BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 10 2013, pages 1308–1316
doi:10.1093/bioinformatics/btt142

Systems biology

Advance Access publication April 16, 2013

EPSILON: an eQTL prioritization framework using similarity measures derived from local networks

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

[1]Department of Information Technology, Ghent University - iMinds, 9050 Gent, Belgium, [2]Department of Microbial and Molecular Systems, KU Leuven, 3001 Leuven, Belgium and [3]Department of Plant Biotechnology and Bioinformatics, Ghent University

Associate Editor: Mario Albrecht

ABSTRACT

Motivation: When genomic data are associated with gene expression data, the resulting expression quantitative trait loci (eQTL) will 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, an extendable framework for eQTL prioritization, which mitigates the effect of highly connected genes and unreliable interactions by constructing a local network before a network-based similarity measure is applied to select the true causal gene.

Results: We tested the new method on three eQTL datasets derived from yeast data using three different association techniques. A physical interaction network was constructed, and each eQTL in each dataset 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. Three similarity measures were evaluated: random walks, the Laplacian Exponential Diffusion kernel and the Regularized Commute-Time kernel. The aim was to predict knockout interactions from a yeast knockout compendium. EPSILON outperformed two reference prioritization methods, random assignment and shortest path prioritization. Next, we found that using a local network significantly increased prioritization performance in terms of predicted knockout pairs when compared with using exactly the same network similarity measures on the global network, with an average increase in prioritization performance of 8 percentage points ($P < 10^{-5}$).

Availability: The physical interaction network and the source code (Matlab/C++) of our implementation can be downloaded from http://bioinformatics.intec.ugent.be/epsilon.

Contact: lieven.verbeke@intec.ugent.be, kamar@psb.ugent.be, jan.fostier@intec.ugent.be

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on November 21, 2012; revised on February 26, 2013; accepted on March 21, 2013

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the last two authors should be regarded as joint Last Authors.

1 INTRODUCTION

Expression Quantitative Trait Locus (eQTL) analysis is the process of associating genetic variation in a population with variable gene expression to identify polymorphic genetic regions affecting gene expression (Michaelson *et al.*, 2009). Frequently, genetic markers are used to sample the genetic diversity between individuals of an organism. Owing to linkage disequilibrium and the spacing of the genetic markers on the genome, these 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 the eQTL (here referred to as *target* genes) is most likely caused by a mutation in a single gene located on the eQTL (the true *causal* gene). Yet, eQTL analysis as such is not able to distinguish this causal gene from the remainder of the genes located on the eQTL (the *candidate causal* genes). Instead, gene prioritization or refinement methods are needed.

A number of prioritization methods that rank candidate genes according to some criterion have been developed in the past [see Tranchevent *et al.* (2011) for an extensive overview]. Typically, these methods target novel (human) disease-gene identification. Information about the disease under study, i.e. in the form of a list of existing disease genes is often needed. Other methods target specific experiments for which additional data, divided in disease/control groups, are required and cannot be readily applied to the results of an eQTL analysis where such groups are not necessarily present. Also, some of these techniques do not allow for the incorporation of a custom gene interaction network, as only predefined networks for a limited number of organisms are available. A smaller number of techniques were developed to tackle the more 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. The higher the network-based similarity between the target and the candidate causal gene, the more likely the candidate corresponds to the true causal gene. Tu *et al.* (2006) developed a method based on random walks (RWs) in a physical interaction network, an approach later refined by Suthram *et al.* (2008), who extended the RW 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.

According to Stojmirović and Yu (2012), it is most important to use directed relations (e.g. protein–DNA interactions) in the interaction network whenever possible. They also 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), before analysis to better reflect the biological context. Otherwise, results of, for example, gene prioritization will be highly dependent on the node degree of the genes in the network. A high node degree can point to functionally useless genes (so-called promiscuous genes). However, genes with a high node degree can also correspond to important hubs (Gillis and Pavlidis, 2011). As a result, 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 overcome this problem, we propose EPSILON, an eQTL prioritization framework using similarity measures derived from local networks. The main assumption underlying EPSILON is that the disturbing influence of highly connected genes and unreliable or irrelevant interactions can be controlled by reducing the global interaction network to a local neighborhood (i.e. a sub-network) connecting the target gene and all candidate causal genes, before calculating a network-based similarity measure. Once the local network is constructed, a similarity measure between the target gene and all candidate causal genes can be calculated, and the candidate with the highest similarity can be selected as the true causal gene. We demonstrate the added value of a local approach for eQTL prioritization by plugging in established prioritization methods such as path finding and RWs in the EPSILON framework. Alternatively, we also investigate a graph node kernel-based similarity measure. We evaluate the performance of all methods using a gold standard dataset derived from a knockout compendium.

2 EPSILON FRAMEWORK

An overview of the EPSILON refinement scheme is presented in Figure 1. As input, the results of an eQTL association analysis are used. Three different association techniques were tested. It is well-known that the results of eQTL analysis are highly dependent on the mapping method used, and we wanted to avoid a prioritization bias toward any method. We used Mixed Models (MM), Non-Parametric Regression (NPR) and Elastic Net (EN) regression for the association analysis, all of which are described in the Section 5. The EPSILON method contains two steps, which are applied to each association found: (i) 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 (ii) calculate a similarity measure that expresses the functional similarity between the target gene and a candidate causal gene. Both steps are elaborated later in the text.

2.1 Extraction of local networks

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 (see later in the text), and an optimal path from each candidate to the target was found using the Dijkstra algorithm (Dijkstra, 1959). All genes and interactions that were found on such a shortest path were included in the sub-network. Furthermore, it was investigated whether enlarging this neighborhood could improve the prioritization results. This was achieved by k times considering if an alternative shortest path exists, i.e. different from any previously found path

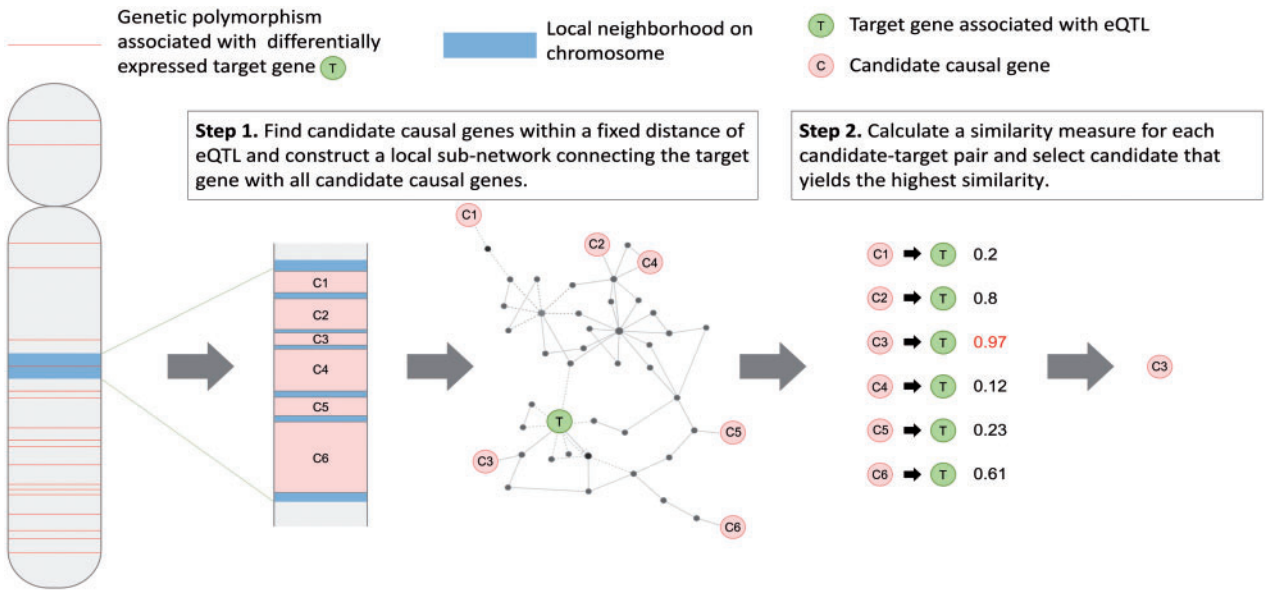

Fig. 1. The EPSILON prioritization flow applied to a fictive eQTL-target pair. First, candidate causal genes (red) overlapping with a local neighborhood (blue) of an eQTL—a genetic polymorphism that is linked to a differentially expressed target gene (green)—are identified, and a local network connecting all the candidates with the single target gene is constructed. Then, for each candidate, a network based similarity measure is calculated, relating the candidate to the target. The candidate that is most similar to the target gene is picked as the prioritized gene

(see Section 5 for a detailed description). All the alternative paths are merged afterwards to form the local sub-network. Because it is expected that the value of k will influence the size of the sub-network, EPSILON was run with different values for k. Our localization method is expected to mediate the influence of highly connected genes. Even if an unwanted hub were included in the local network connecting a candidate with the target, it will have at most $2k$ connections (in the unlikely case where all k paths enter and leave the hub through different interactions), instead of potentially tens or hundreds of connections. Evidently, the higher the quality of the underlying network, the higher the quality of the local networks will be, but by using a shortest/cheapest path-based approach, unreliable or irrelevant (when connecting a specific causal gene with the target gene) connections will only be followed when there is no more reliable or relevant alternative available. The algorithm will not remove shortcuts (indirect interactions between two genes/proteins that do not reflect a true physical relation) from the local networks, but only when they are present as high-confidence interactions.

A naive nearest neighborhood method (take neighbors, neighbors of neighbors, etc.) was tested, but the presence of hubs in the network rendered this approach useless because the resulting connected sub-networks were almost as large as the global network. Results for this approach were therefore not considered in further analysis.

Because the choice of the measure for strength or confidence influences the outcome of path finding methods, we evaluated three different weight-value schemes. Standard and similar to Tu *et al.* (2006) and Suthram *et al.* (2008), we used expression correlation as edge weights. As a first alternative, we used the confidence scores obtained in the network construction process (see Section 5), and, finally, we tested a simple qualitative weighting scheme of setting edge weights to 1 when an interaction is present and 0 otherwise. Several network-based prioritization methods (e.g. Suthram *et al.*, 2008) apply, before the application of the actual prioritization method, an extra constraint to the interaction network: based on biological considerations, it is demanded that a target gene be reached through a directed protein–DNA (transcription factor) interaction. We tested whether applying this transcription factor-based filtering (TF-filtering) before local sub-network construction could improve prioritization performance.

2.2 Network similarity measures

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. Shih and Parthasarathy, 2012; Suthram *et al.*, 2008; Tu *et al.*, 2006) propose a random walk (RW) approach, in which an RW is initiated a 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. The probability of choosing a particular direction is a function of a weight value associated with each interaction. Obviously, the candidate causal gene that yields the highest number of *arrivals* in the target gene is chosen to be the true causal gene.

Next to integrating RWs 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). As already demonstrated by Köhler *et al.* (2008), Lavi *et al.* (2012), Nitsch *et al.* (2010) and Qi *et al.* (2008), kernels operating on nodes in a biological network can be used as functional gene similarity measures. In this study, we evaluated two well-known kernels, the Regularized Commute-Time (RCT) kernel and the Laplacian Exponential Diffusion (LED) kernel. Both are described in the Section 5. The LED and RCT kernels operate on undirected networks. Kernels operating on directed networks do exist, but their calculation does not scale well for large graphs (Mantrach *et al.*, 2010). Even though we use *undirected* kernels, the local sub-network on which the kernels are calculated is constructed using path finding techniques operating on a *directed* global network. The calculation of the kernels results in a similarity matrix containing the similarity between all possible pairs of genes in the network. From this matrix, the candidate-target pair with the highest similarity is easily found.

3 RESULTS

3.1 Benchmark dataset

To investigate whether reducing the global network to a local network, before applying a network-based similarity measure, can aid in identifying the true causal genes in a set of eQTL, we applied several refinement strategies to the well-characterized dataset of Brem and Kruglyak (2005). They profiled simultaneously the genotype and expression phenotype of 112 yeast segregants. eQTL were determined as described later in the text, and for each eQTL, candidate causal genes were identified (see Section 5). Prioritization resulted for each eQTL-target combination in a single prioritized causal gene.

Each prioritized causal gene–target gene combination was evaluated against the experimental knockout dataset of Hughes *et al.* (2000), as described in Ourfali *et al.* (2007), Suthram *et al.* (2008) and Yeang *et al.* (2004) (In Supplementary Fig. S2, we present a graphical overview of the evaluation strategy.) Here, for each experiment, the knocked-out gene is considered a true causal gene. Genes that are affected by the knockout are assumed to be potential targets. Each knockout experiment thus results in a list of presumably true causal-target gene combinations. For all causal-target knockout combinations in such a list, we look for couples of genes that also appear in the result of an association analysis. This means that we look for those associations for which one of the candidate causal genes is the same as the causal gene in a knockout pair, and for which the target gene is identical to that of the same knockout pair. To make sure that we are evaluating the gene prioritization, and not the quality of the network, the number of combinations is further reduced by demanding that there exist a path in the directed interaction network between the causal and the target gene. These so-called true combinations of causal-target genes derived from the knockout dataset, that also appear in an eQTL-analysis, determine the maximal reference performance. A performant gene prioritization procedure is assumed to approximate this list of gold standard knockout pairs. Performance here is defined as the amount of overlap of

the prioritization results with the target list of knockout pairs or alternatively: *performance = number of retrieved knockout pairs*. Relative performance is then defined as *performance/maximum number of retrievable knockout pairs*. Because a different association technique will result in different associations, and consequently in a different reference list of knockout pairs, absolute performance depends on the association technique used.

3.2 Association analysis

We applied three commonly used association techniques to the SNP and expression data (*Saccharomyces cerevisiae*) of Brem and Kruglyak (2005): non-parametric regression (NPR), mixed models (MM) and elastic net regression (EN) regression (see Section 5).

The genetic variability in the yeast dataset is expected to be low, as a limited number of segregants derived from a cross of only two parental strains were tested. Also, gene expression was determined under a single experimental condition. Consequently, it is likely that only a limited number of true eQTL will be detected. Remarkably, the three association techniques yield different results, as is illustrated in Figure 2. Only 662 eQTL are detected by all three methods, even when for each method, the 10 000 most reliable associations are selected. In general, the NPR results in a high number of associations, both *cis* and *trans*. The MM almost exclusively produces *cis* associations, which confirms the findings of Kang *et al.* (2008) and Listgarten *et al.* (2010). The EN produces a much sparser association map, with less *cis* than *trans* associations.

To make the evaluation of the tested gene refinement strategies independent of the outcome of the association analysis, we tested all described gene prioritization schemes on the results of each association method separately. For each association method, the associations that were prioritized were chosen in such a way (using *P*-values, see Section 5) that the size of the gold standard reference list was of the same order of magnitude (see Supplementary Table S1). As only a small number of knockout pairs can be found overlapping with the EN associations, even with an elevated *P*-value cutoff, we also included the unfiltered EN results (containing all eQTL found by the EN feature selection technique) in any further analysis.

3.3 EPSILON prioritization results

To demonstrate the added value of the EPSILON approach, i.e. extracting a local sub-network before applying a network-based similarity measure, we compared the results obtained using different similarity measures in combination with both local networks and the global network. All local networks were based on the *k*-trials shortest path network construction method outlined above with $k = 5$. The global network corresponded to the original interaction network (and not a local network with a very high *k* value). As an established benchmark strategy, we used a shortest path-based method, which comes down to taking the candidate causal gene that lies closest to the target gene (see Section 5). As a baseline, we performed a random assignment, i.e. randomly picking a candidate from the list of available candidate causal genes. All refinement methods except random assignment were tested with and without TF-filtering [suggested by Suthram *et al.* (2008)], meaning that before local network construction or to the application of a global similarity measure, the global network is modified in such a way that target genes only contain incoming transcription factor interactions. The full prioritization result matrix is shown in Supplementary Table S1. In what follows, we highlight the insights that can be drawn from these results.

Figure 3 shows, in relative terms, the difference in prioritization performance between the use of a local network over a global network. In all cases, over all association mapping techniques, pre-filtering steps and prioritization methods, using a local network is beneficial, with a maximum and minimum gain of 15 and 2 percentage points, both for the EN associations prioritized with a RCT kernel, but respectively with and without TF-filtering. In general, using a local network yields significantly better performance compared with using the global network with an average gain of 8 percentage points (paired two-tailed *t*-test, $P < 10^{-5}$). On average (over all association techniques and TF-filtering possibilities), the RCT kernel yields the highest performance (58%), and the RWs yield the worst (55%); yet, these differences were not statistically significant ($P = 0.13$ three-way analysis of variance). The EN associations were more difficult to prioritize than the associations obtained with MMs and NPR, with a maximum performance of respectively 41, 63 and 55%

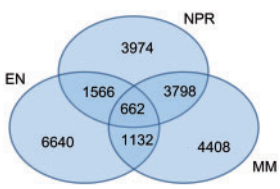

Fig. 2. Agreement between association techniques for the 10 000 most reliable associations per method. NPR, non-parametric regression; MM, mixed model; EN, elastic net

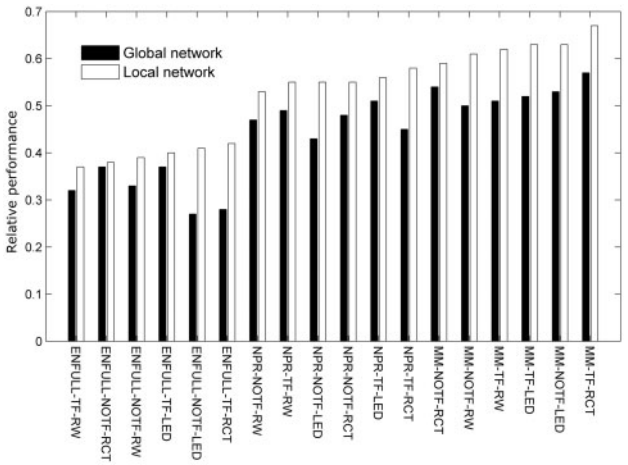

Fig. 3. Effect of using local networks versus using global networks on the relative performance. NPR, non-parametric regression; MM, mixed model; ENFULL, elastic net full result; LED, Laplacian Exponential Diffusion kernel; RCT, Regularized Commute-Time kernel; RW, Random Walk; TF, using transcription factor-based filtering. Relative performance is defined as maximum number of retrieved knockout pairs/maximum number of retrievable knockout pairs

Table 1. Relative prioritization performance for the NPR eQTL

Method	Relative performance
Random assignment	0.18
Shortest path	0.50
Global RW	0.47
Global LED kernel	0.43
Global RCT kernel	0.48
Local RW $k=5$	0.53
Local LED kernel $k=5$, $\alpha=0.001$	0.55
Local RCT kernel $k=5$, $\alpha=0.9$	0.55

Note: All results are calculated using expression-correlation based edge weights, without using TF-filtering. RW, random walk; LED, Laplacian Exponential Diffusion kernel; RCT, Regularized Commute-Time kernel.

(no TF-filter). The average performance (averaged over prioritization methods and association techniques) is only slightly better when TF-filtering is applied than when it is not (50% without TF-filtering, 51% with TF-filtering). Moreover, this difference is not statistically significant (paired two-tailed t-test, $P=0.15$).

As a typical example, we present the results for NPR without TF-filtering in Table 1. The largest overlap with the target list is obtained using a local LED or RCT kernel (55%). As expected, all methods show a much higher overlap with the target list compared with random assignment. The shortest path method performs slightly better (50%) than the global methods; yet, it shows a smaller overlap with the target list than the local methods.

3.4 Size of the local neighborhood

We evaluated the added value of using a relaxed local neighborhood. This was achieved by varying the parameter k in the k-trials shortest path neighborhood construction algorithm between 1 and 20. Local networks generated with different values of k for a single eQTL-target combination are displayed in Figure 4. It can clearly be seen how, for small values of k, local networks remain sparse. Nodes with a high node degree in the global network sometimes (but not always) become hubs again in the local network for large values of k. Indeed when $k \to \infty$, the local network will resemble the global network more closely (see Section 5).

Experiments were run for the eQTL obtained with the NPR using the RW similarity measure and the LED and RCT kernels. The results are displayed in Figure 5. A number of observations can be made. First, the lowest performance is obtained for $k=1$, corresponding to a local network constructed using a single shortest path. Second, although the maximum overall performance is recorded for the RCT kernel, this kernel seems to be more sensitive to the value of k, as more performance variation can be observed. Third, the RCT kernel behaves as expected, i.e. performance is suboptimal for low values of k, reaches a maximum for intermediate values and drops when k becomes too large, and consequently the local networks lose their local characteristics. The RW measure and the LED

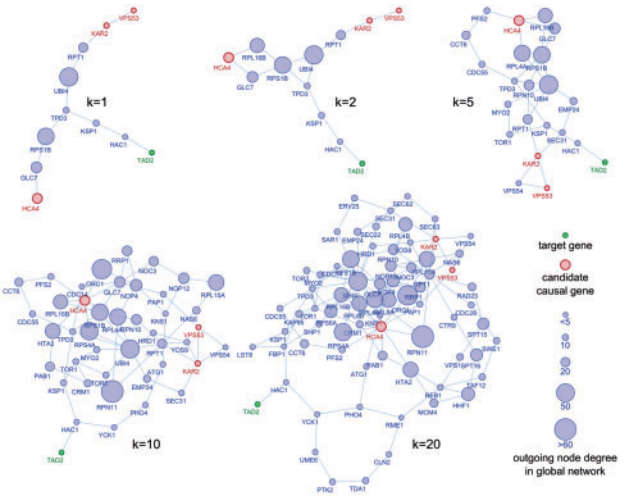

Fig. 4. Example local networks generated with the k-trials shortest path method for different values of k. The diameter of the genes is proportional to the node degree in the global network. Green nodes are target genes, red nodes are candidate causal genes

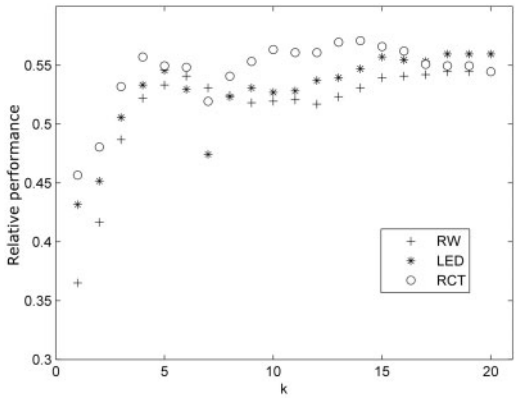

Fig. 5. Influence of the k parameter on prioritization performance for the three local similarity measures. RW, random walk; LED, Laplacian Exponential Diffusion kernel; RCT, Regularized Commute-Time kernel. eQTL were obtained with NPR, edge weights are expression correlation, no TF-filter

kernel do not show this performance drop, but the results of the application of global similarity measures (Table 1) indicate that both suffer from a performance drop when the underlying interaction network becomes too large. Finally, a remarkable performance drop around a k-value of 7 can be observed, for all methods. Closer investigation of the eQTL data and the local networks constructed indicates that this is due to misprioritization of a single eQTL that is shared by a number of genes. In this case, the error is caused by the inclusion of a gene that is relatively distant in a directed network, but close in an undirected network. As explained in the Section 5, the application of kernels implies an undirected network. When the neighborhood again becomes larger, the method quickly recovers from this error.

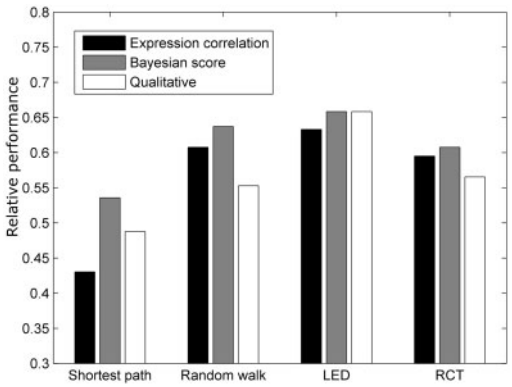

Fig. 6. Influence of the weight type (Expression correlation, Bayesian score and Qualitative) on prioritization performance. eQTL dataset = MM, $P<0.1$, $k=5$. RCT, Regularized Commute-Time kernel; LED, Laplacian Exponential Diffusion kernel

3.5 Edge weight type

The choice of edge weight is non-trivial and an issue for shortest path, RW and kernel-based approaches. We tested three scenarios: (i) edge weights are set proportional to the expression correlation; (ii) weight-values are set to the Bayesian classifier score obtained in the global network construction process (see Section 5); and (iii) values (in the adjacency matrix, see Section 5) are simply set to one if an interaction is present.

In Figure 6, the influence of the choice of weights on the prioritization performance is illustrated. Three edge weight types, expression correlation, Bayesian score and a qualitative approach (with edge weights set to one) and two kernel types, an RW and a shortest path method are evaluated using the MM associations. From the figure, it is clear that the Bayesian score outperforms the expression correlation approach, but differences are small in terms of maximum relative performance (LED kernel: 62% for Expression correlation, 66% for Bayesian score). Remarkably, the qualitative approach where edge weights are simply set to 1 still yields comparable results.

3.6 Comparison with other techniques

To put the result of prioritization with EPSILON in perspective, we compared our approach with two alternative techniques that can be used to perform eQTL prioritization: Information Transduction Module (ITM) Probe (Stojmirović and Yu, 2009, 2012) and eQTL Electrical Diagrams (eQED) (Suthram *et al.*, 2008) (for the implementation details of both methods, see Section 5). ITM Probe was tested in normalized channel mode, in which an RW is executed that has, at each step, a fixed probability of terminating. In this way, ITM Probe is implicitly generating a localized biological context. We obtained results that are in general on par with EPSILON (LED kernel, $k=5$, $\alpha=0.001$, see Table 2). For the MM associations, EPSILON showed increased performance over ITM Probe, with 63 and 55% overlap, respectively, with the target knockout list. Second, we compared EPSILON with eQED. The eQED uses an electric circuit model (operating on a global network) to prioritize candidate causal genes. Unfortunately, eQED does not support the use of phosphorylation interactions (PI), and a thorough comparison was impossible because the network used by EPSILON could not be used as input for eQED. We did perform a limited comparison (on the NPR associations), using a reduced network (see Section 5). We found an overlap of 65% for EPSILON (LED kernel, $k=5$, $\alpha=0.001$) and 54% for eQED.

Table 2. Relative performance comparison of EPSILON (LED kernel, $k=5$, $\alpha=0.001$) with ITM Probe (No TF-filter)

	EPSILON	ITM Probe
NPR	0.55	0.53
EN all features	0.41	0.38
MMs	0.63	0.55

4 DISCUSSION

In this manuscript, we presented EPSILON, a framework for eQTL prioritization that ranks a number of candidate causal genes, overlapping an eQTL based on their functional similarity with the target gene associated with the eQTL. The method consists of two steps: (i) local network construction and (ii) the application of a network-based similarity measure. EPSILON is a modular framework that allows for the incorporation of existing network construction procedures and network similarities. We evaluated a *k*-trials shortest path network construction method together with RW and kernel-based similarity measures using a gold standard dataset derived from a yeast knockout compendium. We were able to show that our approach, the combination of local networks with network-based similarity measures, 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. Highly connected genes (in the global network) may still be present in the local networks, but the number of interactions those genes participate in will be significantly reduced. An additional explanation of the difference between a local and global approach is the inclusion of directional information in the sub-network construction, even though direction information is not used for the calculation of the similarity measures. There are indications that the LED kernel outperforms the RW similarity measure, but more research is needed to confirm this.

We observed large differences in prioritization performance when different eQTL datasets were used (MM, NPR and EN). Especially the EN eQTL seemed to be more difficult to prioritize. Possibly, the quality of the EN associations was lower than that of the other association techniques.

All experiments were run with and without TF-filtering. TF-filtering is a technique used by several other authors where it is demanded that a target gene be reached through transcription factor interactions. We were not able to detect any significant effect of TF-filtering on the results obtained.

We investigated the sensitivity of the local similarity measures to the value of the parameter k, controlling the size of the local sub-network in the network construction algorithm. All network similarity measures yield a near optimal performance for k-values of 5, the value we recommend to use as the running time of the local network construction and the similarity measure calculation increases significantly with higher values of k.

Next, we investigated the influence of different edge weight types on the prioritization process. Three variants were evaluated: expression correlation, a Bayesian score and a qualitative weighting scheme where all edges were set to one. Remarkably, in some cases, optimal results were obtained with the qualitative weighting scheme. This indicates that the mere presence of an interaction is more important than its associated weight. In the absence of reliable edge weights, one can thus revert to qualitative edge weights, simply expressing the presence or absence of interactions.

Finally, EPSILON was compared with 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 PIs present in the global network.

Importantly, we should note that the intuitive network construction method we present could be replaced by a more advanced algorithm; yet, our approach has the advantage of simplicity and performance. The network-based similarities too could be replaced by more specialized variants. For instance, the eQED method of Suthram *et al.* (2008) could be integrated within the EPSILON framework, given that the eQED model was extended to incorporate other types of interactions.

5 MATERIALS AND METHODS

5.1 eQTL analysis

To detect associations between SNP and gene expression data, we used the *S. cerevisiae* data of Brem and Kruglyak (2005). The SNP dataset consists of 2956 individual markers for 112 segregants from a laboratory strain and a wild strain. As strong linkage disequilibrium can be observed, the dataset exhibits substantial redundancy. We reduced the dataset to 404 haplotype blocks using estimates of the recombination probability at each marker. Reducing individual markers to haplotype blocks spanning multiple SNPs inherited together avoids part of the multiple testing problem associated with many eQTL detection methods, but will increase the genomic region covered by the eQTL found.

The gene expression data contained full genome expression values (6130 genes) for the same 112 segregants. Before further analysis, the expression data were re-normalized. Associations were detected using three methods. As a reference, we used the NPR technique (Wilcoxon rank regression) as applied by Brem *et al.* (2002). Next, analogous to Suthram *et al.* (2008), the MM approach of Kang *et al.* (2008) was adopted. This method is known to correct for possibly hidden confounds like population structure or laboratory batches. Finally, we choose EN regression as a state-of-the-art association technique (Basu *et al.*, 2011; Michaelson *et al.*, 2010; Wu *et al.*, 2009). The EN regression was executed to retrieve a maximum of 10 associations per target gene.

For the MM regression, we used the implementation of Kang *et al.* (2008), which can be found at http://genetics.cs.ucla.edu/ice. For the non-parametric and the MM regression, a matrix of association probabilities was obtained in a straightforward way. Probabilities were filtered using $P<0.05$ and $P<0.1$ for the NPR and the MMs, respectively. The thresholds for the P-values were chosen so that the remaining associations contained a comparable number of knockout pairs that can be used for evaluation purposes, as explained in the Sections 3 and 4. For the EN, probabilities were approximated by applying a multiple regression to the selected predictors, a strategy suggested by Wu *et al.* (2009). EN probabilities were then filtered based on their approximated P-value ($P<0.1$). Because this resulted in a low number of remaining associations, we also ran the prioritization method on the full EN dataset, i.e. without any P-value filtering.

5.2 Generating candidate causal genes

For each eQTL, we enumerated all genes that overlap (within 5 kb) the region of the chromosome spanned by the haplotype block that associates with a certain target gene. This approach differs slightly from the one adopted by Suthram *et al.* (2008), where a locus contains all genes that are related to the same closest marker (within a buffer of 10 kb). Our approach will necessarily yield larger loci with a higher number of candidates.

5.3 Network construction

We built a physical interaction network for *S. cerevisiae* containing protein–protein interactions (PPI), transcription factor–DNA interactions (TF–DNA) and PI. In the network, a gene name can represent both a translated protein and a chunk of DNA. Protein interaction data were downloaded from The Biological General Repository for Interaction Data sets database (Stark *et al.*, 2011). After removing self-interactions, this PPI dataset consisted of 49 381 unique interactions. TF–DNA interactions, predicted both *de novo* as from several ChIP-chip experiments were obtained from Beyer *et al.* (2006). Their dataset consisted of 7817 high confident TF–DNA interaction predictions. Kinase–substrate interaction data were obtained from Ptacek *et al.* (2005). With the use of proteome chip technology, the authors identified 4183 *in vitro* phosphorylation events involving 1325 different proteins in yeast with a low false positive rate. Only high-quality PPI interactions were allowed in the final interaction network. This was achieved by using a supervised Bayesian classifier that was trained on a set of known interactions derived from literature (Reguly *et al.*, 2006). As inputs, the classifier uses features derived from expression data, known and predicted domain interactions, network topology and phylogenetic profile similarity. For a more elaborate discussion of the network construction, see Supplementary Data. Our final network contains 4375 genes (connected by 35 569 interactions) that are also present in the eQTL datasets. Contrary to Suthram *et al.* (2008) who selected only regulatory proteins contained in 11 Munich Information Center for Protein Sequences (MIPS) categories, we did not filter the network to only contain interactions found in regulatory pathways. For all experiments except one, we used the inverse of the absolute value of the expression correlation between the two genes involved in the interaction for the edge weights. In a single experiment, we evaluated the choice of weight value by comparing it to two other variants: (i) edge weights are set to $1/(score - minimumScore + 0.1)$, using the score obtained by the Bayesian classifier discussed before and (ii) the edges are simply given a value of one.

5.4 Validation data

To evaluate the proposed method, we used the knockout dataset of Hughes *et al.* (2000). The data were preprocessed by Ourfali *et al.* (2007) and can be downloaded from http://cs.tau.ac.il/~roded/SPINE.html. Each time a variant of the proposed method or a reference method is evaluated, the knockout dataset is filtered to contain only the knockout pairs that are also found in the network and the eQTL datasets (see Sections 3 and 4).

5.5 k-trials shortest paths neighborhoods

An intuitive algorithm was constructed that tries k times to find an alternative path between each candidate causal gene and the target gene using

the Dijkstra shortest path algorithm (Dijkstra, 1959). In what follows, we will use path cost in stead of path length, as this allows for easier interpretation of the algorithm. Each time, the cheapest path is made more expensive by multiplying each edge weight on the path with a constant factor f (set to 1.2 in this study). The algorithm results in at most k alternative paths connecting the target with each of the candidates. Often, values of $k > 1$ will not result in exactly k alternative paths, as it is possible that no (cheap) alternative paths exist. For a k value of 5, the average number of alternative paths was between 2 and 3, for the datasets we used. Once the alternative paths are found, a local neighborhood is constructed by merging all paths found into a single sub-network. Large values of k will result in larger sub-networks that will be more similar, but not necessarily identical, to the global network.

The parameter k is an upper bound of the number of paths that will be included in the local network, and the parameter f controls how different the alternative paths should be. Values of f very close to 1 will likely yield a single cheapest path or small variants of the cheapest path (unless paths with equal cost exist, see later in the text), as the same path is found over and over, even when it is made more expensive. The parameter f controls how much more costly path k can be for it to be picked, when compared with the first cheapest path found (i.e. f^k). The k-trials shortest paths algorithm can be considered a heuristic variant of the more complicated k-shortest path problem (see e.g. Hershberger *et al.*, 2007), but the latter will always yield k paths (at least when k paths exist), even when some of those paths are less relevant. Also, by making the entire path more expensive, the k-trials algorithm is pushed to find real alternative paths instead of variants of an existing path. Paths with equal costs were not treated in a special way because this is unlikely to happen when expression correlation or a Bayesian score is used for edge weights. Furthermore, it such paths would be found (which is more likely when, for example, all edge weights are set to 1), sufficiently large values of k will alleviate the problem.

5.6 RWs and kernels

The RW similarity measure is an estimate of the probability that a target gene is reached when initiating an RW starting from different candidate causal genes. For each local network, new RWs were initiated from the target gene until the sum of the arrival count over all candidate causal genes was equal to 10 000. Using the target gene as the starting point of the RW is possible because we used undirected networks as input for the RW procedure. Once a candidate is reached, the RW is aborted; therefore, no walk visits a candidate gene more than once. The candidate with the highest arrival count is chosen to be the true causal gene. The approach described earlier in the text is less accurate and slow for large global networks. As an alternative to simulating RWs, it is possible to calculate the final number of visits in each network node based on the (Moore-Penrose) pseudo-inverse of the weighted Laplacian matrix $L = D - A$, with D the diagonal degree matrix ($D(i,i) = \sum_{j=1}^{n} A(i,j)$, n is the number of genes) and A the adjacency matrix of the interaction network. For each eQTL, the global matrix A was updated in such a way that the candidate causal genes in the eQTL had no outgoing connections. This corresponds to terminating a RW once a candidate gene is reached.

Next to RWs, we integrated kernel-based similarity measures in the EPSILON framework. A kernel calculated on a graph provides a quantitative measure of similarity between two nodes (i.e. genes) in a graph, which is also a vector inner product in a possibly high dimensional feature space. Depending on the type of kernel, the similarity measure can be interpreted in different physical ways. Fouss *et al.* (2006) provides a comprehensive overview of different kernel types and their interpretation. We decided to use two kernel types that have been used for gene prioritization before (Nitsch *et al.*, 2010), be it in a different context and methodology. The first is the LED Kernel that is calculated as follows:

$$K_{LED} = e^{-\alpha L} \tag{1}$$

K_{LED} is the kernel matrix that contains the evaluation of the kernel function for each gene–gene combination. It is calculated on the weighted Laplacian matrix. The adjacency matrix has been made symmetrical by taking the per element maximum of the non-symmetrical adjacency matrix and its transpose. $K_{LED}(i,j)$ contains, at time $t = \alpha$, the quantity found in node i when a unit quantity starts diffusing from node j at $t = 0$ (Fouss *et al.*, 2006).

Analogous to Nitsch *et al.* (2010), the second kernel we use is the RCT Kernel:

$$K_{RCT} = (D - \alpha A)^{-1} \tag{2}$$

Here, K_{RCT} is the kernel matrix, D the degree matrix defined above, and A the adjacency matrix. $K_{RCT}(i,j)$ corresponds to the probability of visiting node i when starting from node j where a random walker has $\alpha(1-\alpha)$ probability of disappearing at each step (Fouss *et al.*, 2006). Both kernels are centered and normalized in feature space by applying the following transformations:

$$K_N(i,j) = \frac{k_{ij}}{\sqrt{k_{ii}.k_{jj}}} \tag{3}$$

$$K_C = HKH\,,\ H = I - \frac{1}{n}ee^T \tag{4}$$

5.7 Comparison with other techniques

We compared EPSILON with ITM Probe (Stojmirović and Yu, 2009, 2012) and eQED (Suthram *et al.*, 2008), two methods that can be used to identify the true causal gene in an eQTL. The stand-alone version of the ITM Probe software (Python, v 1.5.2) was downloaded from ftp://ftp.ncbi.nlm.nih.gov/pub/qmbpmn/qmbpmn-tools/src and slightly modified (input/output routines only) to accommodate our experimental setup. We ran the model in *normalized channel* mode. For each eQTL-target pair, we set the ITM Probe source gene to the target gene, and the ITM Probe sinks to the candidate causal genes for the eQTL. The final causal gene corresponds to the sink with the maximum value in the resulting H matrix. In normalized channel mode, the ITM Probe model is controlled by a single parameter, the damping factor *df*. We ran the model with *df* values ranging from 0.1 to 1.0 and recorded for each association technique the optimal result. The model was fed with the exact same networks (no TF-filtering) as used to evaluate EPSILON.

Next to ITM Probe, we investigated whether it was possible to compare EPSILON with eQED. The software was downloaded from http://www.stanford.edu/ssuthram/eQED/. eQED does not allow for the distinction between directed transcription factor interactions and directed PIs. This distinction is needed because eQED always applies TF-filtering, treating all directed interactions as TF-interactions. Consequently, the interaction network used in this study could not be used as input for eQED. We tested eQED using a network without PIs and the NPR eQTLs and reran EPSILON using the same reduced network.

5.8 Parameter tuning and running time

The calculation of the two kernel types, K_{LED} and K_{RCT}, requires the estimation of a parameter α. An extensive parameter sweep was performed. We found an optimal, general applicable value of the parameter α to be 0.001 for the LED kernel and 0.9 for the RCT kernel. We also found that the kernel-based similarity measure is not sensitive to this parameter α, as performance deteriorates only when α is changed with an order of magnitude. Testing indicated that near optimal performance can be obtained with a value of 5 for the k parameter in the k trials shortest path neighborhood construction algorithm (see Fig. 5) and 1.2–2.0 for the multiplier constant. As the running time of the algorithm increases considerably with higher values of k, all experiments were

executed with the lowest k value that results in optimal performance, i.e. $k = 5$. For the multiplier, 1.2 was chosen.

Even though the time complexity of calculation of the kernels used will be of order $O(V_l^3)$ with V_l the number of genes in the local network, the running time of EPSILON in practice (for reasonably small values of k yielding small local networks, and for a single eQTL-target prioritization task) is determined by the path-finding algorithm. We implemented the Dijkstra path-finding algorithm using a Fibonacci heap, resulting in a time complexity of $O(k(E_g + V_g.log(V_g)))$ with V_g the number of genes in the global network, and E_g the number of interactions in the global network. With $k = 5$, it takes 40 seconds to prioritize 1000 eQTL-target combinations using a single core of a 64 bit Intel i3 processor.

ACKNOWLEDGEMENTS

The authors thank the anonymous reviewers for the constructive criticism and many suggestions that improved the quality of the manuscript.

Funding: (i) Ghent University Multidisciplinary Research Partnership 'Bioinformatics: from nucleotides to networks'; (ii) Fonds Wetenschappelijk Onderzoek-Vlaanderen (FWO) G.0329.09; and (iii) Katholieke Universiteit Leuven funding: PF/10/010 (NATAR). The computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by Ghent University, the Hercules Foundation and the Flemish Government—department EWI. John Boyer from the IBM Canada Software Lab is the author of the original Dijkstra implementation the authors based their path finding algorithms on.

Conflict of Interest: none declared.

REFERENCES

Basu,S. *et al.* (2011) Multilocus association testing with penalized regression. *Genet. Epidemiol.*, **35**, 755–765.

Beyer,A. *et al.* (2006) Integrated assessment and prediction of transcription factor binding. *PLoS Comput. Biol.*, **2**, e70.

Brem,R.B. and Kruglyak,L. (2005) The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc. Natl Acad. Sci. USA*, **102**, 1572–1577.

Brem,R.B. *et al.* (2002) Genetic dissection of transcriptional regulation in budding yeast. *Science*, **296**, 752–755.

Dijkstra,E.W. (1959) A note on two problems in connexion with graphs. *Numer. Math.*, **1**, 269–271.

Fouss,F. *et al.* (2006) An experimental investigation of graph kernels on a collaborative recommendation task. In: *Proceedings of the 6th IEEE International Conference on Data Mining (ICDM): 18–22 December 2006*, p. 863–868. IEEE Computer Society, Hong Kong.

Gillis,J. and Pavlidis,P. (2011) The impact of multifunctional genes on "guilt by association" analysis. *PloS ONE*, **6**, e17258.

Hershberger,J. *et al.* (2007) Finding the k shortest simple paths. *ACM Trans Algorithms*, **3**, article no. 45.

Hughes,T.R. *et al.* (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.

Kang,H.M. *et al.* (2008) Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. *Genetics*, **180**, 1909–1925.

Köhler,S. *et al.* (2008) Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet.*, **82**, 949–958.

Lavi,O. *et al.* (2012) Network-induced classification kernels for gene expression profile analysis. *J. Comput. Biol.*, **19**, 694–709.

Listgarten,J. *et al.* (2010) Correction for hidden confounders in the genetic analysis of gene expression. *Proc. Natl Acad. Sci. USA*, **107**, 16465–16470.

Mantrach,A. *et al.* (2010) The sum-over-paths covariance kernel: a novel covariance measure between nodes of a directed graph. *IEEE Trans. Pattern Anal. Mach. Intell.*, **32**, 1112–1126.

Michaelson,J.J. *et al.* (2009) Detection and interpretation of expression quantitative trait loci (eQTL). *Methods*, **48**, 265–276.

Michaelson,J.J. *et al.* (2010) Data-driven assessment of eQTL mapping methods. *BMC Genomics*, **11**, 502.

Nitsch,D. *et al.* (2010) Candidate gene prioritization by network analysis of differential expression using machine learning approaches. *BMC Bioinformatics*, **11**, 460.

Ourfali,O. *et al.* (2007) SPINE: a framework for signaling-regulatory pathway inference from cause-effect experiments. *Bioinformatics*, **23**, i359–i366.

Ptacek,J. *et al.* (2005) Global analysis of protein phosphorylation in yeast. *Nature*, **438**, 679–684.

Qi,Y. *et al.* (2008) Finding friends and enemies in an enemies-only network: a graph diffusion kernel for predicting novel genetic interactions and co-complex membership from yeast genetic interactions. *Genome Res.*, **18**, 1991–2004.

Reguly,T. *et al.* (2006) Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae. *J. Biol.*, **5**, 11.

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

Stark,C. *et al.* (2011) The BioGRID interaction database: 2011 update. *Nucleic Acids Res.*, **39**, D698–D704.

Stojmirović,A. and Yu,Y.-K. (2009) ITM Probe: analyzing information flow in protein networks. *Bioinformatics*, **25**, 2447–2449.

Stojmirović,A. and Yu,Y.-K. (2012) Information flow in interaction networks II: channels, path lengths, and potentials. *J. Comput. Biol.*, **19**, 379–403.

Suthram,S. *et al.* (2008) eQED: an efficient method for interpreting eQTL associations using protein networks. *Mol. Syst. Biol.*, **4**, 162.

Tranchevent,L.-C. *et al.* (2011) A guide to web tools to prioritize candidate genes. *Brief. Bioinform.*, **12**, 22–32.

Tu,Z. *et al.* (2006) An integrative approach for causal gene identification and gene regulatory pathway inference. *Bioinformatics*, **22**, e489–e496.

Voevodski,K. *et al.* (2009) Spectral affinity in protein networks. *BMC Syst. Biol.*, **3**, 112.

Wu,T.T. *et al.* (2009) Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, **25**, 714–721.

Yeang,C.-H. *et al.* (2004) Physical network models. *J. Comput. Biol.*, **11**, 243–262.

Zotenko,E. *et al.* (2008) Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality. *PLoS Comput. Biol.*, **4**, e1000140.