Molecular Biosystems

Accepted Manuscript

This is an *Accepted Manuscript*, which has been through the RSC Publishing peer review process and has been accepted for publication.

Accepted Manuscripts are published online shortly after acceptance, which is prior to technical editing, formatting and proof reading. This free service from RSC Publishing allows authors to make their results available to the community, in citable form, before publication of the edited article. This *Accepted Manuscript* will be replaced by the edited and formatted *Advance Article* as soon as this is available.

To cite this manuscript please use its permanent Digital Object Identifier (DOI®), which is identical for all formats of publication.

More information about *Accepted Manuscripts* can be found in the [Information for Authors](#).

Please note that technical editing may introduce minor changes to the text and/or graphics contained in the manuscript submitted by the author(s) which may alter content, and that the standard [Terms & Conditions](#) and the [ethical guidelines](#) that apply to the journal are still applicable. In no event shall the RSC be held responsible for any errors or omissions in these *Accepted Manuscript* manuscripts or any consequences arising from the use of any information contained in them.

PheNetic: Network-based interpretation of unstructured gene lists in *E. coli*Dries De Maeyer¹, Joris Renkens², Lore Cloots¹, Luc De Raedt², Kathleen Marchal^{1,3}¹Center of Microbial and Plant Genetics, Kasteelpark Arenberg 20, B-3001, Leuven, Belgium²Department of Computer Science, Katholieke Universiteit Leuven, Celestijnenlaan 200A, B-3001 Heverlee, Belgium³Department of Plant Biotechnology and Bioinformatics, Ghent University, Technologiepark 927, 9052 Gent, Belgium

Abstract

At the present time, omics experiments are commonly used in wet lab practice to identify leads involved in interesting phenotypes. These omics experiments often result in unstructured gene lists of which interpretation in terms of pathways or mode of action is challenging. To aid in the interpretation of such gene lists, we developed PheNetic, a decision theoretic method that exploits publicly available information, captured in a comprehensive interaction network to obtain a mechanistic view on the listed genes. PheNetic selects from a comprehensive interaction network the sub-networks highlighted by these gene lists. We applied PheNetic on an *Escherichia coli* interaction network to reanalyse a previously published KO compendium, assessing gene expression of 27 *E. coli* knock-out mutants under mild acidic conditions. Being able to unveil previously described mechanisms involved in acid resistance demonstrated both the performance of our method and the added value of our integrated *E. coli* network. PheNetic is available at <http://bioi.biw.kuleuven.be/~driesdm/phenetic/>.

Keywords

Microbial systems biology, *E. coli* network, network-based data integration,

Introduction

Omics experiments (e.g. gene expression profiling experiments) are customarily used in wet lab practice to identify leads involved in interesting phenotypes. The interpretation of the gene lists, resulting from these omics experiments is challenging. Currently, most studies perform enrichment analysis on these gene lists to identify overrepresented pathways and/or functional classes¹. Enrichment analysis however, does not provide insights in the molecular interactions that result in the observed phenotype. In addition, it heavily relies on a priori information collected in databases (such as Gene Ontology lists, KEGG pathways)^{2,3}.

To allow for a more mechanistic interpretation of in-house generated omics datasets, increasingly network-based approaches are being used^{4,5}. These methods integrate in-house generated gene lists with publicly available information on the organism of interest. This public information is typically represented as an interaction network that ideally covers multiple molecular interaction layers^{5,4,6}. Different methods have been developed to query these networks with in-house data, all of which rely on the assumption that genes close to each other in the interaction network are related to the same process (guilt-by association). Cluster techniques^{7,8} test to what extent genes from an input list group together in the interaction network. Path-finding approaches^{9–15} subdivide the input gene lists in causes (e.g. genes carrying a mutation) and effects (e.g. genes affected by the mutation) which they attempt to connect through active paths in the interaction network.

These network-based analysis approaches have previously been developed for the analysis of experimental datasets in human and yeast e.g. for the interpretation of results from differential expression analysis^{9,12,13}, genome wide association studies¹⁶ and eQTL analysis (expression quantitative trait loci)^{11,17}. Their application in prokaryotes has remained largely unexploited. In the present paper, we therefore developed a network-based approach for the analysis of *E. coli* datasets. A comprehensive *E. coli* interaction network derived from publicly available omics data was constructed and queried using a novel path-finding approach, PheNetic. We showed the potential of PheNetic in interpreting in house generated datasets by reanalysing a previously published knock-out (KO) expression profiling experiment in combination with the constructed *E. coli* interaction network.

Results

Method description

PheNetic is a sub-network selection algorithm to interrogate an interaction network, compiled from publicly available data, with sets of cause-effect pairs resulting from in-house experiments (see Figure 1). Here, a cause corresponds to a mutation that is expected to trigger an alteration in downstream genes. If the alteration affects the expression level, the downstream genes will be visible in an expression profiling experiment (and referred to as effects). The network, in which nodes represent genes and edges the interactions between the nodes, will be used as a scaffold to connect causes to their effects. Every edge in the network is assigned a probability that expresses our belief in the interaction being truly present in the organisms interactome⁹ and each node is annotated with a probability that reflects its centrality in the network (see Material and Methods). The goal is to extract from this interaction network, the sub-network involved in transducing the perturbation from the causes to their corresponding effects. This sub-network will comprise genes related to the processes highlighted by the in house dataset.

Because of the probabilistic nature of the network, we can obtain for each path in the interaction network a probability (see Material and Methods). This probability is determined by the probabilities of the edges, expressing the belief in the edge, and the nodes, expressing the network centrality of the node, composing the path. The latter term penalizes paths through highly connected or hub nodes. By doing so the amount of redundancy between paths is quantified in the probability of a path. Highly probable paths will thus avoid hub nodes as these have low centrality probabilities. Given the probabilities on the paths we can formulate the sub-network selection problem as a path-finding problem in the decision theoretic framework of DT ProbLog¹⁸. Briefly, we first determine for each cause-effect pair the set of most likely paths that connect them. Subsequently, we merge these paths into a sub-network in which causes (i.e. mutated genes) are connected to the most and preferentially strongest differentially expressed effects using the most probable paths in a parsimonious way (using the smallest sub-network).

This is achieved by assigning a reward to the selected sub-network based on the cause-effect pairs that are connected in the sub-network. Cause-effect pairs connected with a high probability, and having a high level of differential expression will obtain higher rewards. On the other hand a cost will be assigned with increasing size of the selected sub-network. By maximizing the reward minus the cost, the most sparse sub-network will be selected that best explains the input data (see Material and Methods). The motivation for selecting the most parsimonious solution is based on the assumption that all cause-effect pairs are involved in the same phenotype and therefore should trigger common paths in the interaction network.

Network analysis in *Escherichia coli*

To use PheNetic on *E. coli* datasets, we compiled a comprehensive interaction network for this organism from publicly available omics datasets and predictions (see Materials and Methods). The network, consisting of 16794 physical or metabolic interactions between 3063 nodes, covers protein-protein, transcriptional and metabolic interactions.

This network was used in combination with PheNetic to reanalyse the KO compendium published by Stincone et al. This dataset¹⁹ profiles the expression of 27 *E. coli* KO strains, known to be involved in acid resistance (referred to as causes). For each KO strain, expression was compared to that of the wild type strain under similar conditions¹⁹, resulting in lists of genes differentially expressed between wild type and KO strain (referred to as effects). As all mutated genes were supposed to be involved in the same acid resistance phenotype, the cause-effect pairs of the 27 different experiments were pooled in a single list of cause-effect pairs which was then interpreted by means of the interaction network (see Material and Methods).

To optimize parameter selection and test algorithmic performance, two benchmark sets were defined consisting of genes previously associated with acid resistance in *E. coli*. A first stringent, but small benchmark consisting of 53 genes was based on literature curated information. A second more relaxed benchmark was composed of genes, reported to be differentially expressed genes under acid conditions in studies, other than the one of Stincone et al., 2011¹⁹. Algorithmic performance was assessed using receiver-operator curve (ROC) plots that evaluate the trade-off between sensitivity (the fraction of the genes in the benchmark recovered by PheNetic) versus the false positive rate (FPR) (fraction of false positives amongst the total number of genes predicted to be involved in acid resistance by PheNetic). To define the FPR we used a rather conservative definition by assuming that all genes predicted to be involved in acid resistance, other than the ones in the benchmark were false positives.

Parameter optimization

The benchmark sets were used to assess the effect of parameters and reward functions on the algorithmic performance.

Firstly, the effect of the number of cause-effect pairs in the input was tested. The obtained results show that more cause-effect pairs as input allows reaching a higher sensitivity for the same FPR on both the differential expression and literature benchmark (see Figure 2). This means that for the same network size more genes associated with acid resistance are selected (see Figure 2). This indicates that PheNetic is very robust towards the definition of what is 'the most differentially expressed gene set', indicating that ideally all cause-effect pairs could be used. In our set up we limited the number of cause-effect pairs per mutant in the input to 1000, as this offered a good trade-off between computation time and benchmark performance. For the acid resistance data, we thus used as input for PheNetic a pool of 27000 cause-effect pairs.

Secondly, the effect of the gene selection cost was tested. As PheNetic imposes a trade-off between explaining as many as possible cause-effect pairs (reward term), while keeping the network sparse (cost term), an increasing cost term results in selecting gradually less genes, and thus smaller selected sub-networks (Supplementary Figure S1). At a high gene selection cost only those genes will be added that are located at the crossroad of several paths between different cause-effect pairs. At a lower cost this selection becomes less stringent and also genes located on less common paths between cause-effect pairs can be selected into the resulting sub-network. Sweeping over the gene selection cost thus provides a way of ranking genes

based on the largest cost or smallest sub-network size at which they were first selected into the sub-network.

Thirdly, the reward function defines the rewards assigned to each individually explained cause-effect pair. Different reward functions were tested: a constant function that assigns an equal reward to all cause-effect pairs and two gene-specific reward functions, that weigh the importance of explaining the cause-effect pair based on either the absolute value of the differential expression of the effect gene or based on the power of this differential expression (here 5th power) (Figure S2). Using the differential expression to weigh the reward of explaining a particular effect clearly outperforms a constant reward function. It favours the selection of genes involved in true signalling paths that connect causes to the most differentially expressed effects (as a better performance on the benchmark is obtained). Empirically we found that using the fifth power of the absolute log ratio improved the performance of PheNetic over just using the log ratio, as this reward function assigns an even higher weight to the effect genes with high differential expression.

Results of all parameter tests were consistent on both benchmark sets. The following parameter settings, resulting in the best performance on the benchmark sets were used in the remainder of the article (that is using as input for each KO strain, the 1000 most differentially expressed genes, and a gene-specific reward function based on the fifth power of the expression ratio of the effect genes). Using these parameters, the algorithm was run for gradually decreasing gene selection costs, which allowed ranking genes based on the highest cost at which they were first selected in the sub-network.

Comparison with state-of-the-art

To assess the performance of PheNetic, its results were compared with those obtained using eResponseNet²⁰. This state-of-the-art method uses a minimum-cost flow optimization algorithm to connect causes to effects in an interaction network. Similar to PheNetic, eResponseNet employs a user-defined parameter (gamma) to tune the size of the selected sub-networks. Performing a sweep over gamma or the gene selection cost for respectively eResponseNet or PheNetic allowed prioritizing genes, according to their relevancy to the studied process. Genes added to the sub-network with the most stringent gamma parameter (eResponseNet) or gene selection cost (PheNetic) can be considered most reliable and receive the highest ranks. The resulting ranked gene lists were compared to those obtained by a differential expression-based ranking (see Materials and Methods). The latter comparison allowed us to assess the added value of using an interaction network over using mere expression data as is the case with a differential expression-based ranking. An overview of the ranked gene lists obtained by the different methods can be found in supplementary file SF3.

The performance of the different methods was compared on both the literature and differential expression benchmark sets. For each algorithm the relation between its sensitivity (as an estimate of the ability to recover genes, previously associated with acid resistance) versus its FPR (as an estimate of the number of falsely predicted genes) was tested as a function of the selected sub-network sizes (see Figure 3). Prior to the comparison, we tuned the parameters of

eResponseNet to achieve maximal performance on the benchmark sets (see Material and Methods). For eResponseNet only results obtained with its best performing parameter settings are shown.

On both benchmark sets, but most pronouncedly on the literature benchmark set, PheNetic obtained a higher sensitivity for the same FPR than eResponseNet and the differential expression-based ranking. This difference in performance is most pronounced for small sub-networks, representing the most reliable set of nodes involved in the process of interest. For larger sub-networks, the three tested methods reach similar performances. The fact that this higher performance of PheNetic is mainly visible on the literature benchmark set, reflects the intrinsic difference of PheNetic with the differential expression-based ranking. The latter method by definition only recovers differentially expressed genes related to the process of interest, whereas both network-based approaches (PheNetic and eResponseNet) also select genes not regulated at their expression level. The latter type of genes are present in the literature benchmark set, but absent from the differential expression benchmark set.

This intrinsic difference between network and differential expression-based ranking methods is also illustrated by their different ability to identify regulators or signal transducers. These classes of genes are known not to require drastic changes in their expression level to modulate their activity. Some well-known acid resistance regulators, such as YdeO²¹, TorR²², RcsB²³ and RpoD²⁴, are indeed only present in the literature benchmark set, and not in the differential expression benchmark set, indicating they are not or barely altered at their expression level. The ability of the different methods to select these regulators into a sub-network is quantified in Table 1. Each regulator is ranked based on the total number of genes that have to be selected by the subsequent method to select the regulator into the solution. Network methods clearly rank the previously mentioned regulators higher than differential-expression based ranking does.

When comparing both network-based methods, the main difference between PheNetic and eResponseNet is the way cause-effect pairs are interpreted as input. PheNetic conserves the cause-effect pairs when searching paths in the network, while eResponseNet considers a flow between all causes and all effects. eResponseNet thus assumes all causes to be related to all effects, irrespective of whether a direct measurement for this relation was available in the experiments. The effect of this assumption explains the serious drop in performance of eResponseNet when used with more cause-effect pairs (50 instead of 10 per KO strain (Figure S3)). In those cases, eResponseNet will start connecting causes to biologically unrelated effects, resulting in a serious increase in the FPR. Using only a limited set of input pairs partially solves this problem, but results in a lower sensitivity compared to what can be recovered by PheNetic (due to the reduction of input data). This also implicates that the results of eResponseNet are less robust than those of PheNetic towards variations in the number of cause-effect pairs used in the input.

Biological relevance of the extracted sub-networks

To assess the biological relevance of the sub-networks extracted by the different methods, we tested to what extent sub-networks selected by either method covered the same functionalities. To allow for a fair comparison, sub-networks of similar size were selected for the different methods (see Material and Methods). This resulted in a sub-network consisting of 287 genes for PheNetic (Figure S5), of 271 genes for eResponNet (Figure S7) and of 293 genes for the differential expression-based ranking (Figure S6). All sub-networks selected by each of the three methods show clear enrichment in GO terms known to be involved in acid resistance, such as response to pH, regulation of pH, amino acid metabolism^{25,26}, flagellar motility²⁶, oxidation of organic compounds²⁶ and cellular homeostasis²⁷. Compared to the sub-networks derived by ResponseNet and the differential expression-based ranking, the sub-network selected by PheNetic showed a higher enrichment in terms related to amino acid catabolism (GO:9063), more specifically in terms associated with metabolism and catabolism of glutamine (GO:9064, GO:9065) and arginine (GO:6525, GO:6527), both well studied acid resistance mechanisms^{24,27}. For both network-based approaches, but not the differential expression-based ranking, the inferred sub-networks were also enriched in terms related to 'regulation and signalling'. These processes consist of genes known to be barely altered themselves at expression level and, therefore only recovered by the network-based approaches. PheNetic promotes the finding of sub-networks, containing nodes, located at the crossroads of paths that explain several cause-effect pairs together. Therefore, it focuses more on processes shared by several mutants than is the case for the sub-networks extracted by the two other methods.

Detailed description of sub-network extracted by PheNetic

Figure 4 gives a detailed view on the sub-network selected by PheNetic using the most optimal parameter settings. For visualisation purposes, the selected sub-network was decomposed into different gene groups (subgroups) centred around genes in the sub-network, belonging to GO categories found to be enriched in the sub-network, together with their direct neighbours in the selected sub-network. It is clear that the three major processes identified in this decomposition clearly overlap with the previous processes identified by Stincone et al.¹⁹: amino acid metabolism^{25,26}, TCA cycle^{25,26,28}, and flagellae and motility^{26,29,30}. This visualization shows how an in-house performed omics experiment can be interpreted in the context of the network: not only the processes relevant to the phenotype of interest, but also their relations become apparent.

The subgroup centred around genes associated with amino acid metabolism clearly shows the link between amino acid metabolism and the TCA cycle. This subgroup not only contains genes involved in glutamine³¹, arginine³² and cadaverine³³ metabolism which have previously been associated with acid resistance, but it also infers tryptophan, threonine/serine, proline and succinate synthesis to be involved in acid resistance. As the genes related to the latter processes are highlighted by the expression data and located in the network neighbourhood of genes and processes already known to be involved in acid resistance, their link to acid resistance is likely.

A next subgroup centred around genes associated with cellular respiration recapitulates anaerobic and micro-aerobic modes of respiration that were previously linked to acid resistance,

for instance the complexes (*nuoBEK*^{28,29}, *sdhCDAB*³⁴, *frdB*³³, *cyoEDCBA*³⁵ and *atpAHF*²⁸) involved in aerobic respiration and oxidative phosphorylation and part of the TCA cycle (*aceEF* and *lpd* complex). This mainly because respiration and oxidative phosphorylation interferes with intracellular H⁺ concentrations³⁵. Interestingly, our approach also inferred a role for nitrate (*narG*, *narH*, *narI* and *narJ*) and nitrite (*nirB* and *nirD*) dependent respiration in acid resistance. Besides through respiration, their role in acid resistance also be mediated through their involvement in ammonium metabolism, a process known to be associated with pH alterations²⁵. In addition, ammonia production is a known pH regulation mechanism in *Helicobacter pylori*³⁶.

Finally the subgroup containing genes associated with flagellae and motility could be further subdivided into genes involved in flagellar assembly, flagellar regulation and motility and sensing. Genes involved in flagellar assembly were largely present in the differential expression benchmark set, but their regulatory genes and genes involved in motility were absent from both benchmarks, indicating that they have not explicitly been linked to acid resistance before.

Besides recovering genes already known to be involved in the process of interest (nodes corresponding to genes present in the benchmark sets are indicated in yellow), we also recovered several novel genes, potentially associated to acid resistance (grey nodes). When focusing on regulators (see Table 1), it is remarkable that PheNetic also prioritises OmpR, one of the regulators that was also by Stincone et al.¹⁹ related to acid resistance using a different inference technique. In addition, PheNetic strongly prioritizes PepA, an aminopeptidase and transcriptional repressor. Its homolog in *Vibrio cholerae* has previously been associated with acid resistance³⁷. Its targets, *carA* and *carB* which are also selected by PheNetic are involved in the conversion of glutamine to glutamate, which is one of the major known acid resistance mechanisms in *E. coli*³⁰. This gene encoding pepA seems barely altered at its expression level, explaining why it might have been missed by previous studies. Another strongly prioritized gene/gene product is TyrR, the main regulator of tyrosine synthesis, which has previously been associated with acid conditions in *Salmonella typhimurium*³⁸. TyrR regulates the amino acid metabolism regulator Mtr, which in its turn regulates the tryptophan or indole metabolism operon of which many genes were retrieved in the sub-network selected by PheNetic. The indole biosynthesis operon was found to be down-regulated in many of the KO strains we analysed, but none of its known regulators ranked well by the differential expression-based ranking method, explaining why it has been largely overlooked in the past. So far tryptophan biosynthesis has been only associated with acid resistance through the tryptophanases TnaA and TnaC^{29,28,26}.

Discussion

In this work, we developed an analysis method that allows interpreting in-house generated omics data in the light of publicly available information, represented as an interaction network. The developed method extracts sub-networks describing the mechanism behind the omics data from this interaction network.

The method was applied to reanalyse a previously published expression study, assessing gene expression of 27 KO strains under mild acid growth conditions in *E. coli*. To this end, an *E. coli* interaction network was compiled, spanning multiple layers of interactions. Applying PheNetic on this KO expression dataset using this *E. coli* interaction network, allowed recovering mechanisms known to be involved in acid resistance.

According to the classification of network inference methods of De Smet et al.³⁹, PheNetic can be considered as an integrative inference scheme, that uses next to expression data also omics derived network information to prioritize genes involved in the process of interest. Comparing PheNetic with classical differential expression-based ranking illustrates the added value of using such an integrative network-based approaches to analyse omics derived gene lists. This integrative inference strategy³⁹ allowed reaching higher sensitivities at lower FPR: false positive genes (e.g. genes that were erroneously found differentially expressed under the tested conditions) are filtered out more easily as these genes cannot be connected to genes related to the process of interest through the network. In addition, major players involved in the process of interest that are only weakly or not differentially expressed can be ranked higher using a network-based method than with an expression-based ranking or an expression-based inference method, if they are located in the network neighbourhood of genes that are differentially expressed. The latter property of a network-based approach such as PheNetic results in an improved prioritization of transcriptional regulators and signalling proteins related to acid resistance, as this class of proteins tends to often be less differentially expressed than structural genes. As such, we could predict novel links between regulators and acid resistance, that were not discovered in previous studies as most of these studies relied on a non-integrative inference strategies^{26,29,28,19}. An additional consequence and advantage of using the interaction network is that PheNetic's performance is robust against the number of cause-effect pairs used as input. This robustness makes the method less sensitive to the choice of an arbitrary cut-off on the number of differentially expressed genes selected per KO strain.

In addition, by pooling cause-effect pairs of multiple KO strains, assumed to display the same trait, PheNetic is able to reliably extract sub-networks that are confirmed by multiple KO strains. PheNetic conserves the relation between each cause and effect in contrast to eResponseNet. This implies that PheNetic has a clear advantage over eResponseNet in searching for the shared molecular mechanism. PheNetic will not only be able to recover mechanisms that are similar in multiple KO strains, but it will also be able to recover mechanisms specific to a single or limited number of KO strains as it does not pool all causes and effect as is done by eResponseNet.

Conclusion

In this study we developed PheNetic a network-based approach that allows retrieving from a comprehensive interaction network, the processes active in an omics-derived gene list. Our method can be generically applied to any model organism for which an interaction network is available, provided its parameters are tuned properly. Applying PheNetic on a real case study in *E. coli*. showed how overlaying expression-based gene lists with a network compiled from publicly available data not only recapitulates genes known to be involved in the process of interest, but also could uncover novel lead genes.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

This work is supported by: 1) Katholieke Universiteit Leuven funding: GOA/08/011, PF/10/010 (NATAR), CREA/08/023 2) Agentschap voor Innovatie door Wetenschap en Technologie (IWT): SBO-BioFrame, 3) Fonds Wetenschappelijk Onderzoek-Vlaanderen (FWO) G.0329.09; 4) Ghent University [Multidisciplinary Research Partnership “M2N”]

Material and Methods

Interaction network

The *E. coli* interaction network was built using an approach similar to those recently published for the construction of yeast interaction networks^{9,17,11} and reviewed in Cloots et al.⁴. For every molecular interaction layer the different interactions were retrieved (protein-protein, metabolic and transcriptional layer) and subsequently merged into an integrated network.

The interactions between molecular entities are represented as a graph $G(N, E)$ in which the nodes N are an abstract representation of both the gene and its gene product, depending on the interaction type. The edges E represent the interactions between the nodes. More specifically, when referring to protein-protein interactions, nodes will correspond to proteins and the interactions are considered to be undirected. For protein-DNA interactions, nodes will either correspond to transcription factors (TFs) or to their targets. Edges are by definition directed and point from TFs to targets. Metabolic interactions are represented as edges between nodes, corresponding to enzymes that act in a consecutive order (that is the product of the first enzyme is the substrate of the next). Metabolic interactions that are irreversible will be considered to be directed, whereas reversible metabolic interactions will be treated as undirected. Note that the definition of directed versus undirected interactions is determined by how the algorithm deals with the interactions, rather than based on a biological concept.

Concretely, for the transcriptional interaction layer 6111 protein-DNA interactions were obtained from RegulonDB⁴⁰ and an additional 295 predicted interactions were obtained from DISTILLER⁴¹, resulting in a total of 6406 protein-DNA interactions. Interactions from RegulonDB were assigned a probability of 0.8, whereas interactions predicted by DISTILLER were assigned a probability of 0.5 (as DISTILLER is an integrative scheme that does not assign a score to individual interactions). Related to the protein-protein interaction layer, we obtained a set of 7613 highly reliable protein-protein interactions from Bacteriome.org^{42,43}. These interactions were obtained by integrating seven computational and three experimental datasets using a Bayesian integration scheme and merged with the high-quality TAP results of Hu et al.⁴⁴. As edge weights we adopted the Bayesian integration scores for the individual edges described in the original publication⁴³. 2775 metabolic interactions were obtained from Notebaart et al.⁴⁵. As these are highly reliable interactions, they were assigned a probability of 0.8. This resulted in an integrated *E. coli* interaction network consisting of 3063 nodes connected by 16794 interactions. In addition, each node in the network is annotated with a probability that reflects its network centrality. This probability was derived from the out-degree distribution of the nodes in the network⁴⁶.

Random networks

We model the interaction network as a random network. This is a network (N, E) with N the set of nodes and with $E \subseteq N \times N$ a set of edges, together with weight functions $w_n: N \rightarrow [0,1]$ and $w_e: E \rightarrow [0,1]$ which assign a probability to each of the nodes and edges. A random network defines a probability distribution over possible sub-networks. Each possible sub-network can be represented as a collection of nodes $N' \subseteq N$ and edges $E' \subseteq E$ that denote the edges and nodes that are present. The probability of a possible sub-network is defined as: $p(N', E') =$

$\prod_{n \in N'} w_n(n) \prod_{n \in N \setminus N'} (1 - w_n(n)) \prod_{e \in E'} w_e(e) \prod_{e \in E \setminus E'} (1 - w_e(e))$. Observe that the resulting sub-networks may contain edges $e = (x, y)$ for which x and/or y does not belong to N' . We will eliminate these improper networks when dealing with decision making.

The probability distribution on the random network is used to determine how strongly connected nodes x and y are. This is defined as: $p(\text{path}(x, y) | N, E)$, the probability that there exists a path between x and y in a randomly sampled network of (N, E) . This random sampling is performed using a function: $\delta_{\text{path}(x, y)}: (2^N, 2^E) \rightarrow \{0, 1\}$ which is equal to 0/1 when the path is 'not present'/'present' in the possible network and where 2^N and 2^E are the power sets of N respectively E . For a path to be present in a sub-network it is required that all nodes and edges are present in the sub-network. We can now define the probability of a path as: $p(\text{path}(x, y) | N, E) = \sum_{(N', E') \in (2^N, 2^E)} p(N', E') \delta_{\text{path}(x, y)}(N', E')$. These probabilities can be computed using ProbLog, a probabilistic Prolog¹⁸.

Decision theoretic sub-network selection

We model the sub-network selection problem as a decision theoretic problem. In this type of problem a set of possible decisions is given and the goal is to select the subset of decisions that maximizes the utility function. Each decision is concerned with the presence or absence of a node in the network. Furthermore, when node n is absent, all edges starting in node n $e(n, _)$ and ending in node n $e(_, n)$ are also absent. The decisions at the level of the nodes are directly connected to the edges these nodes are involved in. This is intended to exclude the improper random networks discussed earlier.

The selected sub-network needs to be sparse, and needs to connect the causes to the effects. This leads to the following optimization function where $D \subseteq N$ is a candidate solution and I is the set of cause-effect pairs used as input for the method.

$$S(D) = \max \left\{ \left(\sum_{(x, y) \in I} f_r * p(\text{path}(x, y) | D, E) \right) - |D| * x_c \right\}$$

The total score of a selected sub-network $S(D)$ consisting of D , the subset of nodes selected from N , equals the sum of the reward of each explained cause-effect pair minus a cost term which imposes network sparseness. The reward (positive score) of explaining a single cause-effect pair depends on the probability that a regulatory path connecting cause (x) to effect (y) exists in the selected sub-network which is defined by $p(\text{path}(x, y) | D, E)$ and the degree to which the effect is differentially expressed as described by the reward function $f_r = \text{abs}(A_{\text{difex}}(x, y))^n$ (with A_{difex} being the degree to which the explained effect (y) is differentially expressed. A regulatory path requires the last edge to represent a gene regulatory interaction, thus imposing that the path connecting the cause to the effect should be

able to alter the expression of the effect. The cost term (negative score) is given by multiplying x_c , the gene selection cost, with the number of selected nodes in the sub-network.

Maximizing the optimization function was solved using DTProbLog¹⁸ which is a decision theoretic extension of the logic programming language, Prolog⁴⁷.

Algorithmic settings

As in previous studies^{14,11,48} the maximum length of the path between a cause and effect gene is restricted to 4 due to computational cost. The estimation of probabilities was performed using ProbLog's 5-best inference technique⁴⁹ which uses the 5-best proofs to approximate a probability. In practice this means that only the 5 most probable paths per cause-effect pair are used in the optimization function. To discourage the selection of well-connected or hub nodes the node centrality probability was capped at a minimum of 0.2. By doing so, hub nodes obtain a smaller reward when selected by the algorithm, but their selection is still possible if enough paths between cause-effect pairs pass through the hub gene.

Experimental datasets

Input data was derived from publicly available microarray experiments testing the transcriptional response of 27 different KO strains related to acid resistance in *Escherichia coli*¹⁹. Lists of differentially expressed genes were obtained from COLOMBOS⁵⁰. The data is available under experiment id GSE 13361 at <http://bioi.biw.kuleuven.be/colombos/> and comprises a total of 61 contrasts assessing gene expression for a total of 27 KO strains versus the wild type under two conditions, pH 5.5 and 7.

Input for PheNetic was generated as follows: for each of the 27 KO strains we derived all cause effect-pairs with the cause referring to the mutated gene in the KO strain and the effect to a gene differentially expressed in the KO versus the WT strain. Per KO strain we ranked the cause-effect pairs according to the absolute value of the expression fold change of the effect gene. Subsequently, we selected per KO strain the C highest ranked cause-effect pairs and pooled them into a merged list of 27 X C cause effect pairs.

Comparison with other methods

eResponseNet was obtained from <http://hanlab.genetics.ac.cn/eResponseNet/> and executed using standard settings^{51,9,20}. The *E. coli* interaction network was converted to the eResponseNet network format using the reliability probabilities as weights on the edges. The reliability probabilities were capped at a maximum of 0.8 to prevent selection of solely highly probable paths as described in Yeager-Lotem et al., 2009⁹. Prior to benchmarking, parameters were optimized as follows: different runs were performed using the 10, 25 and 50 most differentially expressed genes per KO strain as effects. The eResponseNet approach defines a parameter gamma that determines the number of nodes in the selected sub-network (referred to as network size). Increasing gamma results in the selection of larger sub-networks. eResponseNet was run for gamma parameters varying between 2 and 11 with a step size of 0.05. This sweep allows ranking the genes based on the gamma value at which the gene was initially selected as described in Huang et al., 2011²⁰.

The differential expression–based ranking was obtained by ranking the differential expressed genes for each KO strain based on their differential expression values (absolute log ratio). The obtained lists were combined into a single merged list in which all original ranks were maintained (as a result several genes obtained the same rank).

Benchmark dataset and performance estimation

To assess the performance we compiled two benchmark datasets: a first benchmark set, referred to as the literature benchmark set was compiled from genes associated with acid resistance in literature (see supplementary file SF1) and contained 53 genes obtained from several studies^{52,53,31,23,54,25,21,30,27,55}. A second benchmark set, referred to as the differential expression benchmark set, consists of 349 genes (see supplementary file SF2) identified to have an altered expression profile in acid conditions derived from different studies. Only genes identified to be significantly differential expressed in at least two of the three experiments were included. The literature and differential expression benchmark sets share an overlap of 16 genes.

To compare the performance of the different algorithm results, the False Positive Rate (FPR) and sensitivity of the result for the benchmark sets were determined. The sensitivity represents the number of genes from the benchmark set that have been identified in the subnetwork selected by the tested algorithm and the FPR represents the number false positives amongst the total number of genes present in the selected sub-network. Every gene not present in the benchmark set was considered false positive.

All three tested algorithms (PheNetic, eResponseNet and the differential expression-based ranking) allow varying the size of the selected sub-networks. Genes selected into the sub-network using a more stringent parameter setting are considered more significantly related to the process of interest. This allows us ranking genes based on the most stringent parameter at which genes are selected into the sub-network. For both the network methods, a sweep over respectively the gene selection cost for PheNetic and the gamma for eResponseNet allow obtaining a ranked gene list which can be compared to the differential expression-based ranking results.

Parameter settings of eResponseNet were tuned on the literature benchmark set. The best performance was found using the 10 most differentially expressed genes (effects) per mutated gene (cause) as input (Figure S3). eResponseNet was originally designed to work with a network consisting of undirected protein-protein interactions and directed DNA-protein interactions. We observed on the literature benchmark that adding the metabolic layer to the interaction network consisting of protein-protein and protein-DNA interactions, either as a combination of directed and undirected edges or as a set of exclusively undirected edges deteriorated the results of eResponseNet (see Supplementary figure S4). This is due to the implementation of eResponseNet: the algorithm searches for regulatory paths in the network. These regulatory paths connect the cause to the effect in the interaction network. eResponseNet requires the last edge of these regulatory path to be a directed one, assuming

this directed edge represent a protein-DNA interaction in the network. By adding directed metabolic edges this assumption is violated. Because it resulted in the best performances, we run eResponseNet on a reduced network without metabolic interactions.

Enrichment analysis

GO enrichment was performed using the BiNGO Cytoscape plugin from ⁵⁶ using a hypergeometric test with a Benjamini and Hochberg False Discovery Rate correction. A p-value cut-off of 0.05 was used to identify enriched processes. Additionally ClueGO ⁵⁷ was used to group and analyse the GO and KEGG enrichments (results not shown).

Network visualization

Networks were visualized and analysed using Cytoscape⁵⁸.

Figure 1 – Schematic overview of the experimental setup. KO genes are referred to as causes and differentially expressed genes as effects (left panel). PheNetic connects causes to effects over an interaction network derived from publicly available data (middle panel) by searching for paths in the interaction network that connect causes (blue dots) to as many possible effects (red and green nodes) in the most parsimonious way (right panel). PheNetic allows extracting from the global interaction network, the molecular mechanism that is activated by the KO experiment.

Figure 2 – The effect of using different numbers of cause-effect pairs per KO strain as input on the performance of PheNetic. Performance comparison was based on a sensitivity–FPR analysis as described in Materials and Methods using the literature (top panel) and the differential expression (bottom panel) benchmark sets. Results are shown for using as input respectively 10 (green), 100 (orange) and 1000 (blue) cause-effect pairs per KO strain. The results were obtained using the exponential reward function (see Material and Methods) and a parameter sweep over the gene selection cost.

Figure 3 – Performance comparison of sub-network selection methods on the acid resistance dataset. The performance of PheNetic (blue), differential expression-based ranking (orange) and eResponseNet (green) was compared using two benchmark sets, one based on literature (top panel) and based on differential expressed genes (bottom panel). The performance is assessed by plotting the sensitivity (the number of benchmark genes in the selected sub-network versus the total number of benchmark genes) versus the false positive rate (FPR, defined as the number of positive interactions amongst the total number of predicted

interactions) for selected sub-networks of increasing size (obtained by a parameter sweep over the gene selection cost).

Figure 4 – Detailed view on the sub-network involved in acid resistance identified by PheNetic. The sub-network was decomposed into different subgroups centred around the overrepresented GO categories. For visualization purposes, genes are grouped based on their annotation in KEGG and GO (shaded areas). Genes contained in both benchmark sets are indicated as yellow nodes.

Table 1 – Ranks assigned to a selection of regulators and signal transducers involved in acid resistance. For an exhaustive ranking of all regulators and genes by the different methods see supplementary material.

For each of the indicated regulators (rows), its rank as assigned by each of the respective algorithms (columns) is shown. The rank is defined as the number of other genes that were selected in the sub-network, prior to selecting the indicated gene. NF (not found) indicates that the indicated gene was not ranked by the method. 'Acid resistance regulators' correspond to those regulators known to be involved in acid resistance (literature benchmark) and not used as cause in the input. The newly associated regulators are regulators that to our knowledge, have not yet been associated with acid resistance in *E. coli*.

Gene name	Differential expression-based ranking	PheNetic	eResponseNet
Acid resistance regulators			
YdeO	2649	393	372
TorR	1808	393	NF
RcsB	1800	185	306
RpoD	1213	99	115
EvgA	557	463	NF
EvgS	1302	855	NF
OmpR	1088	99	277
Newly associated regulators			
PepA	1045	149	57
TyrR	2060	149	NF

References

1. V. K. Ramanan, L. Shen, J. H. Moore, and A. J. Saykin, *Trends in genetics* □: *TIG*, 2012, **28**, 323–32.
2. P. Khatri and S. Drăghici, *Bioinformatics (Oxford, England)*, 2005, **21**, 3587–95.
3. J. a Bernstein, P.-H. Lin, S. N. Cohen, and S. Lin-Chao, *Proceedings of the National Academy of Sciences of the United States of America*, 2004, **101**, 2758–63.
4. L. Cloots and K. Marchal, *Current opinion in microbiology*, 2011, **14**, 599–607.
5. L. Cloots, D. De Maeyer, and K. Marchal, in *Handbook of Bio- and Neuroinformatics*, ed. C. Brown, Springer Verlag, 2012.
6. A. Sánchez-Rodríguez, L. Cloots, and K. Marchal, *Current Bioinformatics*, (in press).
7. D. Nitsch, J. P. Gonçalves, F. Ojeda, B. de Moor, and Y. Moreau, *BMC bioinformatics*, 2010, **11**, 460.
8. L. P. C. Verbeke, L. Cloots, P. Demeester, J. Fostier, and K. Marchal, *Bioinformatics*, 2012.
9. E. Yeger-Lotem, L. Riva, L. J. Su, A. D. Gitler, A. G. Cashikar, O. D. King, P. K. Auluck, M. L. Geddie, J. S. Valastyan, D. R. Karger, S. Lindquist, and E. Fraenkel, *Nature genetics*, 2009, **41**, 316–23.
10. N. Novershtern, A. Regev, and N. Friedman, *Bioinformatics (Oxford, England)*, 2011, **27**, i177–85.
11. O. Ourfali, T. Shlomi, T. Ideker, E. Ruppín, and R. Sharan, *Bioinformatics (Oxford, England)*, 2007, **23**, i359–66.
12. L. García-Alonso, R. Alonso, E. Vidal, A. Amadoz, A. de María, P. Minguez, I. Medina, and J. Dopazo, *Nucleic acids research*, 2012, 1–13.
13. M. Bailly-Bechet, C. Borgs, A. Braunstein, J. Chayes, A. Dagkessamanskaia, J.-M. François, and R. Zecchina, *Proceedings of the National Academy of Sciences of the United States of America*, 2010, **108**, 882–887.
14. C.-H. Yeang, T. Ideker, and T. Jaakkola, *Journal of computational biology* □: *a journal of computational molecular cell biology*, 2004, **11**, 243–62.

15. Z. Tu, L. Wang, M. N. Arbeitman, T. Chen, and F. Sun, *Bioinformatics (Oxford, England)*, 2006, **22**, e489–96.
16. P. Jia, L. Wang, A. H. Fanous, C. N. Pato, T. L. Edwards, and Z. Zhao, *PLoS computational biology*, 2012, **8**, e1002587.
17. S. Suthram, A. Beyer, R. M. Karp, Y. Eldar, and T. Ideker, *Molecular systems biology*, 2008, **4**, 162.
18. G. Van Den Broeck, I. Thon, M. Van Otterlo, and L. De Raedt, *24th AAAI Conference on Artificial Intelligence*, 2010.
19. A. Stincone, N. Daudi, A. S. Rahman, P. Antczak, I. Henderson, J. Cole, M. D. Johnson, P. Lund, and F. Falciani, *Nucleic acids research*, 2011, 1–17.
20. J. Huang, Y. Liu, W. Zhang, H. Yu, and J.-D. J. Han, *Bioinformatics (Oxford, England)*, 2011, **27**, 2319–20.
21. N. Masuda and G. M. Church, *Molecular microbiology*, 2003, **48**, 699–712.
22. C. Bordi, L. Théraulaz, V. Méjean, and C. Jourlin-Castelli, *Molecular microbiology*, 2003, **48**, 211–23.
23. M. P. Castanie-Cornet, T. a Penfound, D. Smith, J. F. Elliott, and J. W. Foster, *Journal of bacteriology*, 1999, **181**, 3525–35.
24. J. W. Foster, *Nature reviews. Microbiology*, 2004, **2**, 898–907.
25. L. M. Stancik, D. M. Stancik, B. Schmidt, D. M. Barnhart, Y. N. Yoncheva, and J. L. Slonczewski, *Journal of bacteriology*, 2002, **184**, 4246–58.
26. L. M. Maurer, E. Yohannes, S. S. Bondurant, M. Radmacher, and J. L. Slonczewski, *Journal of bacteriology*, 2005, **187**, 304–19.
27. Y. Sun, T. Fukamachi, H. Saito, and H. Kobayashi, *Journal of bacteriology*, 2011, **193**, 3072–7.
28. E. T. Hayes, J. C. Wilks, P. Sanfilippo, E. Yohannes, D. P. Tate, B. D. Jones, M. D. Radmacher, S. S. BonDurant, and J. L. Slonczewski, *BMC microbiology*, 2006, **6**, 89.
29. G. Kannan, J. C. Wilks, D. M. Fitzgerald, B. D. Jones, S. S. Bondurant, and J. L. Slonczewski, *BMC microbiology*, 2008, **8**, 37.
30. H. Richard and J. W. Foster, *Journal of bacteriology*, 2004, **186**, 6032–41.

31. B. M. Hersh, F. T. Farooq, D. N. Barstad, D. L. Blankenhorn, and J. L. Slonczewski, *Journal of bacteriology*, 1996, **178**, 3978–81.
32. R. Iyer, C. Williams, and C. Miller, *Journal of Bacteriology*, 2003, **185**, 6556–6561.
33. E. R. Olson, *Molecular microbiology*, 1993, **8**, 5–14.
34. A. G. Oglesby, E. R. Murphy, V. R. Iyer, and S. M. Payne, *Molecular microbiology*, 2005, **58**, 1354–67.
35. P. A. Cotter, V. Chepuri, R. B. Gennis, and R. P. Gunsalus, *Journal of bacteriology*, 1990, **172**, 6333–8.
36. M. Pflock, S. Kennard, N. Finsterer, and D. Beier, *Journal of biotechnology*, 2006, **126**, 52–60.
37. J. Behari, L. Stagon, and S. B. Calderwood, *Journal of bacteriology*, 2001, **183**, 178–88.
38. K. R. Park, J. C. Giard, J. H. Eom, S. Bearson, and J. W. Foster, *Journal of bacteriology*, 1999, **181**, 689–94.
39. R. De Smet and K. Marchal, *Nature reviews. Microbiology*, 2010, **8**, 717–729.
40. S. Gama-Castro, H. Salgado, M. Peralta-Gil, A. Santos-Zavaleta, L. Muñoz-Rascado, H. Solano-Lira, V. Jimenez-Jacinto, V. Weiss, J. S. García-Sotelo, A. López-Fuentes, L. Porrón-Sotelo, S. Alquicira-Hernández, A. Medina-Rivera, I. Martínez-Flores, K. Alquicira-Hernández, R. Martínez-Adame, C. Bonavides-Martínez, J. Miranda-Ríos, A. M. Huerta, A. Mendoza-Vargas, L. Collado-Torres, B. Taboada, L. Vega-Alvarado, M. Olvera, L. Olvera, R. Grande, E. Morett, and J. Collado-Vides, *Nucleic Acids Research*, 2011, **39**, D98–D105.
41. K. Lemmens, T. De Bie, T. Dhollander, S. C. De Keersmaecker, I. M. Thijs, G. Schoofs, A. De Weerd, B. De Moor, J. Vanderleyden, J. Collado-Vides, K. Engelen, and K. Marchal, *Genome biology*, 2009, **10**, R27.
42. C. Su, J. M. Peregrin-Alvarez, G. Butland, S. Phanse, V. Fong, A. Emili, and J. Parkinson, *Nucleic Acids Research*, 2008, **36**, D632–D636.
43. J. M. Peregrín-Alvarez, X. Xiong, C. Su, and J. Parkinson, *PLoS computational biology*, 2009, **5**, e1000523.
44. P. Hu, S. C. Janga, M. Babu, J. J. Díaz-Mejía, G. Butland, W. Yang, O. Pogoutse, X. Guo, S. Phanse, P. Wong, S. Chandran, C. Christopoulos, A. Nazarians-Armavil, N. K. Nasser, G. Musso, M. Ali, N. Nazemof, V. Eroukova, A. Golshani, A. Paccanaro, J. F. Greenblatt, G. Moreno-Hagelsieb, and A. Emili, *PLoS biology*, 2009, **7**, e96.

45. R. a Notebaart, B. Teusink, R. J. Siezen, and B. Papp, *PLoS computational biology*, 2008, **4**, e26.
46. A.-L. Barabási and Z. N. Oltvai, *Nature reviews. Genetics*, 2004, **5**, 101–13.
47. I. Bratko, *Programming in Prolog for Artificial Intelligence*, Pearson Education, 4th editio., 2011.
48. A. Joshi, T. Van Parys, Y. Van De Peer, and T. Michoel, *Genome biology*, 2010, **11**, R32.
49. A. Kimmig, B. Demoen, L. De Raedt, V. S. Costa, and R. Rocha, *Theory and Practice of Logic Programming*, 2011, **11**, 235–262.
50. K. Engelen, Q. Fu, P. Meysman, A. Sánchez-Rodríguez, R. De Smet, K. Lemmens, A. C. Fierro, and K. Marchal, *PLoS ONE*, 2011, **6**, e20938.
51. A. Lan, I. Y. Smoly, G. Rapaport, S. Lindquist, E. Fraenkel, and E. Yeger-Lotem, *Nucleic acids research*, 2011, **39**, W424–9.
52. P. Small, D. Blankenhorn, D. Welty, E. Zinser, and J. L. Slonczewski, *Journal of bacteriology*, 1994, **176**, 1729–37.
53. C. Kirkpatrick, L. M. Maurer, N. E. Oyelakin, Y. N. Yoncheva, R. Maurer, and J. L. Slonczewski, *Journal of bacteriology*, 2001, **183**, 6466–77.
54. I. R. Boot, P. Cash, and C. O’Byrne, *Antonie van Leeuwenhoek*, 2002, **81**, 33–42.
55. M. D. Johnson, N. a Burton, B. Gutiérrez, K. Painter, and P. a Lund, *Journal of bacteriology*, 2011, **193**, 3653–6.
56. S. Maere, K. Heymans, and M. Kuiper, *Bioinformatics (Oxford, England)*, 2005, **21**, 3448–9.
57. G. Bindea, B. Mlecnik, H. Hackl, P. Charoentong, M. Tosolini, A. Kirilovsky, W.-H. Fridman, F. Pagès, Z. Trajanoski, and J. Galon, *Bioinformatics (Oxford, England)*, 2009, **25**, 1091–3.
58. M. E. Smoot, K. Ono, J. Ruscheinski, P.-L. Wang, and T. Ideker, *Bioinformatics*, 2011, **27**, 431–432.