

Biological Network Integration and Mining for Microbial Community Analysis

Curtis Huttenhower
Department of Biostatistics
Harvard School of Public Health
Boston, MA 02115
chuttenh@hsph.harvard.edu

Abstract

A variety of genome-scale functional data is available for many microbial species, but determining the biological functionality and metabolic potential of a sequenced organism remains a significant challenge. Biological network integration and mining algorithms provide a means of assembling this body of data, understanding it from a systems level, and applying it to the study of uncharacterized species and communities. We compare supervised and unsupervised Bayesian approaches to biological network integration; this process provides maps of functional activity and genomewide interactomes in over 100 areas of cellular biology, using information from ~5,000 genome-scale experiments pertaining to 13 microbial species. In combination with graph alignment, these network manipulation tools provide a means for analyzing the functional activity unique to particular pathogens, transferring putative functional annotations to uncharacterized organisms, and potentially inferring interactomes using weighted network integration for metagenomic communities.

1 Introduction

High-throughput sequencing has greatly reduced the cost and difficulty of obtaining microbial genomes, but determining the biological functionality and metabolic potential of a sequenced organism remains a significant challenge. This challenge is magnified in metagenomic and metatranscriptomic communities, which might sample only small fragments of genes from thousands of organisms. Existing computational tools for functionally annotating a newly sequenced genome or community rely heavily on sequence homology, are limited in throughput by requirements for manual curation and prior knowledge, and do not yet leverage the thousands of experimental results publicly available for a diversity of characterized organisms.

Large scale functional genomic data integration has succeeded in predicting both functional annotations and interactomes in organisms ranging from yeast to human [1-4]. The strengths of such integrative methods include the ability to make data-driven predictions based on tens of thousands of experimental results, to operate in the space of biological networks (e.g. physical, genetic, regulatory, or functional interactomes) rather than genomic sequences, and to employ scalable machine learning to overcome sparse prior knowledge. Here, we present computational methodology for biological network integration, allowing A) the rapid, scalable integration of arbitrary experimental data modeled as biological networks and B) a system of functional mapping to derive high-level pathway activity and associations from genome-scale data.

In order to take advantage of large collections of genomic data, they must be integrated, summarized, and presented in a biologically informative manner. We provide a means of mining thousands of whole-genome experimental interactomes by way of functional maps. Each map represents a body of data, probabilistically weighted and integrated, focused on a particular biological question. These questions can include, for example, the function of a gene, the relationship between two pathways, or the overall metabolic or functional activity present in a dataset or genome. Each functional map, based on an underlying predicted interaction network, summarizes an entire collection of genomic experimental results in a biologically meaningful way.

While functional maps can readily predict functions for uncharacterized genes [5], it is important to take advantage of the scale of available data to understand entire pathways and processes. Cross-talk and co-regulation among pathways, processes, and metabolic functions can be mapped by analyzing the structure of underlying functional relationship networks. Similarly, associations between distinct but interacting biological processes (e.g. mitosis and DNA replication) can be quantified by examining functional relationships between groups of genes, allowing the identification of proteins key to interprocess regulation. We demonstrate this functional mapping methodology using a compendium of 4,894 bacterial expression conditions spanning 13 species (see Table 1) and discuss its future application to the characterization of newly sequenced microbes and metagenomes.

Table 1: Organisms and data compendia analyzed in this study

Organism	Dsets.	Conds.	Organism	Dsets.	Conds.
<i>Bacillus anthracis</i>	1	8	<i>Helicobacter pylori</i>	26	809
<i>Bordetella bronchiseptica</i>	9	63	<i>Mycobacterium tuberculosis</i>	31	641
<i>Campylobacter jejuni</i>	352	681	<i>Pseudomonas aeruginosa</i>	34	324
<i>Clostridium botulinum</i>	1	11	<i>Staphylococcus aureus</i>	151	828
<i>Enterococcus faecalis</i>	3	31	<i>Vibrio cholerae</i>	14	275
<i>Escherichia coli</i>	81	1061	<i>Yersinia enterocolitica</i>	11	104
<i>Francisella tularensis</i>	8	58			

2 Results

2.1 Scalable biological network integration accurately predicts microbial interactomes

For each of the 13 organisms discussed above, we generated two predicted functional interactomes integrating all available experimental datasets. A supervised interactome was predicted using Bayesian data integration, in which each dataset's discretized coexpression values were used to train a naive classifier using curated relationships from the KEGG catalog [6]. Additionally, an unsupervised interactome was predicted by averaging over all datasets' normalized coexpression networks; this is detailed in the following section.

As evaluated in Figure 1A, the ability of the resulting supervised interaction networks to accurately recapitulate KEGG coannotations correlates approximately with the number of available expression conditions. In some outliers, the number of effective datasets is limited due to losses during automated extraction from ArrayExpress (e.g. only ~15 of the *S. aureus* datasets contribute usefully to the integrated network). Similarly, in species with limited expression data, the majority of mispredicted interactions are false negatives with no available data to drive predictions (e.g. some ~8,000 gene pairs in *B. anthracis*). However, in species with sufficient data, many high-confidence false positives can easily be characterized as underannotations in KEGG; the top *E. coli* predictions, for example, include a variety of interactions between RNA polymerase, elongation factors, protein translocases, and the ribosome that represent true biological interactions not captured by KEGG pathways. Network analysis can easily extract these dense subgraphs within each species for future curation [4].

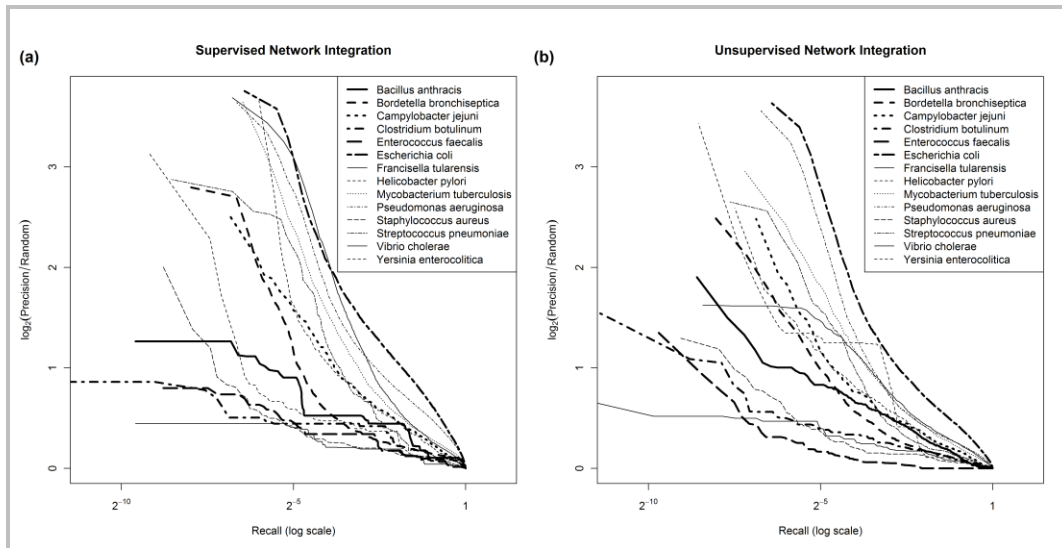


Figure 1: Accuracy of predicted microbial functional interactomes as evaluated against the KEGG [6] catalog. A) Supervised Bayesian integration of experimental datasets. B) Unsupervised biological network integration.

2.2 Unsupervised integration produces functional interactomes in the absence of a curated gold standard

Completely unsupervised biological network integration also predicts most species' functional interactomes with remarkable accuracy, as shown in Figure 1B. As discussed below, unsupervised network integration introduces much greater variability into the predicted interactome, but the rank order of confident predictions is largely preserved. This accuracy is robust to removal of large confounding biological pathways such as translation and the ribosome ([7], data not shown). AUCs decrease by an average of 0.017 in the unsupervised evaluations, the major benefit to supervised network integration arising in cases where noisy datasets are effectively downweighted by the Bayesian classifier (e.g. several *H. pylori* datasets have individual AUCs below 0.5 due to data processing, microarray platform age, and low numbers of conditions).

While overall AUCs are consistently improved by supervised network integration, unsupervised network analysis provides several organisms with a boost in the low recall, high precision region of predictive biological interest. One such example is *Y. enterocolitica*, in which the three expression datasets usable after processing all provide individually accurate functional predictions (data not shown). Many of the predictions with the greatest magnitude of change between the two integration algorithms are uncharacterized, but several include transporters (e.g. *ysaV*, *ysaN*, *yst1M*, *yst1F*, *yst1C*, and others), metabolic enzymes, chemotactic proteins, and flagellar components with clear relationships overly downweighted by the supervised integration process. Organisms in which this is the case generally have too little experimental data for completely accurate supervised learning to take place; for example, no comparable examples are produced for *E. coli* or *P. aeruginosa*.

2.3 Functional mapping characterizes species-specific metabolic and functional potential

Functional mapping is a network mining tool that further summarizes compendia of genomewide interactomes (e.g. from many biological contexts [4] or, in this case, species) as a set of annotated process- and pathway-level associations. As detailed in Methods, functional mapping relies on the aggregate analysis of edges within or spanning gene sets of interest, which can be drawn from prior knowledge (e.g. KEGG pathways) or extracted using unsupervised clustering. While the functional activity scores shown here are based on score ratios for simplicity, an algorithm for deriving bootstrap p-values is discussed below. An

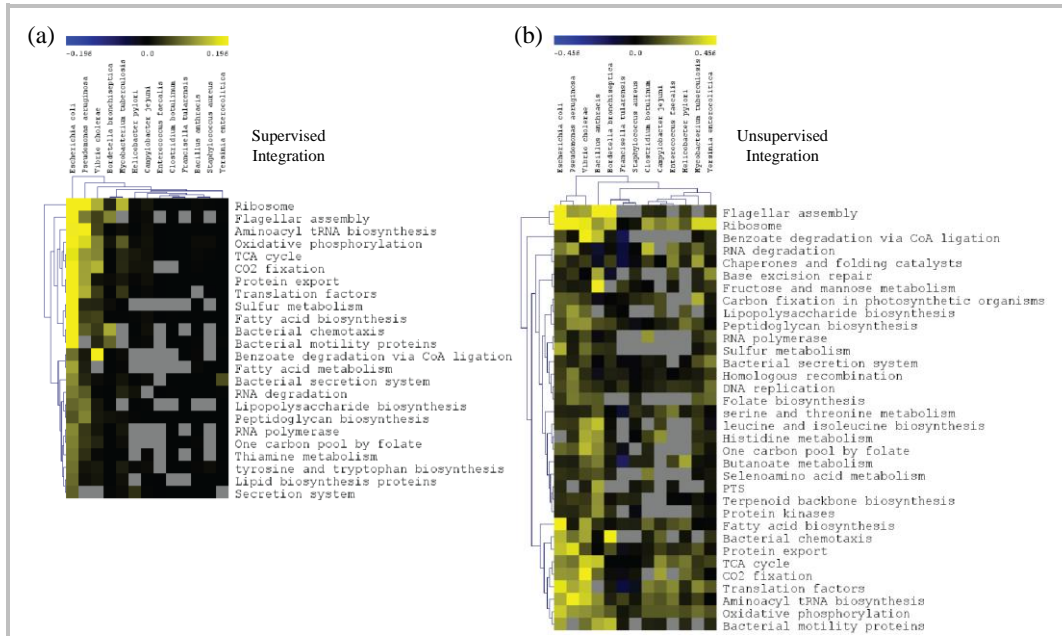


Figure 2: Functional activity predicted by the A) supervised and B) unsupervised network integrations. Cell color indicates the cohesiveness of each biological process within each species' network on a logarithmic scale; dynamic range is four standard deviations around mean, and processes with insufficient data or low detectable activity have been omitted.

example of functional mapping results characterizing biological pathway activity for these 13 microbial species within selected KEGG pathways is shown in Figure 2.

Functional maps derived from the supervised interactomes (Figure 2A) emphasize mainly easily detectable biological processes from well-characterized species. The ribosome and translational processes such as tRNA synthesis are known to have strong expression signals in microorganisms [7], and organisms well-annotated in the KEGG catalog have by far the strongest detectable activity. One outlier is the strong signal detected for anaerobic benzoate degradation in *V. cholerae*, driven by consistent activity in the *sdh* and *frd* operons across nearly all available data. While these are significant components of the tricarboxylic acid cycle and are strongly conserved in many microbes [8-9], there is no literature evidence to suggest that they might play a unique role in cholera, and this unusual activity could be followed up experimentally.

The unsupervised interactomes show a much more heterogeneous pattern of functional activity (Figure 2B), due mainly to the substantially higher variability introduced by giving equal weight to all datasets during the integration process. However, while this clearly introduces additional noise (note several pathways with below-baseline cohesiveness in some organisms, denoted by blue cells), it also emphasizes several areas of biological interest potentially hidden by the emphasis on characterized organisms in the supervised interactomes. These include a link between the TCA cycle/pentose phosphate metabolism and *V. cholerae*'s anaerobic growth (linked to carbon fixation by orthology to photosynthetic microbes), fructose metabolism in *B. anthracis* (a spore surface component [10]), and a collection of RNA helicases in *C. botulinum* (which are differentially active relative to *C. sporogenes* [11]). Unsupervised biological network integration thus provides a means of exposing accurate, novel biology from large functional data compendia, even in the absence of prior knowledge regarding species of interest.

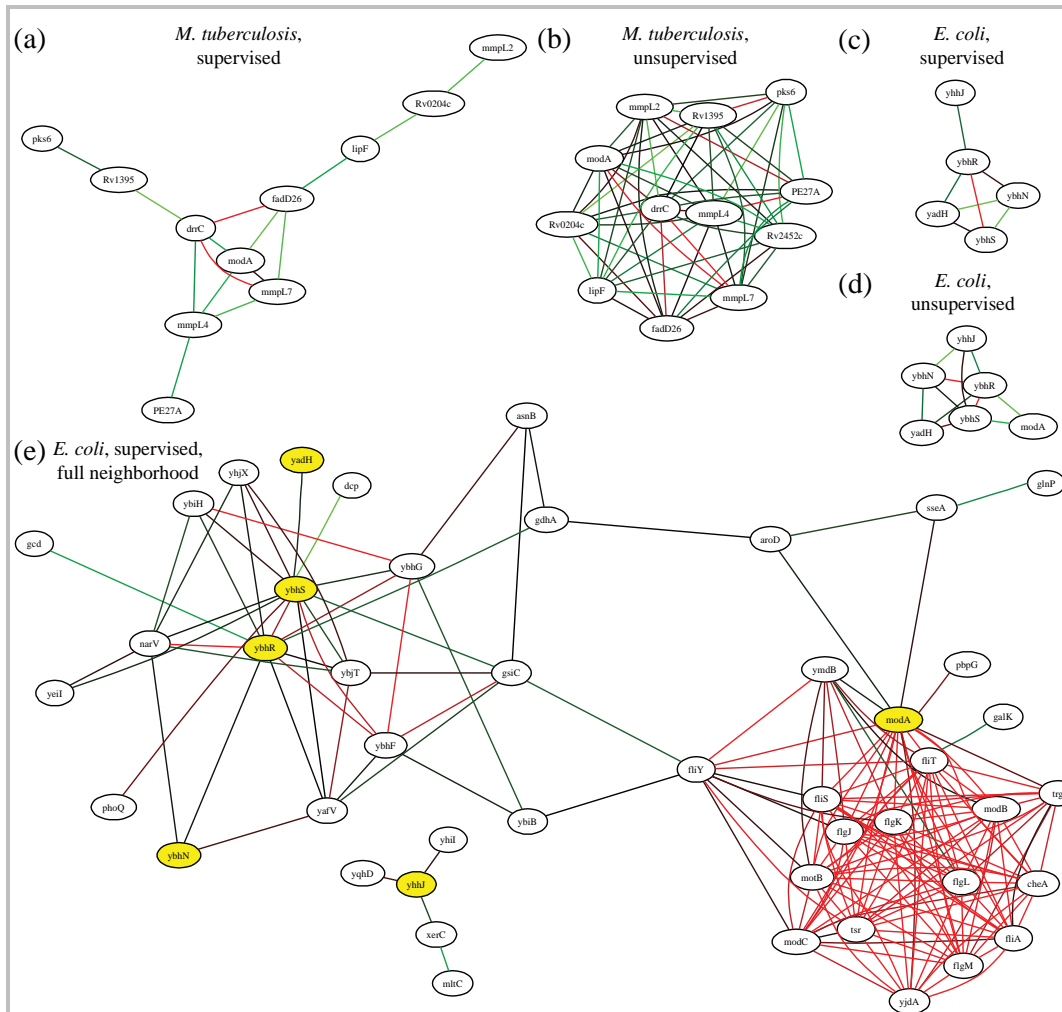


Figure 3: Example of the variation observed in predicted interactomes between species and network integration algorithms. In all subgraphs save E, a normalized edge weight threshold of 0.25 was used. A) The subgraph over 11 virulence-linked genes from [14] in the supervised *Mycobacterium tuberculosis* network. B) The same 11 genes in the unsupervised *M. tuberculosis* network. C) The subgraph over the six orthologs of these genes in the *Escherichia coli* supervised network (the sixth ortholog is omitted from this connected component). D) The subgraph over these six orthologs in the unsupervised *E. coli* network. E) The full network neighborhood surrounding these six orthologs in the supervised *E. coli* network as determined by the HEFAlMp network query algorithm [4].

2.4 Network comparisons highlight functional specialization

By aligning predicted interactomes using sequence-based orthology [12] or graph alignment [13], network-based functional information can be transferred between species in a richer context than by sequence similarity alone. For example, Figure 3 demonstrates the subgraphs surrounding 11 genes linked to virulence in *M. tuberculosis* [14] in a variety of interactome contexts. In the supervised tuberculosis interactome (Figure 3A), most data is downweighted, and these genes are only loosely functionally related. However, they do demonstrate strong coexpression in many datasets, as evidenced by their high connectivity in the unsupervised interactome (Figure 3B); this represents a case in which a clustering analysis could easily have uncovered this important biological feature based on network integration. KEGG provides only six known orthologs to these proteins in *E. coli*, but they are in turn tightly clustered in both *E. coli* interactomes (Figures 3C and D). Using the HEFAlMp graph search algorithm to visualize the entire network neighborhood around these

orthologs (Figure 3E) reveals two distinct clusters, the *ybh* and related operons (an uncharacterized ATP-binding cassette transporter) and the flagellar operons *flg* and *fli*. These are linked by a collection of amino acid transporters and modifiers, providing a rapid, computational derivation of the experimental hypothesis originally offered in [14] that the virulence cluster likely deals with small molecule, drug, and host metabolite transport.

3 Conclusions

Here, we present network integration and analysis tools allowing the supervised (i.e. based on prior biological knowledge) and unsupervised integrations of over 700 experimental datasets from 13 microbial species. This methodology is completely automated, relying on the extraction of expression data from repositories such as ArrayExpress, its conversion into normalized coexpression networks, and the integration of these networks into species-specific functional interactomes either by unsupervised averaging or by supervised Bayesian learning. Finally, biological activity in the resulting interactome compendium was further summarized using functional mapping, revealing significant pathway coregulation and interspecies variability.

The key methodologies driving this analysis are efficient large scale network alignment and subgraph comparisons. The former allows arbitrary experimental data to be modeled as biological networks - possibly with a large proportion of missing nodes (genes) or edges - and weighted either uniformly or using learned probability ratios in a Bayesian framework. The latter allows compendia of functional networks, which would otherwise be unwieldy for direct biological analysis, to be further summarized as association and cohesiveness measures between and within pathways and processes of interest. The combination of these features with sequence-based interspecies orthology or direct graph alignment algorithms provides an immediate means for biological hypothesis generation, for example regarding the factors driving virulence or host interactions in differentially pathogenic strains of a single species or the functionality of uncharacterized genes in newly sequenced organisms.

Finally, one of the most important areas for future applications of this work is in the analysis of metagenomic communities. As high-throughput sequencing is increasingly used to collect short DNA sequences directly from uncultured environmental samples, the need to functionally characterize community activity and individual microbial community members will grow dramatically [15]. By combining sequence similarity, graph alignment, local subgraph analysis, and large scale functional data integration, the tools presented here for weighted biological network integration can be used to transfer functional maps and partial interactomes from laboratory-based experimental results to environmental metagenomes and metatranscriptomes. When used in the analysis of the human microflora or of pathogen populations with variable genetic repertoires, this has the potential to provide rapid computational hypothesis generation for the characterization of microbial community roles in human disease.

4 Methods

We provide genomewide functional interactomes predicted for 13 bacterial species using efficient Bayesian integration of 722 genomic datasets modeled as whole-genome interaction networks [16]. Functional associations between biological processes from KEGG [6] were derived by further integration and analysis of these networks in a context-sensitive manner.

4.1 Data collection and gold standard generation

We integrated 722 expression datasets spanning 13 microbes drawn from the ArrayExpress database [17]. Each experimental result was modeled as an interaction network and initially processed as described in [16]. Each dataset D was converted from expression values to gene pair similarity scores using Pearson correlation normalized using Fisher's z-transform and subsequently z-scored:

$$fisher_D(g_i, g_j) = \frac{1}{2} \log \left(\frac{1 + \rho_D(g_i, g_j)}{1 - \rho_D(g_i, g_j)} \right)$$

$$w_D(g_i, g_j) = \frac{\text{fisher}_D(g_i, g_j) - \mu_D}{\sigma_D}$$

After z-scoring, each expression dataset was quantized using the binnings $(-\infty, -1.5)$, $[-1.5, -0.5)$, $[-0.5, 0.5)$, $[0.5, 1.5)$, $[1.5, 2.5)$, $[2.5, 3.5)$, $[3.5, \infty)$; these represent steps of one standard deviation in z-score space.

Unsupervised network integration was performed by averaging the resulting interactomes within each species S :

$$w_{US}(g_i, g_j) = \frac{1}{|\mathcal{D}_S|} \sum_{D \in \mathcal{D}_S} w_D(g_i, g_j)$$

To perform supervised network integration, we generate a gold standard of known functionally related and unrelated gene pairs. Biological processes of interest were selected from KEGG [6] and an answer set was derived from these processes as described in [16]. Gene pairs coannotated to any term were considered to be related. A gene pair was unrelated in the gold standard if A) the two genes were both annotated to some term in the positive term set, B) the genes were not coannotated to any of these terms, and C) the terms to which the genes were annotated did not overlap with hypergeometric p-value less than 0.05. All other gene pairs were omitted from the standard (i.e. they were neither related nor unrelated for training and evaluation purposes).

4.2 Bayesian analysis

One naive Bayesian classifier was learned per organism of interest; experiments with other network structures were shown to provide negligible performance improvements [16]. Briefly, a global classifier was learned in which the class to be predicted was gene pair functional relationships (as defined in the gold standard) and each dataset formed one node in the network. All Bayes network manipulation was performed using the Sleipnir C++ library for computational functional genomics [18]. Each naive Bayesian classifier directly implies a functional relationship network in which nodes represent genes and edge weights consist of the posterior probabilities of functional relationships between gene pairs. This results in a supervised integrated functional interactome predicted for each species S as:

$$w_{SS}(g_i, g_j) = P[\text{FR}_{i,j} | \mathcal{D}_S] \propto P[\text{FR} | \mathcal{D}_S] \prod_{D \in \mathcal{D}_S} P[w_D(g_i, g_j) | \text{FR}]$$

where $w_D(g_i, g_j)$ is discretized as described above.

4.3 Functional relationship and dataset enrichment predictions

As described above, for the purposes of this analysis, a biological process was defined as a set of related genes. The strength of a predicted functional relationship between two processes F and G was calculated as the average edge weight in the global interaction network within the edge set:

$$E_{F,G} = \{(g_i, g_j) | g_i \in F, g_j \in G, g_i, g_j \notin F \cap G\}$$

Similarly, the functional cohesiveness of a process was measured as the ratio of the average edge weight in the process to the average edge weight incident to the process:

$$\text{cohes}(F) = \frac{2|G| \sum_{g_i, g_j \in F} w_F(g_i, g_j)}{|F-1| \sum_{g_i \in F} \sum_{g_j \in G} w_F(g_i, g_j)}$$

where F is the function of interest, G is the genome, and $w_F(g_i, g_j)$ is the edge weight between genes g_i and g_j .

For the purpose of predicting gene function based on "guilt by association" with known genes in some process, the connectivity of a gene to a process was assessed as follows. Each gene/process pair was assigned a functional association score equal to the ratio of its average

probability of functional relationship to the process over the process's cohesiveness:

$$assoc(g_i, F) = \frac{\sum_{g_j \in F} w_F(g_i, g_j)}{|F| cohes(F)}$$

4.3 Functional mapping associations and p-values

As described in [4], these cohesiveness and association scores can also be converted from empirical ratios into p-values, although this methodology has been omitted from this analysis in the interest of brevity. Briefly, we can define a functional association score made up of four parts. The score between two gene sets within a process is the average probability of all edges *between* them, essentially their association. Their *background* score in a process is the average probability of all edges incident to either set. The *baseline* score is the average probability of an edge in the integrated network. The score *within* a single gene set is the average edge probability assuming nodes are self-connected with baseline strength, and the score within two gene sets is their unweighted average. The *between* and *baseline* scores are divided by the *background* and *within* scores to calculate two gene sets' functional association, which is thus increased if they are more interconnected and decreased if they are more self-connected. Thus for any two gene sets G_1 and G_2 in species S , we define:

$$between_S(G_1, G_2) = \frac{1}{|G_1| |G_2|} \sum_{g_i \in G_1, g_j \in G_2} w_S(g_i, g_j)$$

$$bgrnd_S(G_1, G_2) = \frac{1}{n} \sum_{g_i} \left(\frac{1}{|G_1|} \sum_{g_j \in G_1} w_S(g_i, g_j) + \frac{1}{|G_2|} \sum_{g_j \in G_2} w_S(g_i, g_j) \right)$$

$$baseline_S = \frac{1}{n} \sum_{g_i, g_j} w_S(g_i, g_j)$$

$$within_S(G_1) = \frac{1}{|G_1|^2} \sum_{g_i, g_j \in G_1} \begin{cases} w_S(g_i, g_j) & i \neq j \\ baseline_S & i = j \end{cases}$$

$$within_S(G_1, G_2) = \frac{1}{2} (within_S(G_1) + within_S(G_2))$$

$$FA_S(G_1, G_2) = \frac{between_S(G_1, G_2)}{bgrnd_S(G_1, G_2)} \cdot \frac{baseline_S}{within_S(G_1, G_2)}$$

This score is converted into a p-value by interpolating over a bootstrapped null distribution, which for any species/network is approximately normal with standard deviation asymptotic in the sizes of the two gene sets. Fitting these empirical curves with a ratio of linear polynomials allows computation of an approximate standard deviation for any pair of gene set sizes, which also allows the conversion of functional association scores into p-values using a normal distribution function.

Acknowledgments

We would like to thank Edoardo Airoldi for initial discussions regarding biological network analysis, Sarah Fortune for invaluable input regarding *Mycobacterium tuberculosis* analysis, and Olga Troyanskaya for the inspiration to investigate experimental data integration.

References

1. Lee, I., et al., *A probabilistic functional network of yeast genes*. Science, 2004. **306**(5701): p. 1555-8.
2. Date, S.V. and C.J. Stoeckert, Jr., *Computational modeling of the Plasmodium falciparum interactome reveals protein function on a genome-wide scale*. Genome Res, 2006. **16**(4): p. 542-9.

3. Myers, C.L. and O.G. Troyanskaya, *Context-sensitive data integration and prediction of biological networks*. Bioinformatics, 2007. **23**(17): p. 2322-30.
4. Huttenhower, C., et al., *Exploring the human genome with functional maps*. Genome Res, 2009. **19**(6): p. 1093-106.
5. Murali, T.M., C.J. Wu, and S. Kasif, *The art of gene function prediction*. Nat Biotechnol, 2006. **24**(12): p. 1474-5; author reply 1475-6.
6. Kanehisa, M., et al., *KEGG for linking genomes to life and the environment*. Nucleic Acids Res, 2008. **36**(Database issue): p. D480-4.
7. Myers, C.L., et al., *Finding function: evaluation methods for functional genomic data*. BMC Genomics, 2006. **7**: p. 187.
8. Kan, B., et al., *Proteome comparison of Vibrio cholerae cultured in aerobic and anaerobic conditions*. Proteomics, 2004. **4**(10): p. 3061-7.
9. Mey, A.R., S.A. Craig, and S.M. Payne, *Characterization of Vibrio cholerae RyhB: the RyhB regulon and role of ryhB in biofilm formation*. Infect Immun, 2005. **73**(9): p. 5706-19.
10. Kudva, I.T., et al., *Identification of a protein subset of the anthrax spore immunome in humans immunized with the anthrax vaccine adsorbed preparation*. Infect Immun, 2005. **73**(9): p. 5685-96.
11. Sebahia, M., et al., *Genome sequence of a proteolytic (Group I) Clostridium botulinum strain Hall A and comparative analysis of the clostridial genomes*. Genome Res, 2007. **17**(7): p. 1082-92.
12. Tatusov, R.L., et al., *The COG database: a tool for genome-scale analysis of protein functions and evolution*. Nucleic Acids Res, 2000. **28**(1): p. 33-6.
13. Flannick, J., et al., *Graemlin: general and robust alignment of multiple large interaction networks*. Genome Res, 2006. **16**(9): p. 1169-81.
14. Camacho, L.R., et al., *Identification of a virulence gene cluster of Mycobacterium tuberculosis by signature-tagged transposon mutagenesis*. Mol Microbiol, 1999. **34**(2): p. 257-67.
15. Cardenas, E. and J.M. Tiedje, *New tools for discovering and characterizing microbial diversity*. Curr Opin Biotechnol, 2008. **19**(6): p. 544-9.
16. Huttenhower, C., et al., *A scalable method for integration and functional analysis of multiple microarray datasets*. Bioinformatics, 2006. **22**(23): p. 2890-7.
17. Parkinson, H., et al., *ArrayExpress update--from an archive of functional genomics experiments to the atlas of gene expression*. Nucleic Acids Res, 2009. **37**(Database issue): p. D868-72.
18. Huttenhower, C., et al., *The Sleipnir library for computational functional genomics*. Bioinformatics, 2008. **24**(13): p. 1559-61.