EPSILON: localized networks for eQTL prioritization

Lieven P.C. Verbeke[1], Piet Demeester[1], Jan Fostier[1], and Kathleen Marchal[2,3]

[1] *IBCN - iMinds, Ghent University, Belgium, lieven.verbeke@intec.ugent.be*
[2] *Department of Microbial and Molecular Systems, KU Leuven, Belgium*
[3] *Department of Plant Biotechnology and Bioinformatics, Ghent University, Belgium*

Abstract

When genomic data is associated with gene expression data, the resulting expression quantitative trait loci (eQTL) will very likely span multiple genes. eQTL prioritization techniques can be used to select the most likely causal gene affecting the expression of a target gene from a list of candidates. As an input, these techniques use physical interaction networks that often contain highly connected genes and unreliable or irrelevant interactions that can interfere with the prioritization process. We present EPSILON, a framework for eQTL prioritization that mitigates the effect of highly connected genes and unreliable interactions. We tested the new method on eQTL data sets derived from yeast data. A physical interaction network was constructed and each eQTL in each data set was prioritized using the EPSILON approach: first a local network was constructed using a k-trials shortest path algorithm, followed by the calculation of a network-based similarity measure. We found that using a local network significantly increased prioritization performance in terms of predicted knockout pairs when compared to using exactly the same network similarity measures on the global network. EPSILON performed on par or better than two alternative eQTL prioritization approaches, ITM-Probe and eQED.

1 Introduction

Due to linkage disequilibrium and the spacing of the genetic markers on the genome, genetic markers represent a region on a chromosome that covers multiple genes rather than a single gene. The variability in expression of the genes found to be associated with an eQTL (here referred to as *target* genes) is most likely caused by a mutation in a single gene located on the eQTL (the *causal* gene). Gene prioritization or refinement methods are needed to distinguish the causal gene from a list of candidate causal genes.

A relatively small number of techniques were developed to tackle the rather specific eQTL prioritization task. All eQTL prioritization methods have in common that they use a physical interaction network to define a similarity measure between a target gene and a set of candidate causal genes. Tu *et al.* (2006) developed a method based on random walks in a physical interaction network, an approach later refined by Suthram *et al.* (2008), who extended the random walk idea with an electric circuit

analogy. Voevodski *et al.* (2009) applied the PageRank algorithm to develop a gene affinity measure, and Stojmirović and Yu (2012) used the mathematical modeling of information flow in a network to rank candidate genes.

Stojmirović and Yu (2012) suggest localizing the network, i.e. excluding distant genes from the network that connects an origin (the target gene) with a set of destinations (the candidate causal genes), prior to analysis in order to better reflect the biological context. Otherwise, results of e.g. gene prioritization will be highly dependent on the node degree of the genes in the network. Simply removing genes from the network with a node degree exceeding an arbitrary threshold, or heuristically downweighting the importance of relations based on the number of connections, risks removing useful genes or important relations (Zotenko *et al.*, 2008). To handle both localization and prioritization simultaneously, we present EPSILON.

2 EPSILON framework

The EPSILON method contains two steps, which are applied to each association found: (1) construct, from an existing global interaction network, a local sub-network that connects the candidate causal genes covered by an eQTL with the target gene and (2) calculate a similarity measure that expresses the functional similarity between the target gene and a candidate causal gene. As input, the results of an eQTL association analysis are used.

To restrict the network around a set of candidate causal genes and a target gene, a shortest/cheapest path approach is applied. All interactions are assigned a cost, and an optimal path from each candidate to the target was found using the Dijkstra algorithm. All genes and interactions that were found on such a shortest path were included in the sub-network. Furthermore, it was investigated if enlarging this neighborhood could improve the prioritization results. This was achieved by k times considering if an alternative shortest path exists, that is different from any previously found path.

Once the local network connecting all candidate causal genes with the target gene is constructed, the EPSILON framework requires the calculation of a network similarity measure between the target gene and all candidates to assess their functional relatedness. In principle, any network-based similarity measure could be integrated. Several authors (e.g. Tu *et al.* (2006), Suthram *et al.* (2008), Shih and Parthasarathy (2012)) propose a random walk (RW) approach, in which a random walk is initiated a very high number of times from a candidate causal gene, and it is measured how many times a random walker is found in the target gene.

Next to integrating random walks in EPSILON, we investigated kernels calculated on graph nodes as an alternative similarity measure. These kernels are an attractive tool for uncovering relations in large networks (Fouss *et al.*, 2006). In this study, we evaluated two well-known kernels, the Regularized Commute-Time (RCT) kernel and the Laplacian Exponential Diffusion (LED) kernel.

3 Results and Discussion

We evaluated EPSILON, a k-trials shortest path network construction method combined with random walk and kernel-based similarity measures, using a gold standard data set derived from a yeast knockout compendium. We applied three commonly used association techniques to the SNP and expression data (Saccharomyces cerevisiae) of Brem and Kruglyak (2005): non-parametric regression, mixed models and elastic net regression. An interaction network was constructed using public databases, containing protein-protein interactions, transcription factors with targets and phosphorylation interactions.

We were able to show that our approach, outperformed random assignment and a shortest path reference method. More interestingly, the global network analogues of the network similarity measures too were outperformed significantly ($p < 10^{-5}$), clearly showing the added value of using local over global networks. We assume that constraining the global network to a local neighborhood around the target gene and all candidate causal genes is effectively reducing the disturbing impact of hubs and promiscuous genes. EPSILON was compared to two other methods, ITM Probe and eQED. We found that EPSILON performed as well or better than ITM Probe. EPSILON clearly outperformed eQED, be it using a reduced network because eQED could not deal with the phosphorylation interactions present in the global network.

4 Acknowledgments

This work is supported by: (1) Ghent University Multidisciplinary Research Partnership 'Bioinformatics: from nucleotides to networks', (2) Fonds Wetenschappelijk Onderzoek-Vlaanderen (FWO) G.0428.13N and (3) Katholieke Universiteit Leuven funding: PF/10/010 (NATAR).

References

Brem, R. B. and Kruglyak, L. (2005). The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(5), 1572–7.

Fouss, F., Yen, L., Pirotte, A., and Saerens, M. (2006). An Experimental Investigation of Graph Kernels on a Collaborative Recommendation Task. In *IEEE International Conference on Data Mining - ICDM*, pages 863–868. Citeseer.

Shih, Y.-K. and Parthasarathy, S. (2012). A single source k-shortest paths algorithm to infer regulatory pathways in a gene network. *Bioinformatics*, **28**(12), i49–i58.

Stojmirović, A. and Yu, Y.-K. (2012). Information flow in interaction networks II: channels, path lengths, and potentials. *Journal of computational biology : a journal of computational molecular cell biology*, **19**(4), 379–403.

Suthram, S., Beyer, A., Karp, R. M., Eldar, Y., and Ideker, T. (2008). eQED: an efficient method for interpreting eQTL associations using protein networks. *Molecular systems biology*, **4**(162), 162.

Tu, Z., Wang, L., Arbeitman, M. N., Chen, T., and Sun, F. (2006). An integrative approach for causal gene identification and gene regulatory pathway inference. *Bioinformatics*, **22**(14), e489–e496.

Voevodski, K., Teng, S.-H., and Xia, Y. (2009). Spectral affinity in protein networks. *BMC systems biology*, **3**(1), 112.

Zotenko, E., Mestre, J., O'Leary, D. P., and Przytycka, T. M. (2008). Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality. *PLoS computational biology*, **4**(8), e1000140.