Pacific Symposium on Biocomputing 12:433-444(2007)

ASSESSING AND COMBINING RELIABILITY OF PROTEIN INTERACTION SOURCES

SONIA LEACH, AARON GABOW, AND LAWRENCE HUNTER
University of Colorado at Denver and Health Sciences Center,
Aurora, CO 80045, USA
E-mail: {Sonia.Leach, Aaron.Gabow, Larry.Hunter}@uchsc.edu

DEBRA S. GOLDBERG
University of Colorado at Boulder,
Boulder, CO 80309, USA
E-mail: Debra.Goldberg@colorado.edu

Integrating diverse sources of interaction information to create protein networks requires strategies sensitive to differences in accuracy and coverage of each source. Previous integration approaches calculate reliabilities of protein interaction information sources based on congruity to a designated 'gold standard.' In this paper, we provide a comparison of the two most popular existing approaches and propose a novel alternative for assessing reliabilities which does not require a gold standard. We identify a new method for combining the resultant reliabilities and compare it against an existing method. Further, we propose an extrinsic approach to evaluation of reliability estimates, considering their influence on the downstream tasks of inferring protein function and learning regulatory networks from expression data. Results using this evaluation method show 1) our method for reliability estimation is an attractive alternative to those requiring a gold standard and 2) the new method for combining reliabilities is less sensitive to noise in reliability assignments than the similar existing technique.

1. Introduction

The recent availability of high-throughput proteomics data has allowed genome-wide construction of network models of relationships among proteins[1–12]. These networks are used in such downstream tasks as inferring protein function, identifying potential protein complexes, or interpreting gene expression data. In the majority of cases, however, the notion of protein interaction is biased to mean physical interaction. As argued in Lee *et al.*[9], the term 'interaction' should instead encompass *any* type of evidence linking pairs of genes, whether it be physical, functional, ge-

Pacific Symposium on Biocomputing 12:433-444(2007)

netic, biochemical, evolutionary, or computational. Integrated networks from diverse sources give biologists more insight into their data when these networks are later used for analysis tasks, since each interaction data type offers an alternative view of the relationships which exist among genes.

The major challenge of integration has been that each individual source varies in terms of accuracy and coverage over the domain. Thus, estimating confidence of a particular interaction must account for the number and reliability of the specific sources contributing evidence for that interaction[a]. For example, a high-throughput method offers evidence of interaction genome-wide, yet with many false positives, so interactions supported solely by that method may be suspect a priori. One answer is to favor interactions supported by multiple data sources. Even then, high confidence in an interaction is not guaranteed since the individual reliabilities of the sources supporting that interaction may be low. As such, a large body of literature is dedicated to estimating error rates of the individual experiment types[2–12]. However, nearly all existing methods quantify reliability with respect to a 'gold standard,' such as the percentage of interacting pairs suggested by the source known to have the same function.

We argue that requiring a gold standard for reliability assessments is disadvantageous for a number of reasons. First, the choice of gold standard depends on the task for which a protein interaction network is used and therefore cannot easily generalize to other tasks without recomputation. For example, using correlated gene expression profiles as a gold standard for reliability assessment may not be appropriate for the task of predicting physical interactions. Also, the need to reserve an information source as a gold standard decreases the amount of data providing evidence of interaction. This point becomes critical in less well-studied organisms where there may be few sources of information of interaction, let alone enough information for a gold standard.

In this paper, we present a new method for assigning reliability to individual data types which does not rely on a gold standard. Rather, our method leverages on the relative frequency of overlap between the sources, rating each source by its average agreement with the *consensus*. Our method is not biased toward sources representing the same notion of interaction found in a predefined gold standard and as such, is generally ap-

[a]Note here we are quantifying reliability of evidence for a *particular* interaction, not building a general predictor of interaction; thus some issues, such as missing data for predictive attributes, are not present in our application.

plicable for use in any downstream task. We compare[b] our method to two popular existing reliability assignment techniques[3,9–12]. We then consider strategies for creating a weighted protein interaction network by probabilistic integration of the individual data sources and their assigned reliabilities. For integration of reliabilities, we use one technique[c] appearing in the biological literature[11,12] as well as a new alternative we identify as applicable from the statistics literature[13]. We propose an extrinsic evaluation strategy which measures performance of each reliability assessment and integration method combination on two downstream tasks: inferring protein function and learning regulatory networks. Our results show that our proposed reliability assessment method is a viable alternative to previous methods and the alternative integration method is less sensitive to incorrect reliability assessments than the existing method.

2. Data Sources and Reliability Assignments

Protein interaction networks represent proteins as nodes and integrate interaction sources to identify connections between them. We focus here on techniques which 1) quantify the accuracy of sources indicating that interaction and 2) combine their reliabilities to obtain a consensus reliability for each interaction. Before addressing the first task, we must first identify a set of genes and a set of interaction sources. To create comparable datasets, we use the 6760 annotated orfs in yeast and choose 6760 mouse genes randomly among those having information in at least three interaction sources. With this strategy we obtain similar coverage, where 50% of our mouse genes have a known pathway, compared to 80% with known function in yeast. We use these sets of 6760 genes for all reported results.

We consider two types of data sources which provide positive assertions of relationships: *explicit sources* which indicate interaction between genes directly, or *implicit sources* from which relationships can be derived by noting when two genes are assigned the same category by the source. For example, yeast two-hybrid assays are an explicit measure of physical interaction while presence of identical sequence motifs implies genes may have related functions or regulators. Though implicit sources individually may be poor indicators of interaction, accumulation of evidence from several implicit sources may reliably indicate interaction in the absence of

[b]The day of submission we identified a similar comparison paper by Suthram *et al.* [14]
[c]This measure has identical use by independent research groups. We do not include logistic regression approaches[6,7] since each group uses very different attribute sets.

more explicit information. This point becomes critical in less well-studied organisms where indirect information may be easier to obtain.

For explicit sources, we include those which experimentally measure physical, biochemical, and genetic interactions[15–20] as well as those which computationally predict gene neighborhoods, gene fusion events or conserved phylogenetic profiles[21]. In an effort to create independent indicators, we categorize information from these sources by type, *e.g.* yeast two-hybrid or gene fusion[d] Reliabilities then are calculated for each type (denoted with capital letters, e.g. Y2H and GENEFUSE). As some indication of diversity of types and their coverage, for yeast, we have 21 distinct types with between 208 (GENETIC) and 26k (HMS-PCI) interactions. For mouse, we use 11 types with between 1 (ELISA) and 2546 (IMMUNOPRECIP) interactions. As implicit interaction sources, we use information about literature references[22], sequence motifs[23–26], protein categories[23], protein complexes[23], cell phase[27], phenotypes[23,22], essentiality[23], cellular location[22,23,28], molecular function[23,28], and biological process or pathway[28–30]. Considering each separately for the 6760 gene sets, we obtain eight implicit yeast sources with between 39k (COPROTCATEGORY) and 179k (COESSENTIAL) interactions. In mouse, we have six sources with between 30k (COLITERATURE) and 1.2M (GO:COMPONENT) interactions.

Having identified our data sources and interaction types, our first task is to assign a reliability score to each type which reflects our confidence in its information. For example, we might assign a low score to a high-throughput explicit type like yeast two-hybrid or an implicit type like co-location, yet assign a high score to a low-throughput assay like x-ray crystallography. Many methods exist for estimating reliability where the accuracy of a source is quantified by agreement to a gold standard, such as correlated gene expression, or shared protein complex membership, function or cellular location. We consider two of the most prevalent techniques which rely on a gold standard and propose a third which does not.

The first measure calculates reliability r_E as the *proportion* of pairs from a source E with a known shared designation according to the gold standard, relative to all pairs annotated by the source[11,12]. We denote this measure PropGS. For example, we might count the proportion of interacting pairs suggested by the source with the same function. The second measure of reliability calculates the *log likelihood* of pairs sharing a designation according

[d]Note, here we do not use any type of interaction evidence which is measured experimentally in another species and transferred to the species of interest using orthologs.

to the gold standard (denoted LogLikGS)[3,9,10]: $r_E = \log(\frac{Pr(L|E)/\neg Pr(L|E)}{Pr(L)/\neg Pr(L)})$. We say two proteins are *linked* (L) if they share the same designation in a gold standard. Then $Pr(L)$ is the prior expectation of linkage, while $\neg Pr(L)$ is the prior expectation of non-linkage. The values $Pr(L|E)$ and $\neg Pr(L|E)$ represent the analogous expectations calculated only among interactions offered by the data source E. Note that PropGS computes $Pr(L|E)$. In both of methods, the gold standard is predetermined and held in reserve of the other information sources.

We propose a third alternative, not utilizing a gold standard, which relies instead on average agreement of a source with the overall *consensus* offered by all sources[e], denoted Cons. Let n_e be the number of sources indicating an interaction (edge) e between a given pair proteins. We calculate reliability as: $r_E = \frac{\sum_{e \in E} n_e}{|E|}$, where $|E|$ is the total number of interactions offered by source E. Applicability of this measure assumes relative sparsity for a good proportion of the sources. Reliability of sources which assert many edges, such as the implicit or high-throughput sources, will then be automatically discounted since many of those edges will not have further support among all experts; the average will thus be taken over many small values of n_e. Unlike the previous alternatives, Cons favors a more diverse notion of interaction since it does not penalize sources whose interaction type differs from that of the gold standard.

Given the set of reliability estimates for each interaction information source, our second task is to create a combined reliability score of each interaction based on the individual reliabilities of the sources contributing information for that interaction. We consider two probabilistic approaches to combine source reliabilities, where interactions are *events* and information sources are *experts*. Both assume independence of experts which we address by separating information by experiment type. The first is a noisy-OR model (NoisyOR), used by several groups[11,12], which interprets r_E as the probability[f] of interaction according to expert E and calculates the consensus *un*reliability of the experts: $Pr(e) = 1 - \prod_E (1 - r_E)$, with the product over experts E contributing edge e. The second is a well-known result in statistics for computing consensus likelihoods from a collection of experts, the *Linear Opinion Pool* (LinOP)[13]: $Pr(e) = \sum_E \alpha_E Pr_E(e)$. The α_E are nonnegative expert weights that sum to 1. Our r_E correspond roughly to α_i, since $Pr_E(e) = 1$ for each edge offered by expert E and no

[e] Assessing reliability using consensus has precedence in medical decision making[31].
[f] We use $r'_E = \frac{r_E}{(\max_E\{r_E\}+1)}$ to create a valid probability value.

information otherwise; we must renormalize over the applicable experts per edge. To our knowledge, this is the first formal identification of LinOP from the statistics community for use in biological problems.

3. Application to Biological Tasks

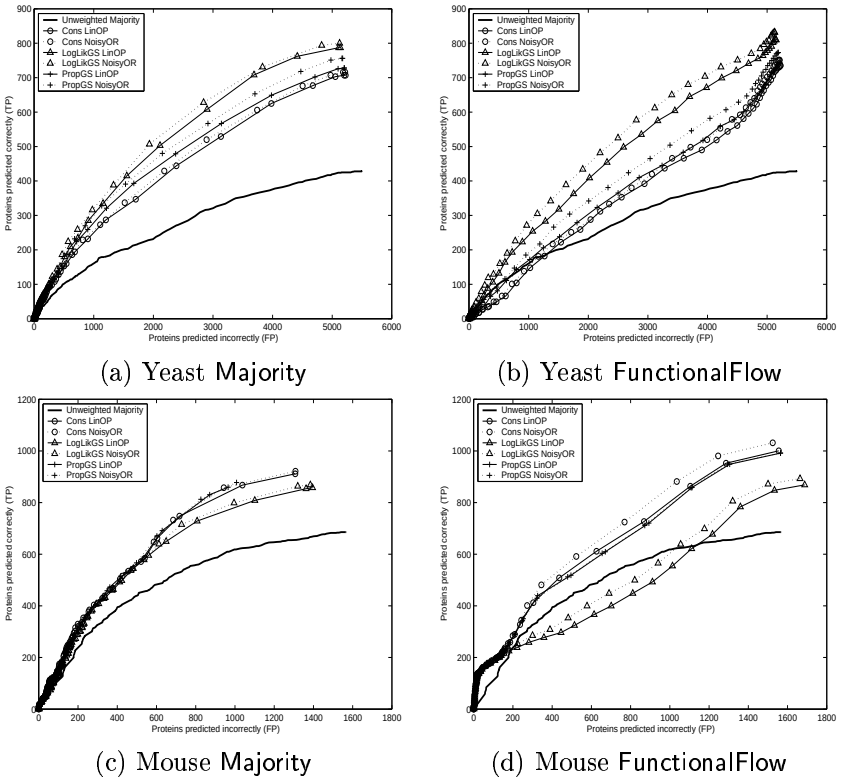
We use PropGS, LogLikGS and Cons to assign reliabilities to interaction sources, then use LinOP and NoisyOR with each assignment to obtain a confidence measure $Pr(e)$ for each interaction e. We compare the various weighting strategies on two tasks which can make use of weighted interactions: inferring protein function and learning regulatory networks from gene expression data. In each task, we evaluate 1) whether incorporating reliability of interaction sources helps, 2) which reliability assignment method is better, and 3) which reliability combination method is better.

3.1. *Protein Function Prediction*

Protein function prediction methods include machine learning[32] and graph theoretic methods[12]. Since most machine learning methods do not use pairwise information, we only consider graph-theoretic approaches. The popular Majority[1] algorithm is based on 'guilt by association' whereby an unknown protein is assigned the (weighted) majority function of its neighbors in a protein network. As a baseline, Unweighted refers to the use of uniform edge weights while Weighted refers to the use of $Pr(e)$ as edge weights. One disadvantage of Majority is revealed when a node is connected to many proteins with unknown function. The FunctionalFlow[12] algorithm overcomes this difficulty by considering a larger neighborhood around a node.

For each of the yeast and mouse genomes, we obtain a set of reliability assignments r_E applying PropGS, LogLikGS and Cons to the available set of interaction types, excluding implicit types based on function or pathway assignments, such as GO:FUNCTION. We use the MIPS Functional Catalog[23] for yeast and KEGG Pathways[29] for mouse as gold standards in PropGS and LogLikGS, as well as gold standards for evaluating performance (described below). For yeast, we find an extremely high correlation between the set of reliabilities assigned by PropGS and LogLikGS ($r = 0.97$), with slightly lower correlation to Cons at $r = 0.870$ and $r = 0.874$, respectively. For mouse, the corresponding values were $r = 0.75$, $r = 0.38$, and $r = 0.64$.

Given a set r_E, we obtain consensus interaction probabilities $Pr(e)$ using LinOP and NoisyOR for use as edge weights in the function prediction



(a) Yeast Majority (b) Yeast FunctionalFlow

(c) Mouse Majority (d) Mouse FunctionalFlow

Figure 1. ROC analysis on a Fixed Topology of 126k edges

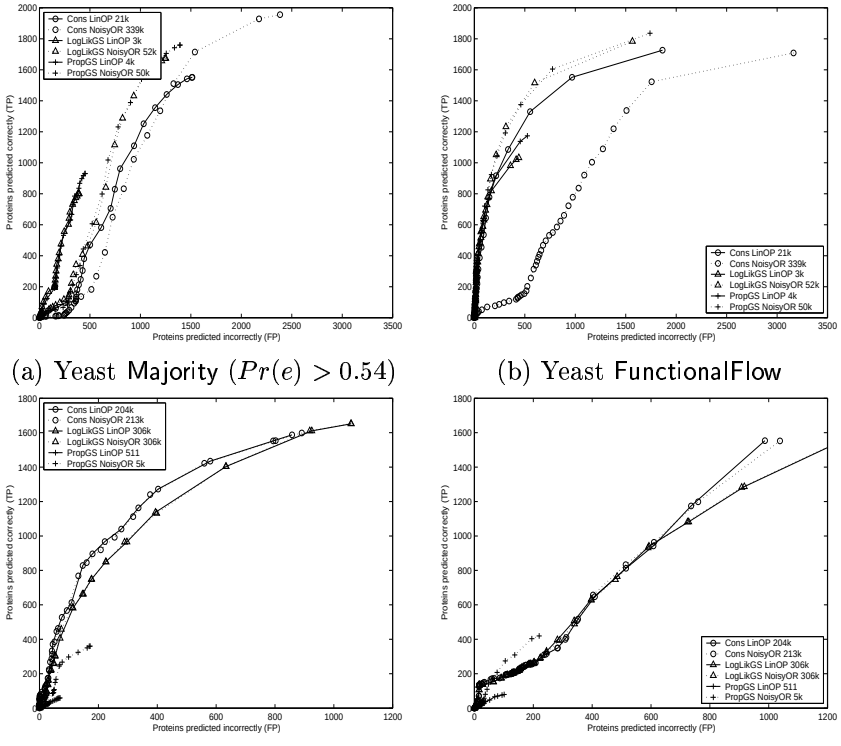
algorithms Weighted Majority and FunctionalFlow. We use two-fold cross-validation in which functions (pathways) for half of the proteins in the graph are hidden and then predicted from function (pathway) assignments to the other half. Correctness of multiple function (pathway) predictions is decided by majority. As done in Nabieva *et al.*[12], we calculate a modified ROC curve, showing the number of incorrect predictions (FP) versus number of correct predictions (TP) as the prediction score threshold varies.

Figure 1 considers relative performance of each $Pr(e)$ assignment to edges for a fixed topology, allowing us to evaluate weighting strategies on the same set of edges. Figure 1(a)-(b) show results in yeast using the prediction algorithms Majority and FunctionalFlow, respectively, while Figure 1(c)-(d) show the equivalent in mouse. A fixed topology in each organism is generated by choosing 126k edges at random among those supported by more than one interaction expert. We choose 126k to make the graph size

Pacific Symposium on Biocomputing 12:433-444(2007)

tractable for multiple runs while requiring support from multiple experts counteracts the effect of promiscuous experts, like COESSENTIAL which asserts $> 50\%$ of the total possible edges in yeast.

To answer the first question of the value of using weights based on relative reliability of experts, we compare the baseline curves Unweighted Majority in Figure 1 to the weighted alternatives. For most of the range of FP, Unweighted Majority identifies fewer TP than the weighted methods. The exceptions are the graphs for FunctionalFlow where PropGS LinOP and the Cons variants in yeast (Fig. 1(a)) and the LogLikGS variants in mouse (Fig. 1(d)) sometimes perform worse than Unweighted Majority. This can occur when a larger neighborhood involves related but not identical functions, in which case, the actual value of the edge weight determines the quality of the weighting strategies – the topic of our second evaluation question.



(a) Yeast Majority ($Pr(e) > 0.54$) (b) Yeast FunctionalFlow

(c) Mouse Majority ($Pr(e) > 0.22$) (d) Mouse FunctionalFlow

Figure 2. ROC analysis using a Fixed Probability Threshold

The LogLikGS reliability assignments show the best performance in yeast, yet the worst performance in mouse. Since LogLikGS is similar to PropGS corrected for background linkage distributions, their relative performance suggests this may be due to different background distributions ($\frac{Pr(L)}{\neg Pr(L)} = 0.15$ in yeast versus 0.01 in mouse). In fact, we found that the numerical edge weights were nearly identical for both methods in yeast, while in mouse, LogLikGS edge weights were generally twice the value of PropGS weights. Also, the yeast graph has a maximum of 77 neighbors while mouse has a maximum of 348, an enormous difference in size of neighborhoods which, together with a difference in weightings, allows FunctionalFlow to propagate a lot more noisy predictions. The Cons variants perform the best overall in mouse, suggesting better overlap of information from sources in mouse compared to those in yeast, even though mouse has fewer sources. Even in yeast, the Cons results, which do not use a function/pathway–based gold standard, are comparable to PropGS which does. In fact, these results suggest capturing a more diverse notion of 'interaction' using Cons still proves successful for the task of function prediction. Together, these results suggest Cons is a valuable alternative to LogLikGS and PropGS in less-studied organisms, where including diverse types of interaction information is critical.

For the third question of whether to use NoisyOR or LinOP to combine source reliabilities, the NoisyOR variants invariably have slightly higher performance than LinOP. For a given interaction, the value assigned by NoisyOR will be greater than by LinOP given the same set of reliability assignments to sources. In this task, this bias causes NoisyOR to make the same prediction as LinOP but at a higher threshold, accounting for the slight vertical shift between the two curves. The effect of this shift in distribution is the subject of the next figure.

The effect of different edge distributions $Pr(e)$ can be seen by fixing a probability threshold and allowing only edges which exceed the threshold. Results using $Pr(e) > 0.54$ in yeast and $Pr(e) > 0.22$ in mouse are shown in Figure 2 (legends indicate graph size per method). Shorter curves mean fewer predictions were made, a comment on the connectivity. As noted above, the LinOP variants will include fewer edges than NoisyOR for a given threshold, though here we see little performance difference between the two for all methods, except Cons (Fig. 2). This difference arises due to the large size of Cons NoisyOR (339k edges) versus the others (mean 26k) in combination with the neighborhood-based FunctionalFlow; for sparse graphs the immediate neighborhood is equivalent to the extended neighborhood,

Pacific Symposium on Biocomputing 12:433-444(2007)

making FunctionalFlow nearly equivalent to Majority. In mouse, LinOP and NoisyOR yield similar graph sizes so we do not see this effect repeated. Again, Cons performs strongly in mouse, suggesting this non gold standard-based approach will be valuable in less well-studied organisms.

3.2. *Learning Regulatory Networks*

Bayesian networks (BN) are a popular modelling formalism for learning regulatory networks from gene expression data (see Pe'er *et. al* [33] for an excellent example). A BN has two components: a directed acyclic graph (DAG) capturing dependencies between variables, and a set of conditional probability distributions (CPDs) local to each node. Nodes represent expression values, arcs represent potential regulatory relationships, and the CPDs quantify those relationships.

Algorithms to learn BNs from data can use prior knowledge about the probability of arcs, such as our $Pr(e)$. Learning performs an iterative search starting from an initial graph, exploring the space of DAGs by removing, adding or deleting a single edge, choosing the best scoring model among these one-arc changes, and terminating when no further improvement in score can be made. Each candidate model is scored with respect to the *log-likelihood* (LL) of the data, e.g. how well the CPDs capture dependencies inherent in the expression data.

To evaluate the quality of a search, we obtain a single performance measure as follows. Given a starting model, we obtain a LL-trace of the best model chosen at each iteration and average the trace over all iterations. We repeat this process for a set of starting models sampled from some distribution, and average the average LL-trace over all models. Starting models are sampled either from an *informed* structural prior (our $Pr(e)$), or an *uninformed* prior which asserts uniform probability over edges. A high average LL trace value for a given prior indicates that searches using that prior consistently explore high-scoring models.

Using the yeast genome, as before we create informed structural priors $Pr(e)$ using all interaction sources (including functional/pathway sources) together with the Cons, PropGS and LogLikGS methods to assign reliabilities (again, KEGG is the gold standard for the latter two) and the LinOP and NoisyOR methods to combine reliabilities. We learn Bayesian networks for 50 genes using a expression dataset covering 1783 yeast microarray experiments (see refs. in Tanay *et al.*[34]). We also create priors using edge reliabilities calculated by other groups, namely STRING[11] (a PropGS NoisyOR (on

Pacific Symposium on Biocomputing 12:433-444(2007)

experts different than ours) for predicting protein complexes) and MAGIC[35] (a hand-crafted BN for predicting function). Both use expression data as experts. As baselines, we include a uniform reliability assignment over experts (Unif5) and two random reliability assignments (Rand1 and Rand2). Figure 3 shows the LL trace averages, scaled to give Uninformed the value $x = 0$.

The worst overall performance by Uninformed demonstrates the value of using priors based on weighted reliabilities. The poor performance of the remaining baseline variants demonstrates the effect of neglecting to assign (Unif) or incorrectly assigning (Rand) reliability to interaction sources. Note NoisyOR performs worse than LinOP for the baseline priors, yet performs better for the non-baseline variants. This repeats the effect seen in the function prediction task where NoisyOR assigns higher values than LinOP. Here, the performance difference indicates that LinOP is more robust to errors in reliability assignment than NoisyOR. The strength of STRING, LogLikGS and MAGIC is due in part to having few high probabilities and many low probabilities in the corresponding $Pr(e)$, in contrast with the more evenly distributed $Pr(e)$ for the other methods. Such conservativism allows the Bayesian learner to strongly preserve only the highest confidence edges while remaining flexible for the others. Performance of the Cons variants is comparable to PropGS for this task as well, demonstrating the utility of our method which does not require a gold standard.

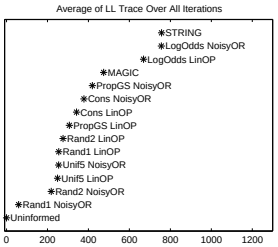


Figure 3. Average of Log-Likelihood trace over all iterations

4. Conclusions

Our results show that the Cons method for assigning reliability to interaction sources is an attractive alternative to existing methods and has the added advantage of not requiring a gold standard for assessment. In the task of predicting protein function, we demonstrated the effectiveness of using weighting strategies, where Cons proved competitive against other methods

which have an unfair advantage of using the same gold standard used for evaluation. For the task involving regulatory networks, we showed that learning greatly benefits from correctly informed estimates of reliability. Again, Cons was comparable to the other methods. We introduced LinOP as an alternative method for combining reliabilities and demonstrated its performance to be comparable to NoisyOR in most tasks and more robust to errors in others.

References

1. B. Schwikowski *et al.*, *Nature Biotech.* **18**, 1257 (2000).
2. H. Hishigaki *et al.*, *Yeast* **18**, 523 (2001).
3. A. M. Edwards *et al.*, *Trends Genet.* **18**, 529 (2002).
4. C. M. Deane *et al.*, *Mol. Cell Proteomics* **1**, 349 (2002).
5. E. Sprinzak *et al.*, *J. Mol. Biol.* **327**, 919 (2003).
6. J. S. Bader *et al.*, *Nature Biotech.* **22**, 78 (2004).
7. Y. Qi *et al.*, *NIPS Workshop on Comp. Bio. and Anal. of Het. Data* (2005).
8. S. Asthana *et al.*, *Genome Res.* **14**, 1170 (2004).
9. I. Lee *et al.*, *Science* **306**, 1555 (2004).
10. D. R. Rhodes *et al.*, *Nature Biotech.* **23**, 951 (2005).
11. C. von Mering *et al.*, *Nucl. Acids Res.* **33**, D433 (2005).
12. E. Nabieva *et al.*, *Bioinformatics* **21**, i302 (2005).
13. C. Genest and J. V. Zidek, *Statistical Science* **1**, 114 (1986).
14. S. Suthram *et al.*, *BMC Bioinformatics* **7**, 360 (2006).
15. I. Xenarios *et al.*, *Nuc. Acids Res.* **30**, 303 (2002).
16. G. Bader *et al.*, *Nuc. Acids Res.* **29**, 242 (2001).
17. H. Hermjakob *et al.*, *Nuc. Acids Res.* **32**, D452 (2004).
18. C. Stark *et al.*, *Nuc. Acids Res.* **34**, D545 (2006).
19. T. I. Lee *et al.*, *Science* **298**, 799 (2002).
20. E. Wingender *et al.*, *Nuc. Acids Res.* **28**, 316 (2000).
21. J. C. Mellor *et al.*, *Nuc. Acids Res.* **30**, 306 (2002).
22. J. T. Eppig *et al.*, *Nuc. Acids Res.* **33**, D471 (2005).
23. H. W. Mewes *et al.*, *Nuc. Acids Res.* **30**, 31 (2002).
24. N. Hulo *et al.*, *Nuc. Acids Res.* **32**, 134 (2004).
25. A. Bateman *et al.*, *Nuc. Acids Res.* **32**, D138 (2004).
26. N. J. Mulder *et al.*, *Nuc. Acids Res.* **33**, D201 (2005).
27. P. T. Spellman *et al.*, *Mol. Bio. Cell* **9**, 3273 (1998).
28. M. Ashburner *et al.*, *Nature Genet.* **25**, 25 (2000).
29. M. Kanehisa *et al.*, *Nuc. Acids Res.* **34**, D354 (2006).
30. K. D. Dahlquist *et al.*, *Nature Genet.* **31**, 19 (2002).
31. S. C. Weller and N. C. Mann, *Medical Decision Making*, **17**, 71 (1997).
32. G. R. G. Lanckriet *et al.*, *PSB* **9**, 300 (2004).
33. D. Pe'er *et al.*, *Bioinformatics* **17 Suppl.1**, S215 (2001).
34. A. Tanay *et al.*, *Molecular Systems Biology* (2005).
35. O. G. Troyanskaya *et al.*, *PNAS* **100** 8348 (2003).