Taking a Walk on the Loci: Simultaneous Gene Prioritization in Multiple Loci using Networks

Sonia Leach[1], Yves Moreau[2]

1 Introduction

Genetic disorders have increasingly been investigated using array-based comparative genomic hybridization (aCGH) techniques for detecting segmental genomic alterations. Often the list of candidate causal genes can be focused to those lying in aberrant genomic regions common among multiple patients. However, in complex diseases, each patient in a cohort may have a distinct set of aberrant loci. The challenge then becomes to prioritize candidate genes within a locus simultaneously with respect to other loci similarly identified as important in other patients.

2 Method

Franke et al., 2006 [3] presented a method for prioritizing genes among multiple loci using a graph distance-based metric on gene networks. Intuitively, since a disorder can be caused by disruption of any member along a given biological pathway, the goal was to identify a subset of genes distinctly contributed by multiple loci that were strongly functionally related. Given a network where edge weights reflect confidence of a functional relationship between the respective gene pair, effect scores were computed for each candidate gene in a given locus by their method as a function of the shortest-path distance to genes on each of the other loci on the gene network (Figure 1). To overcome the bias of hubs evident in shortest-path approaches, they ranked candidate genes using an empirically derived p-value computed from randomizations.

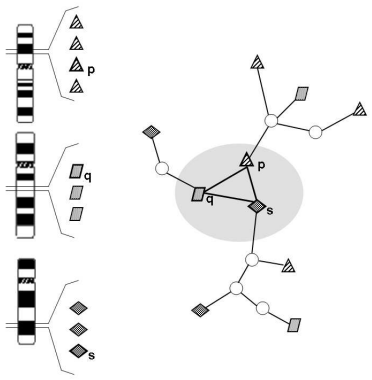

Figure 1: Network-based Multiple Loci Prioritzation (adapted from [3]). The genes p, q and s are the strongest candidates in their respective loci since they are found near each other in the gene network.

Despite the hub correction in Franke et al., 2006 [3], using a shortest-path distance measure makes it difficult to extract a biologically relevant subnetwork among the top candidates, thereby limiting the interpretability and usefulness for biologists. In this work, we investigate two network-based distance measures as alternatives to shortest-path, namely 1) a Laplacian diffusion kernel method [2] where two nodes are considered closer in the network if they are connected by multiple paths and 2) a random walk method [1] which exploits the input set of nodes of interest to give importance only to the subgraph connecting the input set. Each of these methods focuses on highlighting a strongly connected subnetwork among candidates, rather than isolated shortest paths between gene pairs.

[1]Dept. of Electrical Engineering (ESAT/SCD/BioI), Katholieke Universiteit Leuven, Leuven, Belgium,
E-mail: sonia.leach@esat.kuleuven.be
[2]Dept. of Electrical Engineering (ESAT/SCD/BioI), Katholieke Universiteit Leuven, Leuven, Belgium,
E-mail: yves.moreau@esat.kuleuven.be

3 Results

We compare the three distance alternatives (shortest-path, diffusion kernel, and random walks), using the human network from the STRING database [4] and the set of 96 diseases considered in Franke et al., 2006 [3] which had at least three associated disease genes in OMIM [5]. Generating a pseudo-locus of 100 genes surrounding each disease gene and recording the fraction of the true disease genes ranked in the top 10 for each disease, the three measures show the same performance on many diseases since the disease genes were immediate neighbors in the network, thereby making the methods equivalent. For the remainder, the diffusion kernel and random walk alternatives identified more of the known disease genes in the top 10. More importantly, we demonstrate new methods for extracting subnetworks among the strong candidate genes from the distance metrics and show that using the two new measures create networks more interpretable and biologically relevant than those derived from shortest-path distance.

References

[1] Dupont, P., Callut, J., Dooms, G., Monette, J-N. and Deville, Y. 2006. Relevant subgraph extraction from random walks in a graph. *Universite catholique de Louvain, UCL/INGI, Number RR 2006-07.*

[2] Fouss, F., Luh, Y., Pirotte, A. and Saerens, M. 2006. An Experimental Investigation of Graph Kernels on a Collaborative Recommendation Task. *Proceedings of the 2006 IEEE International Conference on Data Mining (ICDM 2006).* pp. 863–868.

[3] Franke, L., van Bakel, H., Fokkens, L., de Jong, E.D., Egmont-Petersen, M. and Wijmenga, C. 2006. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am J Hum Genet.* Jun;78(6):1011–25.

[4] Jensen, L.J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M., Bork, P. and von Mering, C. 2009. STRING 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.* Jan;37(Database issue):D412-6.

[5] Online Mendelian Inheritance in Man, OMIM (TM). McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD), World Wide Web URL: http://www.ncbi.nlm.nih.gov/omim/.