Integrated Protein Interaction Networks for 11 Microbes

Balaji S. Srinivasan[1,2], Antal F. Novak[3], Jason A. Flannick[3], Serafim Batzoglou[3], and Harley H. McAdams[2]

[1] Department of Electrical Engineering
[2] Department of Developmental Biology
[3] Department of Computer Science, Stanford University, Stanford, CA 94305, USA

Abstract. We have combined four different types of functional genomic data to create high coverage protein interaction networks for 11 microbes. Our integration algorithm naturally handles statistically dependent predictors and automatically corrects for differing noise levels and data corruption in different evidence sources. We find that many of the predictions in each integrated network hinge on moderate but consistent evidence from multiple sources rather than strong evidence from a single source, yielding novel biology which would be missed if a single data source such as coexpression or coinheritance was used in isolation. In addition to statistical analysis, we demonstrate via case study that these subtle interactions can discover new aspects of even well studied functional modules. Our work represents the largest collection of probabilistic protein interaction networks compiled to date, and our methods can be applied to any sequenced organism and any kind of experimental or computational technique which produces pairwise measures of protein interaction.

1 Introduction

Interaction networks are the canonical data sets of the post-genomic era, and more than a dozen methods to detect protein-DNA and protein-protein interactions on a genomic scale have been recently described [1, 2, 3, 4, 5, 6, 7, 8, 9]. As many of these methods require no further experimental data beyond a genome sequence, we now have a situation in which a number of different interaction networks are available for each sequenced organism. However, though many of these interaction predictors have been individually shown to predict experiment[6], the networks generated by each method are often contradictory and not superposable in any obvious way [10, 11]. This seeming paradox has stimulated a burst of recent work on the problem of network integration, work which has primarily focused on *Saccharomyces cerevisiae*[12, 13, 14, 15, 16, 17]. While the profusion of experimental network data [18] in yeast makes this focus understandable, the objective of network integration remains general: namely, a summary network

A. Apostolico et al. (Eds.): RECOMB 2006, LNBI 3909, pp. 1–14, 2006.

for each species which uses all the evidence at hand to predict which proteins are functionally linked.

In the ideal case, an algorithm to generate such a network should be able to:

1. Integrate evidence sets of various types (real valued, ordinal scale, categorical, and so on) and from diverse sources (expression, phylogenetic profiles, chromosomal location, two hybrid, etc.).
2. Incorporate known prior information (such as individually confirmed functional linkages), again of various types.
3. Cope with statistical dependencies in the evidence set (such as multiple repetitions of the same expression time course) and noisy or corrupted evidence.
4. Provide a decomposition which indicates the evidence variables which were most informative in determining a given linkage prediction.
5. Produce a unified probabilistic assessment of linkage confidence given all the observed evidence.

In this paper we present an algorithm for network integration that satisfies all five of these requirements. We have applied this algorithm to integrate four different kinds of evidence (coexpression[3], coinheritance[5], colocation[1], and coevolution[9]) to build probabilistic interaction networks for 11 sequenced microbes. The resulting networks are undirected graphs in which nodes correspond to proteins and edge weights represent interaction probabilities between protein pairs. Protein pairs with high interaction probabilities are not necessarily in direct contact, but are likely to participate in the same functional module [19], such as a metabolic pathway, a signaling network, or a multiprotein complex. We demonstrate the utility of network integration for the working biologist by analyzing representative functional modules from two microbes: the eukaryote-like glycosylation system of *Campylobacter jejuni* NCTC 11168 and the cell division machinery of *Caulobacter crescentus*. For each module, we show that a subset of the interactions predicted by our network recapitulate those described in the literature. Importantly, we find that many of the novel interactions in these modules originate in moderate evidence from multiple sources rather than strong evidence from a single source, representing hidden biology which would be missed if a single data type was used in isolation.

2 Methods

2.1 Algorithm Overview

The purpose of network integration is to systematically combine different types of data to arrive at a statistical summary of which proteins work together within a single organism.

For each of the 11 organisms listed in the Appendix[1] we begin by assembling a training set of known functional modules (Figure 1a) and a battery of different predictors (Figure 1b) of functional association. To gain intuition for what our

[1] Viewable at http://jinome.stanford.edu/pdfs/recomb06182_appendix.pdf

algorithm does, consider a single predictor E defined on a pair of proteins, such as the familiar Pearson correlation between expression vectors. Also consider a variable L, likewise defined on pairs of proteins, which takes on three possible values: '1' when two proteins are in the same functional category, '0' when they are known to be in different categories, and '?' when one or both of the proteins is of unknown function.

We note first that two proteins known to be in the same functional module are more likely to exhibit high levels of coexpression than two proteins known to be in different modules, indicated graphically by a right-shift in the distribution of $P(E|L = 1)$ relative to $P(E|L = 0)$ (Figure 1b). We can invert this observation via Bayes' rule to obtain the probability that two proteins are in the same functional module as a function of the coexpression, $P(L = 1|E)$. This posterior probability increases with the level of coexpression, as highly coexpressed pairs are more likely to participate in the same functional module.

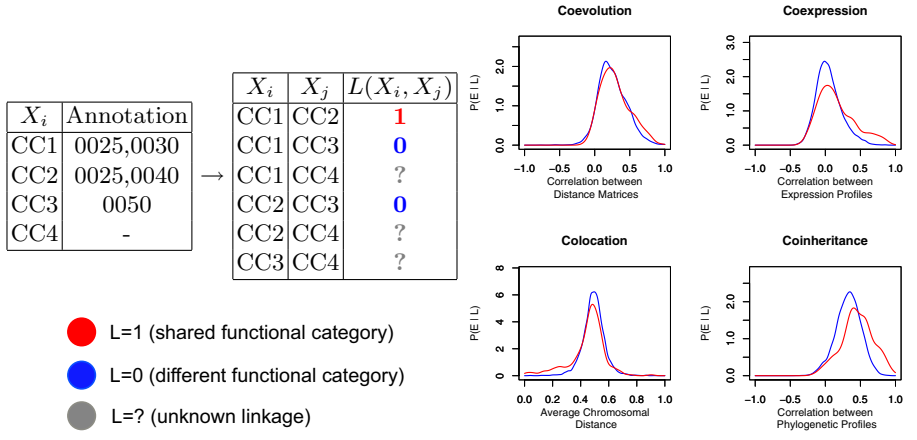
If we apply this approach to each candidate predictor in turn, we can obtain valuable information about the extent to which each evidence type recapitulates known functional linkages – or, more precisely, the efficiency with which each predictor *classifies* pairs of proteins into the "linked" or "unlinked" categories. Importantly, benchmarking each predictor in terms of its performance as a binary classifier provides a way to compare previously incomparable data sets, such as matrices[6] of BLAST[20] bit scores and arrays of Cy5/Cy3 ratios[3]. Even more importantly, it suggests that the problem of network integration can be viewed as a high dimensional binary classifier problem. By generalizing the approach outlined above to the case where E is a vector rather than a scalar, we can calculate the summary probability that two proteins are functionally linked given all the evidence at hand.

2.2 Training Set and Evidence Calculation

It is difficult to say *a priori* which predictors of functional association will be the best for a given organism. For example, microarray quality is known to vary widely, so coexpression correlations in different organisms are not directly comparable. Thus, to calibrate our interaction prediction algorithm, we require a training set of known interactions.

To generate this training set, we used one of three different genome scale annotations: the COG functional categories assigned by NCBI[21], the GO[22] annotations assigned by EBI's GOA project[23], and the KEGG[24] metabolic annotations assigned to microbial genomes. In general, as we move from COG to GO to KEGG, the fraction of annotated proteins in a given organism decreases, but the annotation quality increases. In this work we used the KEGG annotation for all organisms other than *Bacillus subtilis*, for which we used GO as KEGG data was unavailable.

As shown in Figure 1a, for each pair we recorded ($L = 1$) if the proteins had overlapping annotations, ($L = 0$) if both were in entirely nonoverlapping categories, and ($L = ?$) if either protein lacked an annotation code or was marked as unknown. (For the GO training set, "overlapping" was defined as overlap



X_i	Annotation
CC1	0025,0030
CC2	0025,0040
CC3	0050
CC4	-

→

X_i	X_j	$L(X_i, X_j)$
CC1	CC2	1
CC1	CC3	0
CC1	CC4	?
CC2	CC3	0
CC2	CC4	?
CC3	CC4	?

(a) Training Set Generation (b) Evidence vs. Training Set

Fig. 1. Training Sets and Evidence. (a) Genome-scale systematic annotations such as COG, GO or KEGG give functions for proteins X_i. As described in the text and shown on example data, we use this annotation to build an initial classification of protein pairs (X_i, X_j) with three categories: a relatively small set of likely linked (red) pairs and unlinked (blue) pairs, and a much larger set of uncertain (gray) pairs. (b) We observe that proteins which share an annotation category generally have more significant levels of evidence, as seen in the shifted distribution of linked (red) vs. unlinked (blue) pairs. Even subtle distributional differences contribute statistical resolution to our algorithm.

of specific GO categories beyond the 8th level of the hierarchy.) This "matrix" approach (consider all proteins within an annotation category as linked) is in contrast to the "hub-spoke" approach (consider only proteins known to be directly in contact as linked) [25]. The former representation produces a nontrivial number of false positives, while the latter incurs a surfeit of false negatives. We chose the "matrix" based training set because our algorithm is robust to noise in the training set so long as enough data is present.

Note that we have used an annotation on individual proteins to produce a training set on *pairs* of proteins. In Figure 1b, we compare this training set to four functional genomic predictors: coexpression, coinheritance, coevolution, and colocation. We include details of the calculations of each evidence type in the Appendix. Interestingly, despite the fact that these methods were obtained from raw measurements as distinct as genomic spacing, BLAST bit scores, phylogenetic trees, and microarray traces, Figure 1b shows that each method is capable of distinguishing functionally linked pairs ($L = 1$) from unlinked pairs ($L = 0$).

2.3 Network Integration

For clarity, we first illustrate network integration with two evidence types (corresponding to two Euclidean dimensions) in *C. crescentus*, and then move to the N-dimensional case.

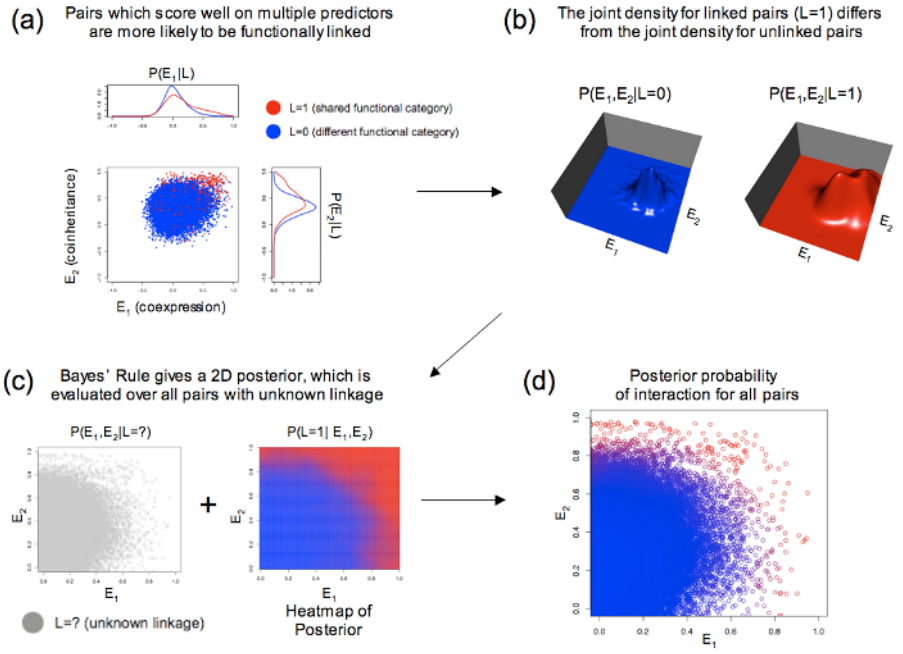


Fig. 2. 2D Network Integration in *C. crescentus*. (a) A scatterplot reveals that functionally linked pairs (red,$L = 1$) tend to have higher coexpression and coinheritance than pairs known to participate in separate pathways (blue,$L = 0$). (b) We build the conditional densities $P(E_1, E_2|L = 0)$ and $P(E_1, E_2|L = 1)$ through kernel density estimation. Note that the distribution for linked pairs is shifted to the upper right corner relative to the unlinked pair distribution. (c) We can visualize the classification process by concentrating on the decision boundary, corresponding to the upper right quadrant of the original plot. In the left panel, the scatterplot of pairs with unknown linkage status (gray) are the inputs for which we wish to calculate interaction probabilities. In the right panel, a heatmap for the posterior probability $P(L = 1|E_1, E_2)$ is depicted. This function yields the probability of linkage given an input evidence vector, and increases as we move to higher levels of coexpression and coinheritance in the upper right corner. (d) By conceptually superimposing each gray point upon the posterior, we can calculate the posterior probability that two proteins are functionally linked.

2D Network Integration. Consider the set of approximately 310000 protein pairs in *C. crescentus* which have a KEGG-defined linkage of $(L = 0)$ or $(L = 1)$. Setting aside the 6.6 million pairs with $(L = ?)$ for now, we find that $P(L = 1) = .046$ and $P(L = 0) = .954$ are the relative proportions of known linked and unlinked pairs in our training set.

Each of these pairs has an associated coexpression and coinheritance correlation, possibly with missing values, which we bundle into a two dimensional vector $E = (E_1, E_2)$. Figure 2a shows a scatterplot of E_1 vs. E_2, where pairs with $(L = 1)$ have been marked red and pairs with $(L = 0)$ have been marked blue.

We see immediately that functionally linked pairs aggregate in the upper right corner of the plot, in the region of high coexpression and coinheritance.

Crucially, the linked pairs (red) are more easily distinguished from the unlinked pairs (blue) in the 2-dimensional scatter plot than they are in the accompanying 1-dimensional marginals. To quantify the extent to which this is true, we begin by computing $P(E_1, E_2|L = 0)$ and $P(E_1, E_2|L = 1)$ via kernel density estimation[26, 27], as shown in Figure 2b. As we already know $P(L)$, we can obtain the posterior by Bayes' rule:

$$P(L = 1|E_1, E_2) = \frac{P(E_1, E_2|L = 1)P(L = 1)}{P(E_1, E_2|L = 1)P(L = 1) + P(E_1, E_2|L = 0)P(L = 0)}$$

In practice, this expression is quite sensitive to fluctuations in the denominator. To deal with this, we use M-fold bootstrap aggregation[28] to smooth the posterior. We find that $M = 20$ repetitions with resampling of 1000 elements from the $(L = 0)$ and $(L = 1)$ training sets is the empirical point of diminishing returns in terms of area under the receiver-operator characteric (ROC), as detailed in Figure 4.

$$P(L = 1|E_1, E_2) = \frac{1}{M}\sum_{i=1}^{M} \frac{P_i(E_1, E_2|L = 1)P(L = 1)}{P_i(E_1, E_2|L = 1)P(L = 1) + P_i(E_1, E_2|L = 0)P(L = 0)}$$

Given this posterior, we can now make use of the roughly 6.6 million pairs with $(L = ?)$ which we put aside at the outset, as pictured in Figure 2c. Even though these pairs have unknown linkage, for most pairs the coexpression (E_1) and coinheritance (E_2) are known. For those pairs which have partially missing data (e.g. from corrupted spots on a microarray), we can simply evaluate over the non-missing elements of the E vector by using the appropriate marginal posterior $P(L = 1|E_1)$ or $P(L = 1|E_2)$. We can thus calculate $P(L = 1|E_1, E_2)$ for every pair of proteins in the proteome, as shown in Figure 2d. Each of the formerly gray pairs with $(L = ?)$ is assigned a probability of interaction by this function; those with bright red values in Figure 2d are highly likely to be functionally linked.

In general, we also calculate $P(L = 1|E_1, E_2)$ on the training data, as we know that the "matrix" approach to training set generation produces copious but noisy data. The result of this evaluation is the probability of interaction for every protein pair.

N-dimensional Network Integration. The 2 dimensional example in *C. crescentus* immediately generalizes to N-dimensional network integration in an arbitrary species, though the results cannot be easily visualized beyond 3 dimensions. Figure 3 shows the results of calculating a 3D posterior in *C. crescentus* from coexpression, coinheritance, and colocation data, where we have once again applied M-fold bootstrap aggregation.

We see that different evidence types interact in nonobvious ways. For example, we note that high levels of colocation (E_2) can compensate for low levels of coexpression (E_1), as indicated by the "bump" in the posterior of Figure 3c. Biologically speaking, this means that a nontrivial number of *C. crescentus* proteins with shared function are frequently colocated yet not strongly coexpressed. This is exactly the sort of subtle statistical dependence between predictors that is crucial for proper classification. In fact, a theoretically attractive property of

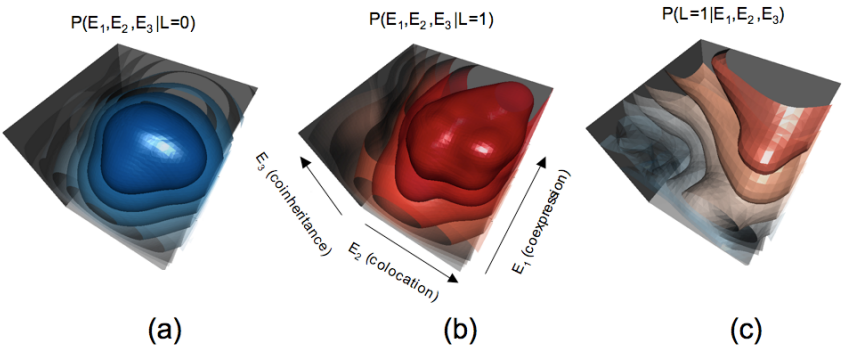
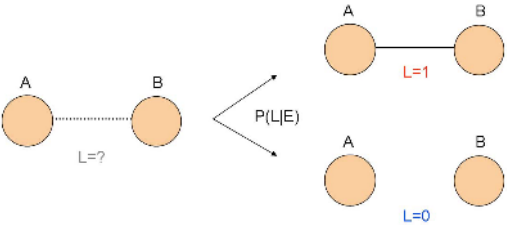


Fig. 3. 3D Network Integration in *C. crescentus*. (a)-(b) We show level sets of each density spaced at even volumetric increments, so that the inner most shell encloses 20% of the volume, the second shell encloses 40%, and so forth. As in the 2D case, the 3D density $P(E|L = 1)$ is shifted to the upper right corner. (c) For the posterior, we show level sets spaced at probability deciles, such that a pair which makes it past the upper right shell has $P(L = 1|E) \in [.9, 1]$, a pair which lands in between the upper two shells satisfies $P(L = 1|E) \in [.8, .9]$, and so on.

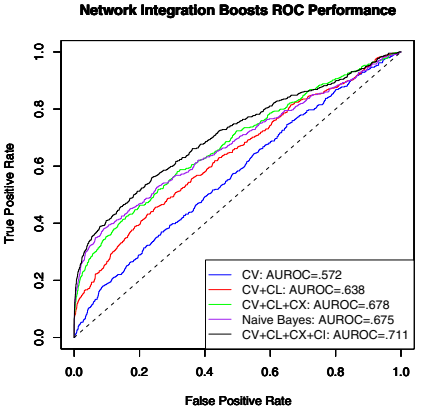
our approach is that the use of the conditional joint posterior produces the minimum possible classification error (specifically, the Bayes error rate [29]), while bootstrap aggregation protects us against overfitting[30].

Until recently, though, technical obstacles made it challenging to efficiently compute joint densities beyond dimension 3. Recent developments[26] in efficient kernel density estimation have obviated this difficulty and have made it possible to evaluate high dimensional densities over millions of points in a reasonable amount of time within user-specifiable tolerance levels. As an example of the calculation necessary for network integration, consider a 4 dimensional kernel density estimate built from 1000 sample points. Ihler's implementation[27] of the Gray-Moore dual-tree algorithm[26] allowed the evaluation of this density at the $\binom{3737}{2} \approx 7,000,000$ pairs in the *C. crescentus* proteome in only 21 minutes on a 3GHz Xeon with 2GB RAM. Even after accounting for the $2M$ multiple of this running time caused by evaluating a quotient of two densities and using M-fold bootstrap aggregation, the resulting joint conditional posterior can be built and evaluated rapidly enough to render approximation unnecessary.

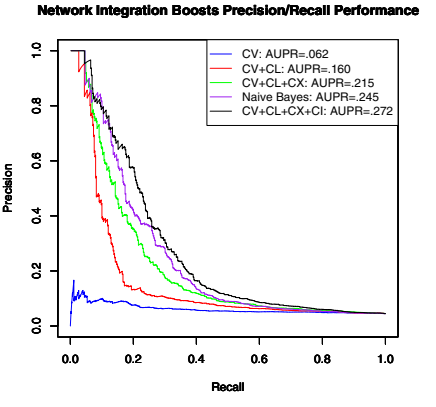
Binary Classifier Perspective. By formulating the network integration problem as a binary classifier (Figure 4), we can quantify the extent to which the integration of multiple evidence sources improves prediction accuracy over a single source. As our training data is necessarily a rough approximation of the true interaction network, these measures are likely to be conservative estimates of classifier performance.



(a) Binary Classifier Paradigm



(b) Receiver/Operator Curves



(c) Precision/Recall Curves

Fig. 4. Network Integration as Binary Classifier. (a) We regard the network integration problem as a binary classifier in a high dimensional feature space. The input features are a set of evidences associated with a protein pair (A, B), and the output is the probability that a pair is assigned to the $(L = 1)$ category. (b) The area under the receiver operator characteristic (AUROC) is a standard measure[29] of binary classifier performance, shown here for several different ways of doing *C. crescentus* network integration. Here we have labeled data types as CV (coevolution), CL (colocation), CX (coexpression), and CI (coinheritance) and shown a successive series of curves for the integration of 1,2,3, and finally 4 evidence types. Classifier performance increases monotonically as more data sets are combined. Importantly, the true four dimensional joint posterior $P(L = 1|CV, CL, CX, CI)$ outperforms the Naive Bayes approximation of the posterior, where the conditional density $P(CV, CL, CX, CI|L = 1)$ is approximated by $P(CV|L = 1)P(CL|L = 1)P(CX|L = 1)P(CI|L = 1)$, and similarly for $L = 0$. For clarity we have omitted the individual curves for the CL (AUROC=.612), CX (AUROC=.619), and CV (AUROC=.653) metrics. Again, it is clear that the integrated posterior outperforms each of these univariate predictors. (c) Precision/recall curves are an alternate way of visualizing classifier performance, and are useful when the number of true positives is scarce relative to the number of false negatives. Again the integrated posterior outperforms the Naive Bayes approximation as a classifier. Note that since the "negative" pairs from the KEGG training set are based on the supposition that two proteins which have no annotational overlap genuinely do *not* share a pathway, they are a more noisy indicator than the "positive" pairs. That is, with respect to functional interaction, absence of evidence is not always evidence of absence. Hence the computed values for precision are likely to be conservative underestimates of the true values.

3 Results

3.1 Global Network Architecture

Applying the posterior $P(L = 1|E)$ to every pair of proteins in a genome gives the probability that each pair is functionally linked. If we simply threshold this result at $P(L = 1|E) > .5$, we will retain only those linkages which are more probable than not. This decision rule attains the Bayes error rate[29] and minimizes the misclassification probability. We applied our algorithm with this threshold to build 4D integrated networks for the 11 microbes and four evidence types listed in the Appendix. Figure 5 shows the global protein interaction networks produced for three of these microbes, where we have retained only those edges with $P(L = 1|E) > .5$.

To facilitate use of these protein interaction networks, we built an interactive netbrowser, viewable at http://jinome.stanford.edu/netbrowser. As a threshold of $P(L = 1|E) > .5$ tends to be somewhat stringent in practice, we allow dynamic, user-specified thresholds to produce module-specific tradeoffs between specificity and sensitivity in addition to a host of other customization options.

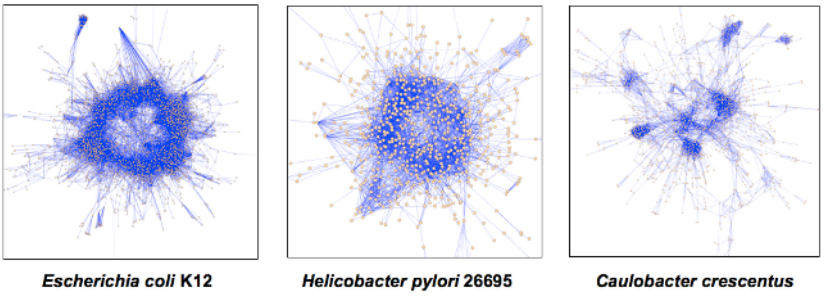


Fig. 5. Global visualization of integrated networks for *Escherichia coli* K12, *Helicobacter pylori* 26695, and *Caulobacter crescentus*. Only linkages with $P(L = 1|E_1, E_2, E_3, E_4) > .5$ are displayed.

3.2 *Campylobacter jejuni*: N-Linked Protein Glycosylation

N-linked protein glycosylation is one of the most frequent post-translational modifications applied to eukaryotic secretory proteins. Until recently[31] this process was thought to be absent from most microbes, but recent work[32] has shown that an operational N-linked glycosylation system does exist in *C. jejuni*. As the entire glycosylation apparatus can be successfully transplanted to *E. coli* K12, this system is of much biotechnological interest[33].

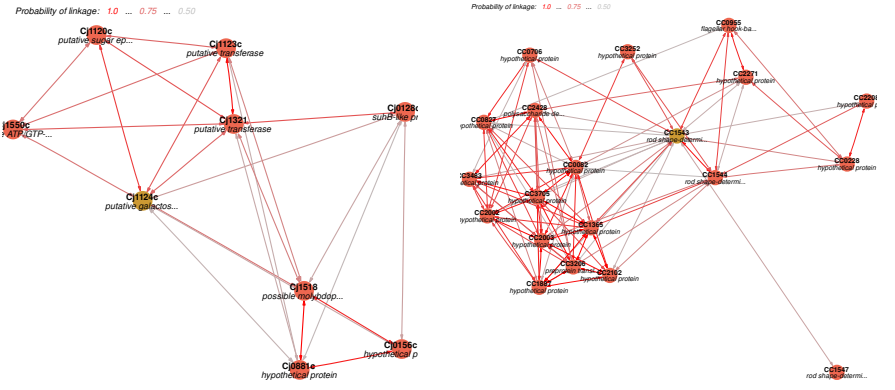
Figure 6a shows the results of examining the integrated network for *C. jejuni* around the vicinity of Cj1124c, one of the proteins in the glycosylation system. In addition to the reassuring recapitulation of several transferases and epimerases experimentally linked to this process[33], we note four proteins which are to our knowledge not known to be implicated in N-linked glycosylation (Cj1518,

Cj0881c, Cj0156c, Cj0128c). Importantly, all of these heretofore uncharacterized linkages would have been missed if only univariate posteriors had been examined, as they would be significantly below our cutoff of $P(L = 1|E) > .5$. As this system is still poorly understood – yet of substantial biotechnological and pathogenic[34] relevance – investigation of these new proteins may be of interest.

3.3 *Caulobacter crescentus*: Bacterial Actin and the Sec Apparatus

Van den Ent's[36] discovery that the ubiquitous microbial protein MreB was a structural homolog to actin spurred a burst of interest[37, 38, 39] in the biology of the bacterial cytoskeleton. Perhaps the most visually arresting of these recent findings is the revelation that MreB supports the cell by forming a tight spiral[37]. Yet many outstanding questions in this field remain, and prime among them is the issue of which proteins communicate with the bacterial cytoskeletal apparatus[40].

Figure 6b shows the proteins from the *C. crescentus* integrated network which have a 50% chance or greater of interacting with MreB, also known as CC1543. As a baseline measure of validity, we once again observe that known interaction partners such as RodA (CC1547) and MreC (CC1544) are recovered by network integration. More interesting, however, is the subtle interaction between MreB and the preprotein translocase CC3206, an interaction that would be missed if data sources were used separately. This protein is a subunit of the Sec machinery,



(a) *C. jejuni*: Glycosylation (b) *C. crescentus*: Bacterial cytoskeleton

Fig. 6. Case Studies. (a) Network integration detects new proteins linked to glycosylation in *Campylobacter jejuni* NCTC 11168. High probability linkages are labeled in red and generally recapitulate known interactions, while moderately likely linkages are colored gray. Moderate linkages are generally not found by any univariate method in isolation, and represent the new biological insight produced by data integration. (b) In *Caulobacter crescentus*, data integration reveals that the Sec apparatus is linked to MreB, a prediction recently confirmed by experiment[35]. Again, moderate linkages revealed by data integration lead us to a conclusion that would be missed if univariate data was used.

and like MreB is an ancient component of the bacterial cell[41]. Its link to MreB is of particular note because recent findings[35] have shown that the Sec apparatus – like MreB – has a spiral localization pattern. While seemingly counterintuitive, it seems likely from both this finding and other work[42] that the export of cytoskeleton-related proteins beyond the cellular membrane is important in the process of cell division. We believe that investigation of the hypothetical proteins linked to both MreB and Sec by our algorithm may shed light on this question.

4 Discussion

4.1 Merits of Our Approach

While a number of recent papers on network integration in *S. cerevisiae* have appeared, we believe that our method is an improvement over existing algorithms.

First, by directly calculating the joint conditional posterior we require no simplifying assumptions about statistical dependence and need no complex parametric inference. In particular, removing the Naive Bayes approximation results in a better classifier, as quantified in Figure 4. Second, our use of the Gray-Moore dual tree algorithm means that our method is arbitrarily scalable in terms of both the number of evidence types and the number of protein pairs. Third, our method allows immediate visual identification of dependent or corrupted functional genomic data in terms of red/blue separation scatterplots – an important consideration given the noise of some data types [43]. Finally, because the output of our algorithm is a rigorously derived set of interaction probabilities, it represents a solid foundation for future work.

4.2 Conclusion and Future Directions

Our general framework presents much room for future development. It is straightforward to generalize our algorithm to apply to discrete, ordinal, or categorical data sets as long as appropriate similarity measures are defined. As our method readily scales beyond a few thousand proteins, even the largest eukaryotic genomes are potential application domains. It may also be possible to improve our inference algorithm through the use of statistical techniques designed to deal with missing data[44].

Moving beyond a binary classifier would allow us to predict different kinds of functional linkage, as two proteins in the same multiprotein complex have a different kind of linkage than two proteins which are members of the same regulon. This would be significant in that it addresses one of the most widely voiced criticisms of functional genomics, which is that linkage predictions are "one-size-fits-all". It may also be useful to move beyond symmetric pairwise measures of association to use metrics defined on protein triplets[8] or asymmetric metrics such that $E(P_i, P_j) \neq E(P_j, P_i)$.

While these details of the network construction process are doubtless subjects for future research, perhaps the most interesting prospect raised by the availability of a large number of robust, integrated interaction networks is the possibility

of comparative modular biology. Specifically, we would like to *align* subgraphs of interaction networks on the basis of conserved interaction as well as conserved sequence, just as we align DNA and protein sequences. A need now exists for a network alignment algorithm capable of scaling to large datasets and comparing many species simultaneously.

Acknowledgments

We thank Lucy Shapiro, Roy Welch, and Arend Sidow for helpful discussions. BSS was supported in part by a DoD/NDSEG graduate fellowship, and HHM and BSS were supported by NIH grant 1 R24 GM073011-01 and DOE Office of Science grant DE-FG02-01ER63219. JAF was supported in part by a Stanford Graduate Fellowship, and SB, AFN, and JAF were funded by NSF grant EF-0312459, NIH grant UO1-HG003162, the NSF CAREER Award, and the Alfred P. Sloan Fellowship.

Authors' Contributions

BSS developed the network integration algorithm and wrote the paper. AFN designed the web interface with JAF under the direction of SB and provided useful feedback on network quality. HHM and SB provided helpful comments and a nurturing environment.

References

1. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D., Maltsev, N.: The use of gene clusters to infer functional coupling. Proc Natl Acad Sci U S A **96** (1999) 2896–2901
2. McAdams, H.H., Srinivasan, B., Arkin, A.P.: The evolution of genetic regulatory systems in bacteria. Nat Rev Genet **5** (2004) 169–178
3. Schena, M., Shalon, D., Davis, R.W., Brown, P.O.: Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science **270** (1995) 467–470
4. Enright, A.J., Iliopoulos, I., Kyrpides, N.C., Ouzounis, C.A.: Protein interaction maps for complete genomes based on gene fusion events. Nature **402** (1999) 86–90
5. Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., Yeates, T.O.: Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A **96** (1999) 4285–4288
6. Srinivasan, B.S., Caberoy, N.B., Suen, G., Taylor, R.G., Shah, R., Tengra, F., Goldman, B.S., Garza, A.G., Welch, R.D.: Functional genome annotation through phylogenomic mapping. Nat Biotechnol **23** (2005) 691–698
7. Yu, H., Luscombe, N.M., Lu, H.X., Zhu, X., Xia, Y., Han, J.D.J., Bertin, N., Chung, S., Vidal, M., Gerstein, M.: Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. Genome Res **14** (2004) 1107–1118
8. Bowers, P.M., Cokus, S.J., Eisenberg, D., Yeates, T.O.: Use of logic relationships to decipher protein network organization. Science **306** (2004) 2246–2249

9. Pazos, F., Valencia, A.: Similarity of phylogenetic trees as indicator of protein-protein interaction. Protein Eng **14** (2001) 609–614 Evaluation Studies.
10. Gerstein, M., Lan, N., Jansen, R.: Proteomics. Integrating interactomes. Science **295** (2002) 284–287 Comment.
11. Hoffmann, R., Valencia, A.: Protein interaction: same network, different hubs. Trends Genet **19** (2003) 681–683
12. Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N.J., Chung, S., Emili, A., Snyder, M., Greenblatt, J.F., Gerstein, M.: A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science **302** (2003) 449–453 Evaluation Studies.
13. Troyanskaya, O.G., Dolinski, K., Owen, A.B., Altman, R.B., Botstein, D.: A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). Proc Natl Acad Sci U S A **100** (2003) 8348–8353
14. Lee, I., Date, S.V., Adai, A.T., Marcotte, E.M.: A probabilistic functional network of yeast genes. Science **306** (2004) 1555–1558
15. Tanay, A., Sharan, R., Kupiec, M., Shamir, R.: Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. Proc Natl Acad Sci U S A **101** (2004) 2981–2986
16. Wong, S.L., Zhang, L.V., Tong, A.H.Y., Li, Z., Goldberg, D.S., King, O.D., Lesage, G., Vidal, M., Andrews, B., Bussey, H., Boone, C., Roth, F.P.: Combining biological networks to predict genetic interactions. Proc Natl Acad Sci U S A **101** (2004) 15682–15687
17. Lu, L.J., Xia, Y., Paccanaro, A., Yu, H., Gerstein, M.: Assessing the limits of genomic data integration for predicting protein networks. Genome Res **15** (2005) 945–953
18. Friedman, A., Perrimon, N.: Genome-wide high-throughput screens in functional genomics. Curr Opin Genet Dev **14** (2004) 470–476
19. Hartwell, L.H., Hopfield, J.J., Leibler, S., Murray, A.W.: From molecular to modular cell biology. Nature **402** (1999) 47–52
20. Schaffer, A.A., Aravind, L., Madden, T.L., Shavirin, S., Spouge, J.L., Wolf, Y.I., Koonin, E.V., Altschul, S.F.: Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res **29** (2001) 2994–3005
21. Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., Rao, B.S., Smirnov, S., Sverdlov, A.V., Vasudevan, S., Wolf, Y.I., Yin, J.J., Natale, D.A.: The COG database: an updated version includes eukaryotes. BMC Bioinformatics **4** (2003) 41
22. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G.: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet **25** (2000) 25–29
23. Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., Binns, D., Harte, N., Lopez, R., Apweiler, R.: The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. Nucleic Acids Res **32** (2004) 262–266
24. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., Hattori, M.: The KEGG resource for deciphering the genome. Nucleic Acids Res **32** (2004) 277–280

25. Bader, G.D., Hogue, C.W.V.: Analyzing yeast protein-protein interaction data obtained from different sources. Nat Biotechnol **20** (2002) 991–997
26. Gray, A.G., Moore, A.W.: 'n-body' problems in statistical learning. In: NIPS. (2000) 521–527
27. Ihler, A., Sudderth, E., Freeman, W., Willsky, A.: Efficient multiscale sampling from products of gaussian mixtures. In: NIPS. (2003)
28. Breiman, L.: Bagging predictors. Machine Learning **24** (1996) 123–140
29. Duda, R., Hart, P., Stork, D.: Pattern Classification. Wiley-Interscience Publication (2000)
30. Bauer, E., Kohavi, R.: An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. Machine Learning **36** (1999) 105–139
31. Szymanski, C.M., Logan, S.M., Linton, D., Wren, B.W.: Campylobacter–a tale of two protein glycosylation systems. Trends Microbiol **11** (2003) 233–238
32. Wacker, M., Linton, D., Hitchen, P.G., Nita-Lazar, M., Haslam, S.M., North, S.J., Panico, M., Morris, H.R., Dell, A., Wren, B.W., Aebi, M.: N-linked glycosylation in Campylobacter jejuni and its functional transfer into E. coli. Science **298** (2002) 1790–1793
33. Linton, D., Dorrell, N., Hitchen, P.G., Amber, S., Karlyshev, A.V., Morris, H.R., Dell, A., Valvano, M.A., Aebi, M., Wren, B.W.: Functional analysis of the Campylobacter jejuni N-linked protein glycosylation pathway. Mol Microbiol **55** (2005) 1695–1703
34. Karlyshev, A.V., Everest, P., Linton, D., Cawthraw, S., Newell, D.G., Wren, B.W.: The Campylobacter jejuni general glycosylation system is important for attachment to human epithelial cells and in the colonization of chicks. Microbiology **150** (2004) 1957–1964
35. Campo, N., Tjalsma, H., Buist, G., Stepniak, D., Meijer, M., Veenhuis, M., Westermann, M., Muller, J.P., Bron, S., Kok, J., Kuipers, O.P., Jongbloed, J.D.H.: Subcellular sites for bacterial protein export. Mol Microbiol **53** (2004) 1583–1599
36. van den Ent, F., Amos, L.A., Lowe, J.: Prokaryotic origin of the actin cytoskeleton. Nature **413** (2001) 39–44
37. Gitai, Z., Dye, N., Shapiro, L.: An actin-like gene can determine cell polarity in bacteria. Proc Natl Acad Sci U S A **101** (2004) 8643–8648
38. Kurner, J., Frangakis, A.S., Baumeister, W.: Cryo-electron tomography reveals the cytoskeletal structure of Spiroplasma melliferum. Science **307** (2005) 436–438
39. Gerdes, K., Moller-Jensen, J., Ebersbach, G., Kruse, T., Nordstrom, K.: Bacterial mitotic machineries. Cell **116** (2004) 359–366
40. Cabeen, M.T., Jacobs-Wagner, C.: Bacterial cell shape. Nat Rev Microbiol **3** (2005) 601–610
41. Vrontou, E., Economou, A.: Structure and function of SecA, the preprotein translocase nanomotor. Biochim Biophys Acta **1694** (2004) 67–80
42. Kruse, T., Bork-Jensen, J., Gerdes, K.: The morphogenetic MreBCD proteins of Escherichia coli form an essential membrane-bound complex. Mol Microbiol **55** (2005) 78–89
43. Vidalain, P.O., Boxem, M., Ge, H., Li, S., Vidal, M.: Increasing specificity in high-throughput yeast two-hybrid experiments. Methods **32** (2004) 363–370
44. McLachlan, G., Krishnan, T.: The EM Algorithm and Extensions. John Wiley and Sons (1996)