BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 17 2007, pages 2322–2330
doi:10.1093/bioinformatics/btm332

Data and text mining

Context-sensitive data integration and prediction of biological networks

Chad L. Myers[1,2] and Olga G. Troyanskaya[1,2,*]

[1]Department of Computer Science, Princeton University, 35 Olden Street and [2]Lewis-Sigler Institute for Integrative Genomics, Princeton University, Carl Icahn Laboratory, Princeton, NJ, USA

Received on April 10, 2007; revised on June 1, 2007; accepted on June 18, 2007

Advance Access publication June 28, 2007

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Several recent methods have addressed the problem of heterogeneous data integration and network prediction by modeling the noise inherent in high-throughput genomic datasets, which can dramatically improve specificity and sensitivity and allow the robust integration of datasets with heterogeneous properties.

However, experimental technologies capture different biological processes with varying degrees of success, and thus, each source of genomic data can vary in relevance depending on the biological process one is interested in predicting. Accounting for this variation can significantly improve network prediction, but to our knowledge, no previous approaches have explicitly leveraged this critical information about biological context.

Results: We confirm the presence of context-dependent variation in functional genomic data and propose a Bayesian approach for context-sensitive integration and query-based recovery of biological process-specific networks. By applying this method to Saccharomyces cerevisiae, we demonstrate that leveraging contextual information can significantly improve the precision of network predictions, including assignment for uncharacterized genes. We expect that this general context-sensitive approach can be applied to other organisms and prediction scenarios.

Availability: A software implementation of our approach is available on request from the authors.

Contact: ogt@genomics.princeton.edu

Supplementary information: Supplementary data are available at http://avis.princeton.edu/contextPIXIE/

1 INTRODUCTION

Recent developments in biological technology have fueled the generation of numerous large genomic and proteomic datasets for several organisms. These data capture a wide range of biological phenomena including gene expression, genetic interactions, physical interactions between proteins and sequence content. Many recent studies have shown that high-throughput data are often quite noisy and have varying degrees of reliability or relevance for understanding biological networks (Bader *et al*., 2004; Deng *et al*., 2003; Sprinzak *et al*., 2003). To address this heterogeneity and harness the wealth of information present in the data, several groups have designed methods for data integration to combine information from multiple sources of genomic or proteomic evidence in order to arrive at accurate and holistic network and gene predictions. For instance, Troyanskaya *et al*. used expert-based Bayesian networks for inferring functional interactions between pairs of proteins given observed experimental data supporting those interactions (Troyanskaya *et al*., 2003). Other studies have extended this idea by applying more sophisticated Bayesian approaches and other methods, most of which automatically learn reliability characteristics from the data given a trusted gold standard (Jaimovich *et al*., 2005; Jansen *et al*., 2003; Lee *et al*., 2004; Qi *et al*., 2005; von Mering *et al*., 2003). In general, all of these methods assess the reliability of input high-throughput genomic data and use these characteristics for more robust integration, which typically offers significant improvement in terms of both sensitivity and specificity in predicting protein–protein interactions or functional relationships.

While these earlier approaches to data integration address the heterogeneity in reliability among different datasets, they all fail to utilize one important source of variation: biological context. Most experiments are designed with a particular process or pathway in mind. For instance, a researcher studying meiosis in yeast might profile gene expression under specific conditions (e.g. in sporulation media) that result in a clear meiotic signal in the data but very little reliable information about the mitotic cell cycle. Furthermore, most experimental technologies target specific biological processes simply because of how they physically measure biological phenomena. Yeast two-hybrid technology for identifying interacting proteins, for example, relies on the two-domain structure of eukaryotic transcription factors to report an interaction. A two-hybrid positive interaction is obtained by fusing one protein to a DNA-binding domain (bait) while another protein is fused to an activation domain such that binding of the two proteins of interest 'switches on' transcription of a reporter gene (Phizicky and Fields, 1995). Thus, while two-hybrid results are generally informative for proteins which can be targeted to the nucleus, we should expect very little reliable information about membrane proteins or proteins with domains that prevent them from entering the nucleus. In fact, if an interaction including such a protein is reported, we should confidently reject it as a false positive.

*To whom correspondence should be addressed.

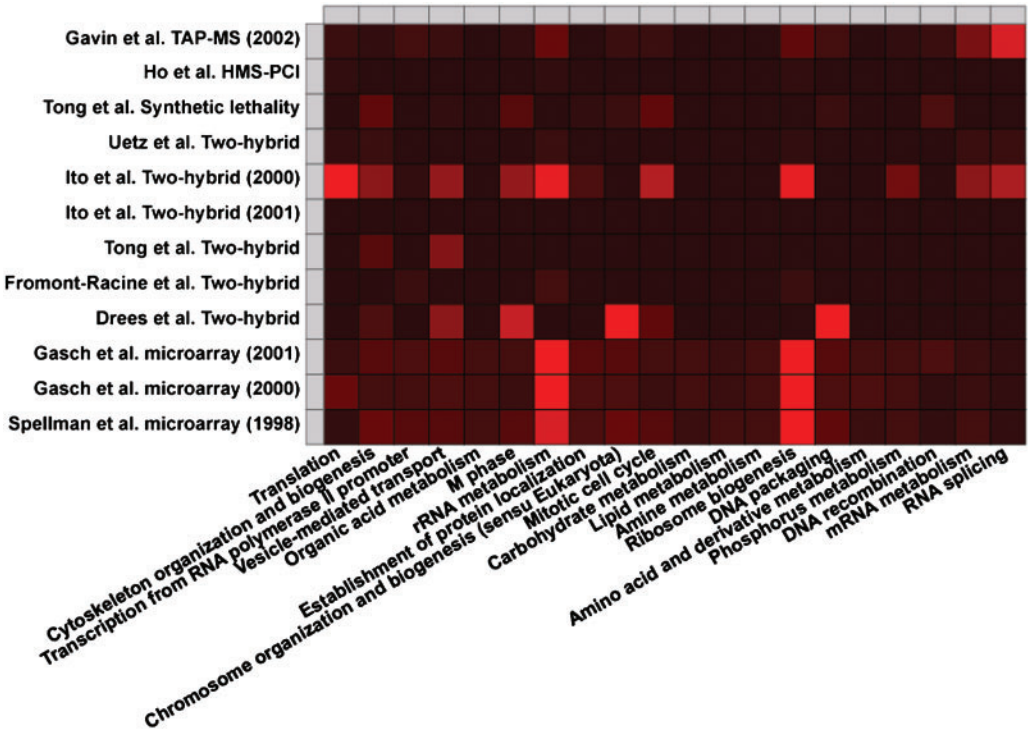


Fig. 1. Dataset relevance across different biological contexts. We measured the relevance of several *S.cerevisiae* genomic datasets for predicting function in a range of biological contexts (GO terms) using our previously published evaluation framework (Myers *et al.*, 2006). A selection of the datasets used in our integration appear on the rows, and contexts appear on the columns. The intensity of each square reflects the area under a precision-recall curve (AUPRC) for each dataset in the corresponding context. The relevance of each dataset varies substantially both in terms of precision and sensitivity across biological processes, and thus the relative weighting of data during integration depends critically on the context. For example, if one were interested in predicting proteins involved in ribosome biogenesis, any of the three gene expression datasets would be informative. If one were interested in chromosome organization, these data might offer little reliable information as compared to one of the two-hybrid datasets (e.g. Drees *et al.*, 2001).

We have explicitly measured context-dependent variation for a wide variety of public, genomic data for *Saccharomyces cerevisiae* (baker's yeast), including a large number of microarray datasets, protein–protein interaction data and sequence data. Specifically, for each source of functional genomic data we measured precision-recall characteristics for a set of experimentally relevant Gene Ontology (GO) terms covering a broad range of biological processes (Myers *et al.*, 2006) (Fig. 1). This analysis demonstrates that most datasets have a broad range of precision-recall characteristics depending on which processes they are compared against. More importantly, we find that the relative ordering of genomic data in terms of quality varies dramatically from process to process, suggesting the degree to which we should trust any dataset depends on the process we are interested in predicting.

While this context-dependent variation is not surprising given the inherent bias of different experimental techniques toward particular processes and different goals and conditions under which the data was measured, to our knowledge, no previous computational approaches for heterogeneous data integration or network prediction have explicitly leveraged this information. We demonstrate here that incorporating information about biological context in the integration and prediction process can significantly boost precision and sensitivity. We develop a system for predicting process-specific networks from diverse genomic data that uses biological context information to improve the recovery of known networks from integrated experimental data. We compare our contextual approach to our earlier work, which uses prior knowledge of gene function as a gold standard, but does not specifically leverage biological context (Myers *et al.*, 2005), and demonstrate that considering context can yield a dramatic benefit. While we illustrate the effect of biological context for a specific method for network prediction here, we demonstrate that such context-specificity has a dramatic effect on dataset reliability and thus we expect that the general idea can be used to improve predictions in a variety of settings and for many organisms.

2 METHODS

The objective of our approach is, given a diverse set of genomic data, to recover a process-specific network starting from a small related set of query proteins. Such algorithms have proven to be practical approaches for expert-driven search of genomic data, largely because they harness

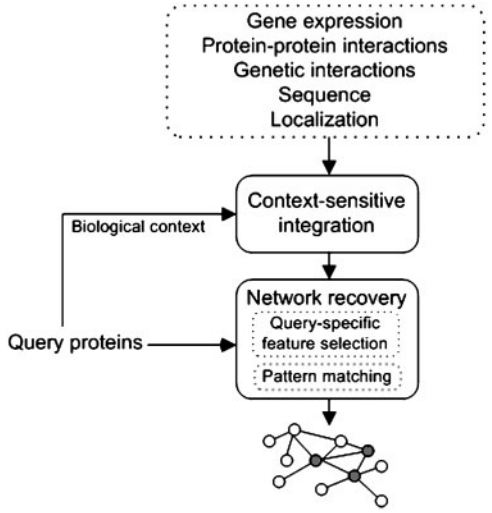


Fig. 2. Overview of method for context-sensitive integration and prediction. Our approach is developed for the scenario where a user enters a query set of proteins and wishes to obtain a relevant network prediction based on a diverse set of experimental evidence. The method consists of two stages, the first a Bayesian network for data integration and the second a network recovery algorithm which uses the probabilistic network from the first stage to recover the network surrounding the entered query. The biological context of a prediction is inferred from the entered query set, and this information is fed into both stages to improve prediction precision.

information from all available evidence in a robust way while also providing an intelligent interface for discovering functional modules and extracting the relevant portion of the interaction network (Myers *et al*., 2005). This general approach of incorporating expert direction in the prediction process is particularly attractive because it offers a convenient method of learning the biological context and leveraging this information to arrive at more precise predictions. Our solution based on this premise can be divided into two distinct components: a data integration phase that forms a probabilistic protein–protein network as supported by experimental data, and a network search algorithm that, given the probabilistic network, recovers additional relevant proteins starting from a query set (Fig. 2). Both phases of the network prediction process utilize information about biological context, which is inferred from the starting query set.

2.1 Bayesian context-specific integration

The integration phase consists of a Bayesian network, which captures the context-dependent reliability variation to integrate the diverse input data. The result of this phase is a probabilistic protein–protein interaction network reflecting the reliability of the supporting data in a given biological context. The input data used here and the details of the Bayesian network are described below.

2.1.1 Genomic input data We have collected genomic data for *S.cerevisiae* from over 6500 publications, including gene expression, literature-curated and high-throughput protein-protein and genetic interactions (Alfarano *et al*., 2005; Stark *et al*., 2006), protein localization data (Huh *et al*., 2003), transcription factor binding site data (Harbison *et al*., 2004; Zhu and Zhang, 1999) and sequence data (SGD, 2006). See the Supplementary Material for a detailed description of how each data type was processed. The processed input data was separated first by experimental method responsible for producing the data, then by publication. To ensure that each input dataset had a reasonable number of observations for learning, publications with fewer than 50 observations were merged with other publications reporting results from the same experimental method. This process resulted in 174 different input data types for Bayesian integration.

2.1.2 Bayesian network The goal of our integration scheme is to harness the information from the diverse data while not sacrificing precision. Furthermore, the integration is designed such that it can model and exploit the context-dependent relevance variation discussed earlier. Because many of the input data types represent functional interactions (either physical or other) between pairs of genes or proteins, we have adopted the approach of predicting functional associations. This approach has been used in several earlier studies (Jaimovich *et al*., 2005; Jansen *et al*., 2003; Lee *et al*., 2004; von Mering *et al*., 2003), and the final integrated protein–protein linkage network is convenient for understanding and predicting network structure, which is our goal here. Several methods for associating proteins directly with processes or functional classes (function prediction) have also been applied successfully (Barutcuoglu *et al*., 2006; Lanckriet *et al*., 2004; Letovsky and Kasif, 2003), but are less appropriate for the goal of network analysis and prediction.

Starting with the goal of predicting functional associations between genes, there are several choices of machine-learning methods that might be appropriate. Here, we employ a Bayesian network because it is robust to diverse forms of input data, and it yields a generative model that is useful in terms of drawing relevant biological conclusions about the properties of the input data. Furthermore, a Bayesian framework is a convenient setting for incorporating contextual information as is illustrated below.

The simplest Bayesian approach for integration is to assume independence between all of the input datasets given knowledge of a functional relationship between any pair of proteins. In practice, this approach is quite powerful for genomic data and is competitive with more sophisticated alternatives, including methods where dependence among datasets is modeled [e.g. tree-augmented Bayesian networks (Friedman *et al*., 1997), see Supplementary Material]. We begin with the naive approach and extend it to include contextual information as illustrated in Figure 3. Each input dataset is modeled with a discrete probability distribution conditioned on the presence or absence of a functional relationship *and* the biological context. Given a gold standard which associates observed data with known functional relationships and biological context (described in detail in the following section), we estimate the conditional distribution for each input dataset by simple counting. With these learned parameters, given a new protein–protein pair with observed data and a corresponding context (derived from the query as described below), we can then infer the probability of functional relationship between the two proteins, i.e.

$$P\left(\mathrm{FR}_{ij}\middle|D_{ij}^{1},D_{ij}^{2},\ldots,D_{ij}^{k},C_{ij}\right)=\alpha P\left(\mathrm{FR}_{ij}\middle|C_{ij}\right)\prod_{n=1}^{k}P\left(D_{ij}^{n}\middle|\mathrm{FR}_{ij},C_{ij}\right)$$

where

$$P\left(D_{ij}^{n}=d\middle|\mathrm{FR}_{ij}=f,C_{ij}=c\right)=\frac{\#\left(D_{ij}^{n}=d\wedge\mathrm{FR}_{ij}=f\wedge C_{ij}=c\right)}{\#\left(\mathrm{FR}_{ij}=f\wedge C_{ij}=c\right)}.$$

Here FR_{ij} refers to the presence or absence of a functional relationship between proteins i and j, D_{ij}^{n}refers to the observed association in dataset n between the proteins i and j, C_{ij} is the biological context of the pair and α is a normalization constant.

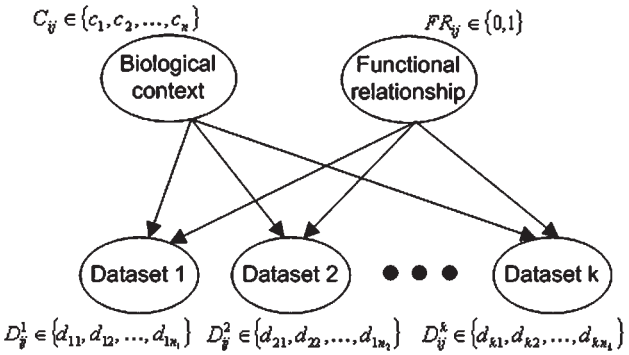


Fig. 3. Bayesian network for context-sensitive integration. The data integration stage of our context-sensitive approach consists of a Bayesian network, which is used to integrate pairwise protein–protein association data to arrive at a single, probabilistic network. Biological context information is incorporated into the integration process by conditioning the probability distributions of each type of observed genomic data on both the presence or absence of a functional relationship between the pair of proteins in question and the biological context of interest. This structure captures both the inherent dataset quality as well as the relevance variation from one biological process to another. Evidence nodes are assumed to be discrete, and conditional probability tables (CPT's) are automatically learned from the data using a gold standard based on the biological process branch of the GO.

2.1.3 Gold standard for Bayesian integration The gold standard used in estimating the parameters for the Bayes net is a critical part of the prediction process. The gold standard used here is based on the biological process branch of the Gene Ontology (Ashburner *et al.*, 2000) as proposed in Myers *et al.* (2006). For the global (non-context-sensitive) approach described here, we directly used the protein–protein pairwise standard for functional relationships published in Myers *et al.* as our global (non-context-sensitive) standard for functional relationship. For the context-sensitive approach, we require a gold standard that associates positive and negative examples of functionally related pairs of proteins to a set of biological contexts. For this, we used the non-redundant set of specific GO terms published in Myers *et al.* (2006), which is a set of terms spanning the entire process ontology at a specificity sufficient for inferring useful functional information as curated by biology researchers. Specifically, we chose the 101 largest of these terms (those with more than 20 annotations), as the space of all possible contexts $(c_1, \ldots, c_n)$. Positive examples for each context were derived by forming all possible pairs of proteins annotated to the corresponding term. Negatives were sampled from the negative gold standard described in Myers *et al.* (2006) as discussed in detail in the Supplementary Material. During the inference process, context is inferred from the entered query proteins by mapping to the term in this comprehensive set containing the maximum number of proteins in the query.

2.2 Context-sensitive network recovery algorithm

The problem of recovering a network from a starting query set given a probabilistic interaction graph of proteins has been addressed in previous work (Asthana et al., 2004; Bader, 2003; Can et al., 2005; Myers et al., 2005). Approaches to this problem range from random walks on the probabilistic network (Can et al., 2005), to methods based in network reliability theory (Asthana et al., 2004), to variations of maximum adjacency (Bader, 2003; Myers et al., 2005). We find that the performance of such methods often depends on the sparsity of the starting network, and it is difficult to find one that always provides superior performance. We describe an approach here that performs favorably on our probabilistic network, but emphasize that the larger point of incorporating biological context is independent of the specific network recovery algorithm used. Our network recovery algorithm consists of two steps: (1) a feature selection step that, given a query set of genes, determines a 'characteristic' interaction profile for that group, and (2) a pattern-matching step that finds additional proteins matching the characteristic profile.

2.2.1 Feature selection Let Q be the query set of proteins of size N_Q chosen out of the entire proteome consisting of N_T proteins, and let $p_{ij} = P\left(\mathrm{FR}_{ij} | D_{ij}^1, D_{ij}^2, \ldots, D_{ij}^k, C_{ij}\right)$ be the probability of functional relationship between proteins i and j in the current biological context. Our goal is to select a set of features which are predictive of proteins related to the query set. Here, we treat each protein's interaction probabilities as a set of features, and thus feature selection is equivalent to finding a set of interaction partners which are common and discriminative of the query set. For each possible feature, k, we compute:

$$N_{Q,k}(t) = \left|\left\{j \in Q : p_{ij} > t\right\}\right|$$
$$N_{T,k}(t) = \left|\left\{j : p_{ij} > t\right\}\right|$$

where t is a threshold on the interaction probabilities. We can then assign a P-value measuring the significance of the association between feature k and the query set using the hypergeometric distribution, i.e.

$$f_k(t) = 1 - \sum_{n=0}^{N_{Q,i}(t)} \frac{\binom{N_Q}{n}\binom{N_T - N_Q}{N_{T,i}(t) - n}}{\binom{N_T}{N_{T,i}(t)}}$$

For each feature, we compute this P-value over a range of interaction probability thresholds and select the minimum. The selected features are then given by $F = \left\{k \in \{1,2,\ldots,N_T\} : \min_t f_k(t) < 0.05\right\}$.

2.2.2 Pattern matching During the pattern-matching phase, we identify remaining genes whose interaction profiles match the characteristic profile determined during the feature selection phase. Given the query set, Q and selected features, we add proteins to the predicted network based on their similarity to the query proteins over the set of relevant features, F. Specifically, each candidate protein, i, is ranked according to the following adjacency score:

$$S_i = \sum_{j \in Q} \sum_{k \in F} p_{ik}\ p_{jk}$$

This metric ensures that only relevant features are used in predicting the final network, and each relevant feature (protein interaction) is weighted by our confidence in that particular interaction. Intuitively, this two-step approach of graph feature selection and pattern-matching identifies a set of informative neighbors in the interaction network and ranks candidate proteins by measuring adjacency to the query set on paths *through* these informative neighbors.

3 RESULTS

We demonstrate the importance of considering biological context for predicting biological networks by comparing our contextual approach with a simpler version that does not use information about biological context. Specifically, we replaced the context-sensitive Bayesian network illustrated in Figure 3 with a simple, naive structure with no context node. For all experiments described here, both approaches start with a query set of proteins and use the same network recovery procedure,

such that the only difference between the two is the presence or absence of contextual information during data integration.

We compared the simpler version of our method (with no contextual information) to existing approaches for network recovery (Asthana et al., 2004; Bader, 2003) in our previous publication (Myers et al., 2005). In summary, the non-contextual version of our method outperforms existing approaches for network recovery in terms of both precision and recall on a wide range of biological processes, complexes and pathways. The details of this comparison are summarized in the Supplementary Material. Evaluation results presented here illustrate further improvement offered by incorporating context information during integration and network recovery.

3.1 Contextual network recovery evaluation

Perhaps the most important question to address with evaluation experiments is: does incorporating biological context information improve network prediction? To answer this question, we performed cross-validation experiments on *S.cerevisiae* data for both our context-sensitive approach and the simpler non-contextual Bayesian integration and search algorithm. Specifically, for each of the GO terms in the evaluation gold standard (Myers *et al*. 2006), we withheld one-half of the annotated proteins for network recovery evaluation. The other half was used in training both Bayesian network configurations (with and without context nodes). Positive and negative examples (protein pairs) for the non-contextual configuration were derived as described in Myers *et al*. (2006). For the context-specific case, we obtained positive protein pairs for each context by considering all pairs between proteins annotated to the corresponding GO terms, except those selected in the corresponding cross-validation fold, as positive examples. To maintain the same ratio of positives to negatives, negative examples were sampled from the negatives described in Myers *et al*. (2006). Details on the training example selection are discussed in the Supplementary Material.

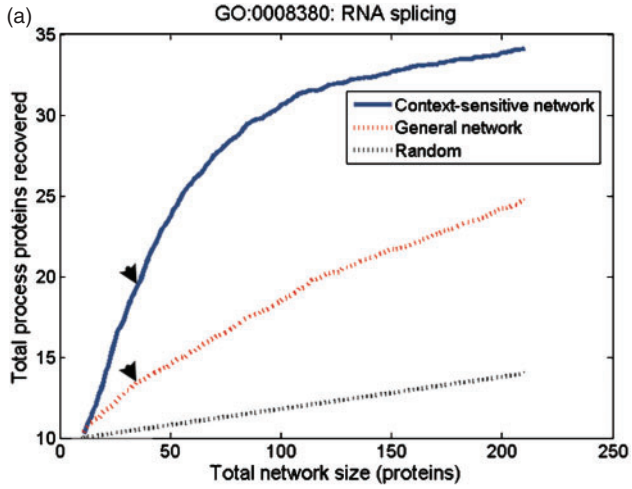
On the proteins held out in each cross-validation fold, query sets of 10 proteins each were randomly sampled from each GO term, and we attempted to recover the remaining proteins with both the context-sensitive and general approaches. All results presented here are averaged over 20 random query set samplings and two folds of cross-validation.

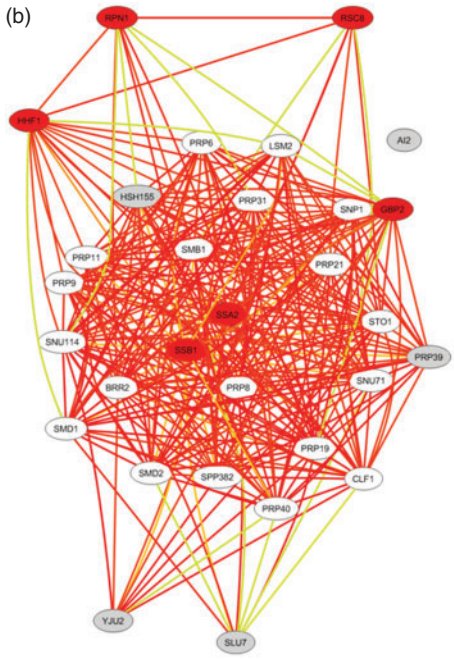
We start by considering network recovery results for the RNA splicing context. Our context-sensitive integration and recovery dramatically improves both the precision and sensitivity of network recovery for RNA splicing proteins (Fig. 4). For example, starting with 10 randomly chosen RNA splicing proteins, the context-sensitive approach recovers an average of 25 proteins correctly in the first 50 predictions, while the global approach only recovers 15 proteins. Figure 4B and C illustrates the results of the same 5-protein query for both methods at the indicated point on the recovery performance curves. For this particular query, the context-sensitive prediction reports only 6 false positives resulting in 80% precision while the global network reports 22 false positives resulting in 27% precision. Both approaches are substantially better than random in terms of predictive power, but the contextual information clearly offers an improvement.

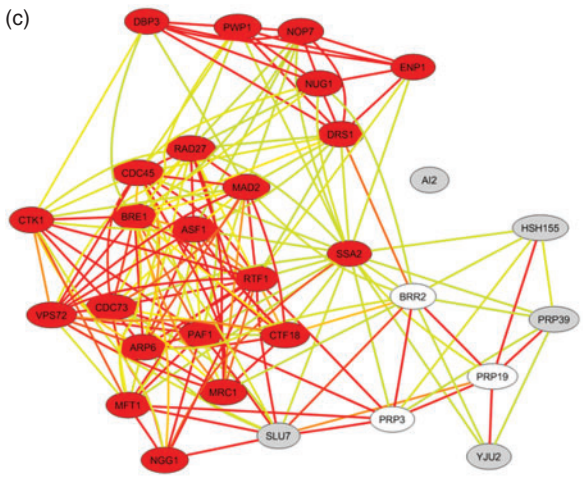
This improvement gained by using contextual information is consistent over a broad range of biological processes. We performed a similar evaluation to that described above for RNA splicing for 101 total GO terms from the evaluation set (Myers et al., 2006). The results of this evaluation for a range of predicted network sizes are summarized in Supplementary Table 1. As each approach added proteins to the predicted network, we measured the number of predicted, held-out true positives and averaged these estimates over several randomly sampled query sets. At each network size increment, we compared the average number of recovered true positive proteins for the context-sensitive versus global approaches and summarized the improvement over the set of evaluation GO terms for which both methods recovered at least 2 true positives (53 out of the 101 evaluation terms). For example, for networks of 40 recovered proteins (from a query of 10 proteins), the context-sensitive approach improved 51% of the GO terms by more than 2 SDs (estimated from random query samplings). Conversely, the context-sensitive approach resulted in a deterioration of the performance by more than 2 SDs on only 8% of the GO terms. The average improvement in the number of true positives recovered across all terms for size 40 networks is 46%. This comparison is summarized in Figure 5. The improvement offered by context-sensitive integration and prediction is consistent across a range of network sizes (see Table 1 in the Supplementary Material for a complete performance comparison).

3.2 Comparing dataset relevance across contexts

After confirming superior performance of the context-sensitive approach for a variety of biological processes, we investigated reasons for this improvement. The most informative aspect of our results is the learned parameters of the context-sensitive Bayesian network, which is designed to capture the relevance variation that motivated our approach. If our original observation of context-dependent relevance variation is correct, we expect to observe differences in the learned conditional probability distributions. To measure these differences, we computed $P(\mathrm{FR}|D_i, C_j)$, the posterior probability of a functional relationship given an observation from a single dataset, D_i, across a range of biological contexts, C_j. To obtain a single measure reflecting the relevance of each dataset in each context, we then found the maximum posterior over all possible quantized observations for a given dataset. Comparing this posterior for several contexts to the same posterior inferred by the non-contextual Bayesian network yields insight into how dataset relevance variation is captured across different contexts. Figure 6 illustrates this comparison for 13 of the total 174 input datasets and two biological contexts: RNA splicing (GO:0008380) and Phosphorus metabolism (GO:0006793). The global network reports dataset relevance (posterior probability of FR) as inferred by the simpler Bayesian network (with no contextual information). As is demonstrated in the figure, there are several datasets for which the posterior from the global network is much larger than both contexts [e.g. ER-Golgi co-localization (Huh *et al*., 2003), Martin *et al*. (2004) microarray] suggesting these datasets are generally quite reliable but contain little information about either RNA







splicing or phosphorus metabolism. Conversely, there are some datasets that appear relatively unreliable on the global scale, but are actually quite precise when examined in a specific context. For instance, all three protein–protein interaction datasets pictured are up-weighted in the RNA splicing context, particularly the Gavin *et al*. TAP-MS (2006) interaction data, which measures a maximum posterior of 0.72 for the RNA splicing context compared to a 0.22 posterior in the simpler Bayesian network. From a biological standpoint, perhaps this is not too surprising since a large portion of the RNA splicing term is composed of the spliceosome complex, which would be readily detectable with physical binding assays. These protein–protein interaction datasets have no extra relevance for the phosphorus metabolism, but all of the microarray datasets included in Figure 6 are up-weighted in the phosphorus metabolism context, particularly the Epstein *et al*. dataset, which profiled several mitochondrial perturbations.

These differences between the global and context-specific posteriors are not limited to these 13 datasets, but occur in many of the datasets included in our integration (see Fig. S4 in Supplementary Material). Interestingly, there are a large number of datasets that have reasonably high posteriors in the global setting with near zero posteriors in the specific contexts. This suggests that many datasets either contain little or very unreliable information for these contexts. This knowledge is actually quite useful for improving predictions for a specific context, because it means we can confidently exclude a number of observations from the corresponding datasets as false positives. Generally, the chances of making a false positive prediction are high simply because there are many more negative examples (proteins) than positive for network prediction problems. Thus, any reliable means of excluding false positives is an effective strategy for improving prediction performance.

3.3 Learning new biology using contextual information

We have shown through cross-validation experiments that using contextual information can generally improve the quality of network prediction, but these results are based on held-out, known annotations for genes or proteins. An interesting

Fig. 4. RNA splicing network recovery example. We compared the ability of the context-sensitive and global approaches to recover known networks of proteins using cross-validation experiments. Specifically, we started with a set of GO terms covering a wide range of biological processes (Myers *et al*., 2006), and measured each method's ability to recover held-out proteins given 10-protein queries from the same process. As proteins are added to the predicted network, we plot the number of true positive proteins present for each method, averaged over 20 query samplings (**a**). On average, the context-sensitive approach recovers more held-out true positive proteins at better precision than the global approach. Specific examples of predicted networks from the context-sensitive and global approaches are pictured in (**b**) and (**c**) respectively (sampled from the recovery curve at the point indicated in a). Query proteins are colored gray, true positives are white and false positives are red. For this particular query, the context-sensitive approach makes 24 of 30 correct predictions (80% precision) while the global approach only makes 8 of 30 correct predictions (27% precision).

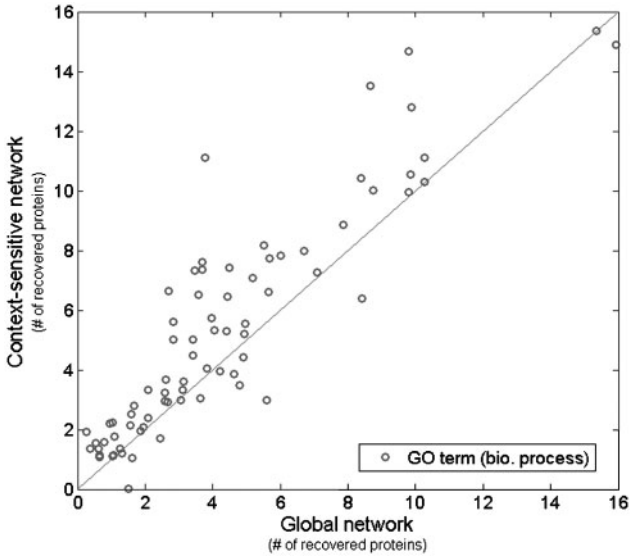


Fig. 5. Network recovery evaluation summary. We compared the ability of the context-sensitive and global approaches to recover known networks of proteins using cross-validation experiments. Specifically, we started with a set of GO terms covering a wide range of biological processes (Myers *et al*., 2006), and measured each method's ability to recover held-out member proteins given 10-protein queries from the same process. As proteins were added to each process-specific network, we measured the number of true positives recovered. Figure 5 compares the number of true positives recovered for the two different methods for networks of 40 proteins on 101 different biological processes. The context-sensitive approach improves recovery by more than 2 SD (estimated from query samplings) for 51% of the terms evaluated and only causes deterioration by more than 2 SD on 8% of the terms. This improvement is consistent across network sizes (see Supplementary Table 1 for a complete comparison).

(and perhaps more biologically relevant) question is, does such an approach help us learn new biology with greater precision? While the true answer to this question requires experimental confirmation of novel predictions, we can derive some hints from our network recovery evaluation.

To compare the ability of the context-sensitive and global approaches to confidently associate previously uncharacterized proteins in *S.cerevisiae* with portions of characterized networks, we performed a similar cross-validation experiment to that described previously. More specifically, on the proteins held out in each cross-validation fold, query sets were randomly sampled from each GO term, and we used both methods to recover the remaining network. For each protein added to the network, we estimated the precision of that particular prediction based on known, held-out proteins for the corresponding cross-validation fold. Precision estimates were smoothed across each ranked list (order in which proteins were recovered for each network), and an uncharacterized gene appearing in any prediction was assigned the corresponding precision. Uncharacterized genes were assumed to be genes annotated to the 'biological process unknown' GO term (GO:0000004) as of 1 May 2006 (Saccharomyces Genome Database, 2006). Figure 7 illustrates the results of this analysis for the two methods by plotting the measured precision (relative to random) versus the number of uncharacterized proteins assigned with *at least* that precision.

The context-sensitive network prediction approach is generally able to make more network predictions at higher confidence. For instance, at 10 times the precision expected by chance, the global scheme is able to predict networks for 118 previously uncharacterized proteins while the context-sensitive approach makes predictions for 214 uncharacterized proteins (81% improvement). Interestingly, the difference between the two approaches is smaller for very high-precision predictions (e.g. >20 fold over random), suggesting there a limited number of uncharacterized proteins whose participation in certain networks is relatively easy to detect and varies little between the two methods. As we relax the precision criteria, however, the context-sensitive approach shows a clear and consistent improvement in precisely predicting uncharacterized genes in networks recovered from known sets of related proteins.

In summary, incorporating contextual information in the data integration and prediction process can significantly improve prediction quality and provide important information about relevance of individual datasets in different contexts. As noted above, there are a very limited number of cases where the context-sensitive approach results in a loss of performance. This is typically due to the size of the GO terms corresponding to these contexts, and for such cases, global (non-context-sensitive) integration should be used (see Supplementary Material for a detailed analysis). Incorporating context into the Bayesian integration phase requires context-specific examples, which can be very few in number for smaller contexts (GO terms). Interestingly, this suggests a trade-off between the number of examples and the specificity of examples, which hints at why contextual information for network prediction is important. Put simply, the more specific we can be about the learning task, the better performance we can expect. This only holds true, however, if we can maintain a statistically representative example set, which requires a minimum number of examples. In general, this problem seems to affect a small minority of contexts evaluated here, and can be avoided by defining contexts more broadly.

We should emphasize that although we have implemented our approach using a Bayesian integration scheme and a particular search algorithm, the overall message of using contextual information is general and could be used to improve a variety of approaches to network prediction. We expect this concept to be particularly true as we begin to develop methods for integration and prediction in higher organisms, where there is not only variation in dataset relevance across biological process, but also across other aspects such as tissues or stages of development. An important consideration, however, is that to take advantage of this information, methods must be formulated in such a way that cross-context variation can actually by incorporated into the process. For instance, in our discussion here, we have assumed a query-based scheme, which inherently provides a straightforward approach to inferring the context of the prediction. Methods like this that allow expert direction are particularly well suited to leveraging contextual information to improve prediction.

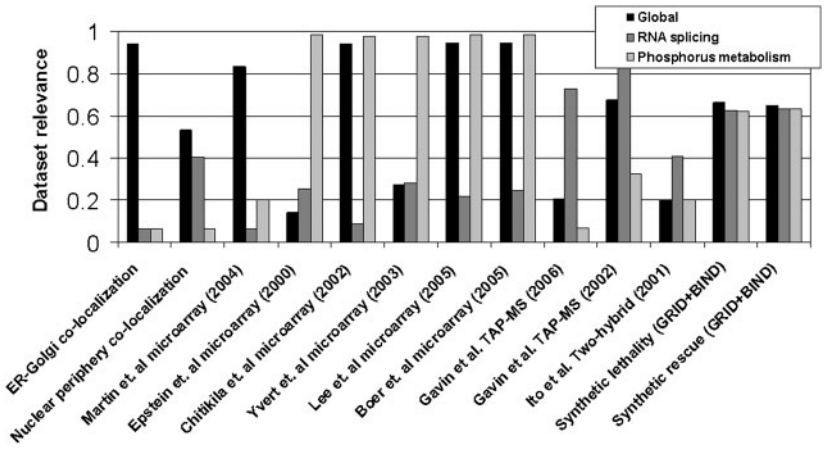


Fig. 6. Bayes net learned dataset relevance. We analyzed the learned parameters of the context-sensitive Bayesian network to understand the improvement achieved by our method. Dataset relevance was measured by computing the maximum posterior probability of functional relationship for each dataset in each context. Figure 6 compares these relevance estimates for the global integration approach to the context-specific approach for RNA splicing and phosphorus metabolism contexts on a sampling of 13 datasets integrated by our approach. Datasets that one might expect to be relevant for predicting RNA splicing proteins are up-weighted relative to the global approach in the RNA splicing context (e.g. Gavin *et al*., 2006 TAP-MS data), and likewise, datasets that are likely relevant for understanding metabolism are up-weighted in the phosphorus metabolism context8 (e.g. Epstein *et al*., 2001 which profiled mitochondrial perturbations). Modeling this variation during data integration helps to exclude false positives from irrelevant datasets that might otherwise result in poor network prediction. A more complete comparison of dataset relevance estimates can be found in the Supplementary Material Figure S6.

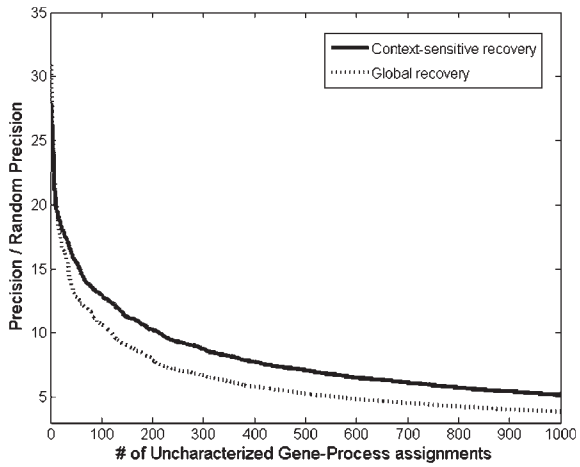


Fig. 7. Precision of network prediction for uncharacterized genes. To assess the potential of context-sensitive prediction for learning new biology in *S.cerevisiae*, we compared the ability of the context-sensitive and global approaches to predict precise networks involving uncharacterized genes. We performed cross-validation analysis as described in Section 3.1, and used held-out known proteins to assess the precision at which uncharacterized genes were predicted in networks across a range of biological processes. Figure 7 plots a range of precision measures (relative to random predictions) versus the number of uncharacterized genes recovered at that precision or higher. The context-sensitive approach tends to predict the involvement of more uncharacterized genes at higher precision than the global approach. For instance, at 10 times the precision expected by chance, the global scheme is able to predict networks for 118 previously uncharacterized proteins while the context-sensitive approach makes predictions for 214 uncharacterized proteins (81% improvement).

In conclusion, we have demonstrated evidence for context-dependent dataset reliability and illustrated a Bayesian integration and network recovery approach that makes use of this variation. Our approach achieves significant improvement in terms of both precision and sensitivity over a broad range of biological processes, and we have shown that it improves the estimated precision on predicted networks for previously uncharacterized genes. Furthermore, this approach provides information about the relevance of different data sources to specific biological processes. Biological context is an important consideration for any network prediction approach, and can be an effective means for managing data heterogeneity, particularly as we move toward developing computational methods for understanding networks in more complex organisms.

ACKNOWLEDGEMENTS

The authors would like to thank Matt Hibbs, Curtis Huttenhower, Florian Markowetz, and Edo Airoldi for insightful discussions. This research is partially supported by NSF CAREER award DBI-0546275 to OGT, NIH grant R01 GM071966, NIH grant T32 HG003284, and NIGMS Center of Excellence grant P50 GM071508. OGT is an Alfred P. Sloan Research Fellow.

Conflict of Interest: none declared.

REFERENCES

Alfarano,C. (2005) The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Res*., **33**, D418–D424.

Ashburner,M. *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet*., **25**, 25–29.

Asthana,S. *et al*. (2004) Predicting protein complex membership using probabilistic network reliability. *Genome Res*., **14**, 1170–1175.

Bader,J.S. (2003) Greedily building protein networks with confidence. *Bioinformatics*, **19**, 1869–1874.

Bader,J.S. *et al*. (2004) Gaining confidence in high-throughput protein interaction networks. *Nat. Biotechnol*., **22**, 78–85.

Barutcuoglu,Z. *et al*. (2006) Hierarchical multi-label prediction of gene function. *Bioinformatics*, **22**, 830–836.

Can,T. *et al*. (2005) Analysis of protein-protein interaction networks using random walks. *Conference on Knowledge Discovery in Data*. Chicago, IL.

Deng,M. *et al*. (2003) Assessment of the reliability of protein-protein interactions and protein function prediction. *Pac. Symp. Biocomput*., **8**, 140–151.

Drees,B.L. *et al*. (2001) A protein interaction map for cell polarity development. *J Cell Biol*, **154**, 549–571.

Epstein,C.B. *et al*. (2001) Genome-wide responses to mitochondrial dysfunction. *Mol Biol Cell*, **12**, 297–308.

Friedman,N. *et al*. (1997) Bayesian network classifiers. *Mach. Learn*., **29**, 131–163.

Gavin,A.C. *et al*. (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature*, **440**, 631–636.

Harbison,C.T. *et al*. (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.

Huh,W.K. *et al*. (2003) Global analysis of protein localization in budding yeast. *Nature*, **425**, 686–691.

Jaimovich,A. *et al*. (2005) Towards an integrated protein-protein interaction network. In *Proceedings of International Conference on Research in Computational Molecular Biology (RECOMB 2005)*, pp. 14–30.

Jansen,R. *et al*. (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, **302**, 449–453.

Lanckriet,G.R. *et al*. (2004) Kernel-based data fusion and its application to protein function prediction in yeast. *Pac. Symp. Biocomput*., **9**, 300–311.

Lee,I. *et al*. (2004) A probabilistic functional network of yeast genes. *Science*, **306**, 1555–1558.

Letovsky,S. and Kasif,S. (2003) Predicting protein function from protein/protein interaction data: a probabilistic approach. *Bioinformatics*, **19** (Suppl. 1), i197–i204.

Martin,D.E. *et al*. (2004) Rank Difference Analysis of Microarrays (RDAM), a novel approach to statistical analysis of microarray expression profiling data. *BMC Bioinformatics*, **5**, 148.

Macatee,H.R., Garner, *et al*. (2001) Genome-wide responses to mitochondrial dysfunction. *Mol Biol Cell*, **12** (2), 297–308.

Myers,C *et al*. (2006) Finding function: evaluation methods for functional genomic data. *BMC Genomics*, **7**, 187.

Myers,C.L. *et al*. (2005) Discovery of biological networks from diverse functional genomic data. *Genome Biol*., **6**, R114.

Phizicky,E.M. and Fields,S. (1995) Protein-protein interactions: methods for detection and analysis. *Microbiol. Rev*., **59**, 94–123.

Qi,Y. *et al*. (2005) Random forest similarity for protein-protein interaction prediction from multiple sources. *Pac. Symp. Biocomput*., **10**, 531–542.

Saccharomyces Genome Database Retrieved 1 May 2006, from ftp://ftp.yeastgenome.org/yeast/.

Sprinzak,E. *et al*. (2003) How reliable are experimental protein-protein interaction data? *J. Mol. Biol*., **327**, 919–923.

Stark,C. *et al*. (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res*., **34**, D535–D539.

Troyanskaya,O.G. *et al*. (2003) A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). *Proc. Natl Acad. Sci. USA*, **100**, 8348–8353.

von Mering,C. *et al*. (2003) STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res*., **31**, 258–261.

Zhu,J. and Zhang,M.Q. (1999) SCPD: a promoter database of the yeast Saccharomyces cerevisiae. *Bioinformatics*, **15**, 607–611.