disease to a list of prioritized genes and each gene relating to a list of prioritized SNPs is shown. This way of structuring the information allows the user to visually infer possible relations among genes, diseases, SNPs and TFBSs. For instance, the frequent appearance of motif Kid3 as best result in many SNPs reveals a direct relation of Kid3 and Alzheimer (Acquaah-Mensah et al., 2012), which would have been otherwise difficult to discover.

Results and Discussion

The proposed methodology can be applied to any prioritization method that can score genes relating to diseases and SNPs related to genes. Due to the lack of information sources relating regulatory variations and diseases, validation becomes a difficult process, along with determining the biological impact of the results obtained. Relations between genes in Table 1 such as APP, TREM1 and TREM2 and Alzheimer disease can be found in the literature (Cruchaga et al. 2012). However, finding information in the literature about the SNPs that appear in table 1 was not possible. The main reason is probably that the focus of research relating to SNPs has relied on coding regions, searching for missense variations in exomic areas of the genome. Our main focus is transcriptional regulation, area from which the amount of available information is much less significant. Moreover, other tools, like SNPRank (Jadamba et al., 2012), focus on coding areas of the genome. This makes it unfeasible to compare the results, since the sets of SNPs obtained should always be different. Other approaches like regSNPs (Teng et al. 2012) focus on regulatory elements, but require experimental evidence from a GWAS and make an inverse process, starting from variations proven to be related to a disease by GWAS results. Furthermore, these tools relate directly diseases and variations, leaving gene information out or implicit.

We believe that DiGSNP can be helpful to researchers, who can see in a glance the relationship between a certain disease and a set of SNPs related to genes, probably involved in the regulation processes that affect the target disease. In addition, the second step of DiGSNP focuses on genomic information and our knowledge about TFBSs and their binding affinity, making it possible for researchers to obtain a set of probable candidates for any disease. Evidence of variation-disease association in the literature is not needed for placing a query in DiGSNP. This feature makes DiGSNP a helpful tool when trying to discover a

set of highly related SNPs and genes to a new query disease to boost further research.

Acknowledgements

This work has been carried out as part of projects P08-TIC-4299 of J. A., Sevilla and TIN2009-13489 of DGICT, Madrid

References

1. Acquaah-Mensah GK, Taylor RC, et al. (2012) PACAP interactions in the mouse brain: implications for behavioral and other disorders. Gene, 491(2):224-231. doi:10.1016/j.gene.2011.09.017.
2. Cruchaga C, Chakraverty S, et al. (2012). Rare Variants in APP, PSEN1 and PSEN2 Increase Risk for AD in Late-Onset Alzheimer's Disease Families. PLoS One 7(2). doi:10.1371/journal.pone.0031039
3. García-Alcalde F, Blanco A, et al. (2010). An intuitionistic approach to scoring DNA sequences against transcription factor binding site motifs. BMC bioinformatics, 11(1): 551. doi:10.1186/1471-2105-11-551.
4. Jadamba E, Shin M. (2012). A SNP Prioritization Method Using Linkage Disequilibrium Network for Disease Association Study. INTELLI 2012 (c), 86-88.
5. Makarov V, O'Grady T, et al. (2012). AnnTools: a comprehensive and versatile annotation toolkit for genomic variants. Bioinformatics, 28(5):724-5. doi:10.1093/bioinformatics/bts032
6. Martinez V, Cano C, et al. (2012). Network-based gene-disease prioritization using ProphNet. EMBnet.journal S18.B (in press)
7. Moreau Y and Tranchevent LC (2012). Computational tools for prioritizing candidate genes: boosting disease gene discovery. Nature Reviews Genetics 13:523-536. doi:10.1038/nrg3253