BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 4 2010, pages 529–535
doi:10.1093/bioinformatics/btp701

Systems biology

Pandora, a PAthway and Network DiscOveRy Approach based on common biological evidence

Kelvin Xi Zhang[1,2] and B. F. Francis Ouellette[2,*]

[1]Graduate Program in Bioinformatics, University of British Columbia, Vancouver, British Columbia, V6T 1Z4 and
[2]Ontario Institute for Cancer Research, MaRS Centre, South Tower, 101 College Street, Toronto, Ontario,
M5G 0A3, Canada

Received on July 15, 2009; revised on November 17, 2009; accepted on December 17, 2009

Advance Access publication December 22, 2009

Associate Editor: John Quackenbush

ABSTRACT

Motivation: Many biological phenomena involve extensive interactions between many of the biological pathways present in cells. However, extraction of all the inherent biological pathways remains a major challenge in systems biology. With the advent of high-throughput functional genomic techniques, it is now possible to infer biological pathways and pathway organization in a systematic way by integrating disparate biological information.

Results: Here, we propose a novel integrated approach that uses network topology to predict biological pathways. We integrated four types of biological evidence (protein–protein interaction, genetic interaction, domain–domain interaction and semantic similarity of Gene Ontology terms) to generate a functionally associated network. This network was then used to develop a new pathway finding algorithm to predict biological pathways in yeast. Our approach discovered 195 biological pathways and 31 functionally redundant pathway pairs in yeast. By comparing our identified pathways to three public pathway databases (KEGG, BioCyc and Reactome), we observed that our approach achieves a maximum positive predictive value of 12.8% and improves on other predictive approaches. This study allows us to reconstruct biological pathways and delineates cellular machinery in a systematic view.

Availability: The method has been implemented in Perl and is available for downloading from http://www.oicr.on.ca/research/ouellette/pandora. It is distributed under the terms of GPL (http://opensource.org/licenses/gpl-2.0.php)

Contact: francis@oicr.on.ca

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

One definition of biological pathways is a defined group of biological entities that are organized in a specified order and perform a specified biological task or function (Viswanathan *et al*., 2008). Cells represent complex structures that can be viewed as organizers of pathways, separating, directing and organizing the inputs and outputs of various pathways. Our understanding of how each pathway works and interacts with other pathways is, however, still far from complete. Using high-throughput techniques, the internal organization of cells can be studied from a systematic perspective. For example, the interactomes of several model organisms such as *Saccharomyces cerevisiae* (Gavin *et al*., 2002, 2006; Ho *et al*., 2002; Ito *et al*., 2001; Krogan *et al*., 2006; Uetz *et al*., 2000), *Drosophila melanogaster* (Formstecher *et al*., 2005; Giot *et al*., 2003) and *Caenorhabditis elegans* (Li *et al*., 2004) have recently been extensively studied in large-scale protein–protein interaction (PPI) studies, providing us with rich data sets from which to map disparate functional modules in these interactomes onto biological pathways at the protein level. To complement these proteomic studies, recent efforts on the generation of large-scale genetic interactome data sets have helped us to interpret pathway organization in *S.cerevisiae* (Meluh *et al*., 2008; Schuldiner *et al*., 2005; Tong *et al*., 2001, 2004), *C.elegans* (Kamath *et al*., 2003; Lehner *et al*., 2006) and *D.melanogaster* (Boutros *et al*., 2004) at the gene to phenotype level. Similarly, at the transcription level, microarray techniques have generated large amounts of data enabling the construction of transcription networks for specific biological pathways under any given biological condition of interest (Curtis *et al*., 2005). In spite of these developments, results to date have yielded few overlapping data sets, making it difficult to infer the organization of pathways. This situation has prompted us to propose and develop a novel computational approach that integrates disparate biological information and predicts specific pathways (defined group of proteins that are organized in a specified order and perform a specified biological task or function) and their organization.

In defining a pair of proteins as the basic unit of a pathway, and by revealing the functional relevance of these pairs, biological evidence can be used to infer their roles in the context of a pathway. It is possible for us to utilize databases containing biological data sets to explore how pathways are organized. Kelley and Ideker (2005) developed a log-odds scoring model that identified 360 pathway pairs and 401 pathways in yeast by incorporating physical and genetic interactions (GIs) (synthetic-lethal and -sick interactions). Their study provides a starting point to reveal pathway organization and function from high-throughput data. Ulitsky and Shamir (2007) proposed a modified methodology based on Kelley and Ideker's approach and identified 140 pathway pairs and 280 pathways that contain more information regarding GIs than the previous method. In both approaches, the connection of each protein pair is scored by the probability of observing this connection at random for the given networks, which might result in limited performance due

[*]To whom correspondence should be addressed.

to inaccurate null hypotheses of the underlying statistical tests. Furthermore, neither of these methods consider the situation where some identified pathways contain both dense physical and dense GIs, resulting in large pathway sizes that need to be further clustered. Instead of employing both physical and GIs, Ma and colleagues (Ma *et al*., 2008) designed a method using synthetic lethal interactions alone. They identified 2590 pathway pairs and 5180 pathways in yeast by searching approximately complete bipartite graphs within the synthetic lethal interaction network. In a recent publication, Brady and colleagues introduced a novel approach that discovered 602 and 1510 pathway pairs by searching stable bipartite subgraphs on two different versions of GI networks (Brady *et al*., 2009). However, since GI data is far from complete, only partial pathway organization can be inferred when using GI data alone, as the proteins outside of GI data sets have been overlooked. Thus, a more comprehensive understanding of the cellular pathway organization requires more heterogeneous data that is functionally associated to complement the GI data.

To address the above limitations, we incorporated four types of functionally associated data in the model organism *S.cerevisiae*: PPIs, GIs, domain-domain interactions (DDIs) and semantic similarity of Gene Ontology (GO) terms. PPI data increases the gene coverage compared to the genetically interacting gene list. However, it has been demonstrated that the quality of large-scale PPI data is limited by its high false-positive and false-negative rates (Pitre *et al*., 2008; Zhu *et al*., 2008). To overcome these limitations, we also included DDIs to provide more biological evidence to protein pairs, as it has been widely accepted that some proteins interact with each other through interactions between their respective domains which are defined as independently structural and/or functional blocks of proteins (Lim *et al*., 1994; McGough *et al*., 2003). Semantic similarities of GO terms provide further evidence to a protein pair in terms of their biological functions. We integrated these four biological data sources for protein pairs with a weighted score that represents pathway relevance between a pair of proteins. We also developed a new graph clustering algorithm to group proteins sharing similar neighborhoods on the weighted network of yeast. By comparing our results to pathway annotations from KEGG (Kanehisa *et al*., 2006), BioCyc (Karp *et al*., 2005) and Reactome (Matthews *et al*., 2009), we found that our approach is able to predict biological pathways with a higher positive predictive value (PPV) compared to other approaches (Brady *et al*., 2009; Kelley and Ideker, 2005; Ma *et al*., 2008; Ulitsky and Shamir, 2007). Our results, which also revealed new members of pathways, provide testable hypotheses for experimental validation. Complemented with other predictive methods, our study makes promising progress in the process of deciphering the entire pathway organization in yeast cells. This approach has application in other eukaryotic systems where large data sets are available.

2 METHODS

2.1 Data sources

We downloaded physical interaction and GI data for *S.cerevisiae* from the BioGRID database (http://www.thebiogrid.org) (Stark *et al*., 2006) version 2.0.49. The BioGRID database is a literature-based repository containing physical interaction and GI data. Interactions are categorized as 'Two-hybrid', 'Affinity Capture-Luminescence', 'Affinity Capture-MS', 'Affinity Capture-RNA', 'Affinity Capture-Western', 'Biochemical Activity', 'Co-crystal Structure', 'Co-fractionation', 'Co-purification', 'Co-localization', 'Far Western', 'FRET', 'PCA', 'Protein-peptide', 'Protein-RNA', 'Reconstituted Complex' in the BioGRID database are selected. For GIs, only interactions labeled as 'synthetic lethality' in BioGRID were selected. After removing redundant interactions, the interaction data contained 43 687 unique physical interactions and 10 735 GIs. We also compiled 7820 DDIs in yeast from two sources: (i) the iPfam database (Finn *et al*., 2005), a DDI database derived from RCSB Protein Data Bank (PDB) crystal structures (http://www.pdb.org); and (ii) the list of predicted DDIs from our previously published GAIA algorithm (Zhang and Ouellette, 2009), a method to identify interacting protein domains.

2.2 Gene ontology similarity scores

The functional relationship of proteins can be estimated from how they share protein annotation in a controlled vocabulary system, such as GO (Ashburner *et al*., 2000). We assigned a semantic similarity score to each protein pair to represent how close they work together in a molecular function. We downloaded the GO terms associated with each protein from the Saccharomyces Genome Database (Nash *et al*., 2007), as of October 2008. Given two groups of GO terms ($G1$ and $G2$) for two query proteins $P1$ and $P2$, semantic similarity between protein pairs was calculated by a similar approach as G-SESAME (Wang *et al*., 2007):

$$Sim(G1,G2)=\frac{\sum\limits_{1\le i\le|G1|}\sum\limits_{1\le j\le|G2|} Sim(Term1,Term2)}{|G1|\times|G2|},$$

where $|G1|$ and $|G2|$ is the number of GO terms associated with $P1$ and $P2$, respectively. The range of semantic similarity scores lies between 0 and 1.The semantic similarity score between two GO terms $t1$ and $t2$ was calculated by the following equation:

$$Sim(t1,t2)=\frac{\sum\limits_{t\in ancestors(t1\cap t2)}(Score_{t1}(t)+Score_{t2}(t))}{\sum\limits_{t\in ancestors(t1)} Score_{t1}(t)+\sum\limits_{t\in ancestors(t2)} Score_{t2}(t)}$$

Score() is the function to measure the edge (semantic relations) connecting two GO terms and defined as:

$$Score_{t1}(t)=\max\{weight\times Score_{t1}(t')\}\text{ if }t\neq t1,$$

where t' is the children of the GO term t. If $t=t1$, the score is 1. The weight score is 0.8 for the 'is-a' relation and 0.6 for the 'part-of' relation as in Wang *et al*. (2007).

2.3 Data integration to a weighted biological network

For each protein pair in the physical and GI data, we assigned a confidence score to each connection by combining four types of biological evidence: physical interaction, GI, DDI and GO term similarity. If a physical interaction connects a pair of proteins, we assigned 1 to it, otherwise 0. If a DDI connects a pair of proteins, we assigned 1 to it, otherwise 0. To minimize GIs within pathways, we assigned 0 to a pair of proteins if a GI connects them, otherwise 1. We followed the previously described method to calculate a GO term similarity score for each pair. An integrated score was calculated by averaging these four scores under the assumption that the score from each type of evidence contributes equally to the association between a pair of proteins. Finally, we generated a biological network in which each protein connects to other proteins by the weighted edges. In total, the resultant network contained 5280 proteins.

2.4 Pathway finding algorithm

We developed a new clustering algorithm based on the weighted network. Given a weighted biological network G in yeast, our algorithm computes the following step to find clusters representing pathways {P} in a similar fashion as previous studies (Huttenhower *et al*., 2007; Mete *et al*., 2008):

Step A. For each protein in the network, a pathway protein label was applied if it had at least n topologically similar proteins. Here, n was set to 2, the minimal size of a pathway being two proteins. Given a protein x, a set of topologically similar proteins Y of protein x was defined by the Jaccard coefficient:

$$Y=\left\{\frac{\{neighbors\ (t_i)\}\cap\{neighbors\ (x)\}}{\{neighbors\ (t_i)\}\cup\{neighbors\ (x)\}}>s:t_i\ is\ a\ set\ of\ neighbors\ T\ of\ x\right\}.$$

Here, s is the threshold of topological similarity scores.

Step B. Each protein labeled as a pathway protein was used as a starting point of a pathway P by iteratively searching topologically similar proteins to it and adding them to P unless it had already been classified.

Step C. Each remaining protein (not labeled as a pathway protein) was added to each pathway if it has connections to multiple pathways; otherwise, it was classified as a non-pathway protein.

Pathway finding algorithm

```
Input: $G, s, n$
Output: {P}
for each $x \in G$ do
$T$ = neighbors($x$) // T is a set of neighbors of x
                for each $t \in T$ do
                    $y=\frac{\{neighbors(t)\}\cap\{neighbors(x)\}}{\{neighbors(t)\}\cup\{neighbors(x)\}}$
                    if $(y > s)$
                            topological_neighbors ← $t$
                    end if
                end for
                if (topological_neighbors >= $n$)
                        pathway_proteins ← $x$
                end if
end for
until each protein $x \in$ pathway_proteins is assigned to a pathway ID do
                assign $x$ to a pathway P
                recursively find topological similar proteins $Y$ of $x$
                until each protein $y \in Y$ is assigned to a pathway ID do
                        assign protein $y$ to P
                end until
end until
return {P}
```

2.5 Evaluation of the algorithm (adjusted rand index)

We utilized the adjusted rand index (ARI) (Hubert and Arabie, 1985) to measure the similarity of our resultant pathway organization to other pathway annotation sources. The ARI has been widely used in determining the agreement between two partitions of any network. Scores lie between 0 and 1, and when the two tested partitions agree perfectly, the score is 1. For each identified pathway from our approach, we compared it to every pathway in three pathway databases [KEGG (Kanehisa *et al.*, 2006), BioCyc (Karp *et al.*, 2005) and Reactome (Matthews *et al.*, 2009)] and calculated the ARI score for each identified pathway. Given a pathway X from our approach and an annotated pathway Y from KEGG or Reactome, the ARI was calculated as:

$$ARI(X,Y)=\frac{2(A\times B-C\times D)}{((A+D)\times(D+B)+(A+C)\times(C+B))}$$

where A, denoted as (X∩Y), is the number of proteins appearing in both pathways X and Y; B, denoted as $[Z-(X\cup Y)]$ is the number of proteins appearing in neither pathway X nor Y given the number of proteins Z. (The number of proteins in this study is 5280.) in yeast; C, denoted as $[X-(X\cap Y)]$, is the number of proteins appearing in pathway X but not in Y; D, denoted as $[Y-(X\cap Y)]$, is the number of proteins appearing in pathway Y but not in X.

The final index score of pathway X is defined as the maximal score compared to all annotated pathways in databases:

$$ScoreARI(X)=\underset{i=0}{\overset{n}{Max}}(ARI(X,Y_i)).$$

We regarded pathway X as a true positive if ScoreARI(X) is ≥ 0.5, which meant that at least half of two tested pathways agree with each other. This cutoff is significantly greater than found by chance (Wilcoxon Rank Sum test, $P < 10^{-4}$).

2.6 Network randomization

Comparable control networks were generated by randomly rewiring a pair of edges to connect different pairs of nodes in the interaction networks and then repeating the rewiring step. The number of the repeats is equal to the total number of the edges in the networks. This method was previously reported and utilized by other groups (Maslov and Sneppen, 2002; Royer *et al.*, 2008). With this approach, the degree distribution of a given interaction network can be preserved. The randomization procedure was repeated 1000 times.

3 RESULTS AND DISCUSSION

3.1 Parameter tuning

Pandora identifies pathways by finding neighboring proteins based on confidence scores of protein pairs derived from multiple types of biological evidence. Only two parameters for this method require tuning: (i) the threshold of confidence scores (c); and (ii) the threshold of topological similarity scores (s). We applied our pathway finding approach using different combinations of c and s. We then evaluated the performance of our approach by calculating the PPV, which is generated by comparing our identified pathways to the Reactome pathways based on ARI scores. Here, PPV is defined as: number of true positives/(number of true positives + number of false positives). From the observation of the performance plot (Fig. 1), we concluded that our approach achieves the best PPV performance if c and s were set as 0.7 and 0.5, respectively. With these settings, the PPV is 12.8% when tested against the Reactome pathway annotations. Identical settings also show good performance for the KEGG and BioCyc pathway annotations (Supplementary Figs S1 and S2). In addition, when c and s were set as 0.7 and 0.5, we also observed the best recall rates obtained by our approach when tested on three pathway databases (Supplementary Figs S3, S4 and S5). The best recall rates for Reactome, KEGG and BioCyc are 6.6, 8.3 and 8%, respectively. We found that with higher c and s, small sub-networks are generated, and consequently lowering the PPV. On the contrary, with lower c and s, the network contains high noise and generates many false positives.

3.2 Summary statistics of identified pathways

Our approach identified 195 biological pathways, which covers 31% (1617 out of 5280) of the yeast proteins, 38% (16 685 out of 43 687) of the physical interactions, 8.3% (890 out of 10 735) of the synthetic lethal interactions and 18% (1407 out of 7820) of the DDIs involving yeast proteins. The relatively high coverage of both physical interactions and DDIs and the low coverage of GIs indicate that the pathways identified in our study tend to have dense physical interactions while the GIs in these pathways are sparse. It is not surprising that we identified fewer pathways than previous methods because more constraints such as GO term similarity scores and DDIs were applied in identification of the pathways to

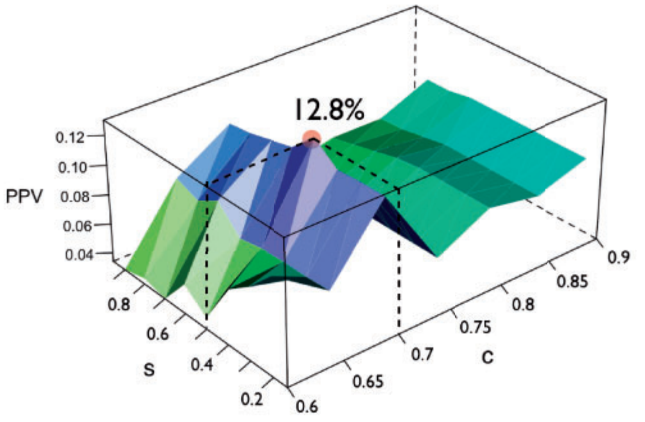

Fig. 1. 3D performance plot tested on different combinations of the threshold of confidence scores (c) and the threshold of topological similarity scores (s). The PPVs of our approach are plotted for different combinations of thresholds when tested against the Reactome pathway annotations. For simplicity, only c ranging from 0.6 to 0.9 and s ranging from 0.2 to 0.9 are tested. The red dot represents the peak showing the best performance of our approach as of 12.8% PPV when c and s set as 0.7 and 0.5, respectively.

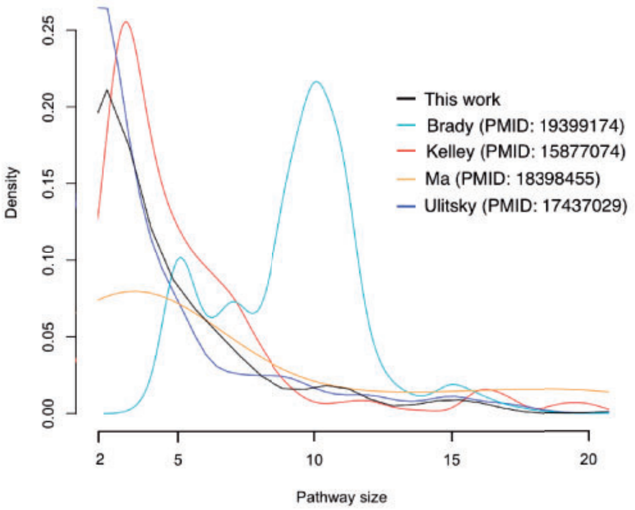

Fig. 2. Distribution of pathway sizes of different approaches. The distribution of pathway sizes of Kelly and Ideker (2005) is represented by the red line; the distribution of pathway sizes of Ulitsky and Shamir (2007) is represented by the blue line; the distribution of pathway sizes of Ma *et al.* is represented by the green line; the distribution of pathway sizes of Brady *et al.* is represented by the brown line and the distribution of pathway sizes of our approach is represented by the black line. All pathways are non-redundant.

ensure the reliability of identified pathways. The size of identified pathways ranged from 2 to 407 proteins, with a strong bias to short pathways. The distribution of pathway size in our study is statistically consistent with that of pathways generated from two previous methods (Kelley and Ideker, 2005; Ulitsky and Shamir, 2007) based on physical interaction data and GI data with the P-value of 0.04 and 2.4×10^{-5}, respectively, by the Wilcoxon Rank Sum test (Fig. 2). However, the distribution is not consistent with that of those approaches (Brady *et al.*, 2009; Ma *et al.*, 2008) based on GIs alone, with the P-value of 0.42 and 0.07, respectively, by the Wilcoxon Rank Sum test. We also found a correlation between the number of protein hubs and the size of the pathway (the Pearson correlation coefficient is 0.79 at P-value $< 2.2 \times 10^{-16}$). In other words, more protein hubs were identified in pathways of larger size. Here, we defined the top 20% proteins in the PPI network of *S.cerevisiae* with high degrees as ‘protein hubs’ as Yu and colleagues presented (Yu *et al.*, 2007). Taken together, we proposed that such a distribution of pathway size reflects a scale-free topological property present in the network, a property that is currently supported by multiple types of biological evidence but not by the GI network alone. A list of the identified pathways and their members found in our study is listed in Supplementary Table S1. We also found that the topological properties of the source PPI network are similar to those of the network of our identified pathways, which indicates that our approach does not appear to have a bias towards the highly connected areas of the source PPI network (Supplementary Table S2).

3.3 Validation of our approach

GO term enrichment analysis was used to measure the cellular functions of identified pathways as performed in previous studies (Carbon *et al.*, 2009; Yi and Stephens, 2008). However, because GO semantic similarity scores have been integrated into our approach as one of types of biological evidence, we used a different evaluation method to measure pathway biological function. We tested our identified pathways on three public pathway databases: KEGG, BioCyc and Reactome. The KEGG database contains manually annotated pathways based on biochemical evidence from the literature, including metabolism, genetic information processing, environmental information processing and cellular processes. BioCyc is a collection of metabolic pathways of 570 organisms and on average pathways in BioCyc are 4.2 times shorter than KEGG pathways. The Reactome database is another manually curated core human biological pathway database. Pathway annotations of organisms other than human are derived by mapping their human counterparts onto these organisms based on protein orthology data. Currently, there are 96, 150 and 381 biological pathways of yeast containing at least two protein members in KEGG, BioCyc and Reactome, respectively. We calculated the ARI scores to quantify the similarity of our 195 resultant pathways and pathway annotations from each pathway database (see Section 2). In this study, we computed the ARI score of each of our identified pathways against every pathway in three pathway databases, and selected the highest resultant score to be the ARI score for the tested pathway. For the KEGG database, we found 4% (8 out of 195) of our identified pathways with ARI scores ≥ 0.5 when tested against the pathways in KEGG. This low percentage, however, is still significantly greater than that found purely by chance (Z-test, $P < 0.001$) with regard to the similarity between the pathways discovered by our approach and the KEGG pathways. For the BioCyc database, there are 5.6% (11 out of 195) pathways with ARI scores ≥ 0.5 when tested against the pathways in BioCyc (Z-test, $P < 4.1 \times 10^{-3}$). For the Reactome database, there are 12.8% (25 out of 195) pathways with ARI scores ≥ 0.5 when tested against the pathways in Reactome (Z-test, $P < 2.6 \times 10^{-4}$). The observed discrepancy on the percentages when tested on three reference databases can be explained by the different ways KEGG, BioCyc and Reactome are curated. KEGG and BioCyc mainly emphasize the metabolic and the signaling pathways, whereas Reactome employs a more general way to collect biological reaction data of pathways. We tested the degree of overlap between these three reference databases using ARI values. We found that there is a 26% overlap between KEGG and BioCyc, possibly due to their similar emphasis on metabolic and signaling pathways. In contrast, there are only 14 and 16% overlaps between Reactome and KEGG, and between Reactome and BioCyc, respectively. This result further addresses the observed discrepancy of PPV when tested on different databases. Furthermore, KEGG relies on Enzyme Commission (EC) numbers to map the physical polypeptides involved in metabolic reactions to public gene/protein annotation databases, and as a result, mis-mapping may lead to the incompleteness of pathway organization.

We also tested whether the proteins within each identified pathway share highly similar phenotypic response patterns. We tested our identified pathways on a data set containing phenotypic response measurements under different treatments (Brown *et al.*, 2006) as used by Ulitsky and Shamir (2007). We found that proteins within the same pathway in our study show significantly higher correlation to phenotypic response patterns compared to that expected by random (the average Pearson correlation coefficient is 0.39 at $P < 4.2 \times 10^{-10}$).

3.4 Comparison between different approaches

Pathway organization derived from biological networks has been widely studied. These approaches are described in previous publications and can be classified into two categories: (i) statistical models with multiple data sources (physical interactions and GIs); (ii) graph-based models with a single data source (GIs). In this study, we also employed a graph-based model, but with diverse lines of biological evidence. To compare the performance of different approaches, we computed the PPV values by calculating the ARI scores between identified pathways from each approach and the pathways from Reactome, KEGG and BioCyc. For the Reactome database, the PPV of Kelley and Ideker (2005), 3.7% (15 out of 404 pathways), is very close to that of Ulitsky and Shamir (2007), which is 3.2% (nine out of 280 pathways). This finding is not surprising because the approach of both methods is identical. Two other approaches also share very similar PPV values: 0.08% (one out of 1297 pathways) for Ma *et al.* (2008) and 0.9% (one out of 108 pathways) for Brady *et al.* (2009) on the more recent version of GI network. Our approach achieves a PPV of 12.8%, indicating that our approach outperforms the other methods when tested on Reactome (Fig. 3). For the KEGG and Biocyc pathway database, performance of the four aforementioned methods follows the same trend as when tested on Reactome (Fig. 3). To compare the performance of different approaches when tested on negative data, we found that all approaches achieves the negative predictive value (NPV) of 100% if tested on randomized pathway data sets, further suggesting better performance of our approach at the same level of NPV. Here, NPV is defined as: number of true negatives/(number of true negatives + number of false negatives).

3.5 Biological examples of predicted pathways

In our study, we have demonstrated that our predicted pathways bear biological meanings as they can be validated by comparing to annotated pathways in Reactome, KEGG and BioCyc. Also, proteins in the same pathway share very similar phenotypic response patterns. The next logical step is to identify usefulness and function of these predicted pathways. We presented several examples to show that biological insights can be inferred from resultant pathways identified in this study. One example is pathway 61 with an ARI score of 0.89 when compared to the 'Orc1 removal from chromatin' pathway in Reactome (Supplementary Fig. S6). Pathway 61 itself is enriched for four GO terms (0000502: proteasome complex/26S proteasome; 0006508: proteolysis and peptidolysis; 0044257: cellular protein catabolism and 0030163: protein catabolism/protein degradation), which is consistent with pathway annotation in Reactome. Ninety-four percent (32 out of 34) of the proteins in pathway 61 are annotated as belonging to the pathway Orc1 removal from chromatin in Reactome; only two proteins (YGL004C, YLR421C) are not included. In fact, YLR421C is a known member of the 26S proteasome (Husnjak *et al.*, 2008; Seong *et al.*, 2007) based on the KEGG annotation while YGL004C is missing from the KEGG pathway, but is a highly related protein (Seong *et al.*, 2007). This example demonstrates the ability of our approach to identify new pathway members, thus providing testable hypotheses for experimental validation. Another interesting example is pathway 20, which is found to match pathway sce03020 'RNA polymerase' in KEGG, with an ARI score of 0.95. Pathway 20 is enriched for the GO term 0030880 (RNA polymerase complex), indicating that it has a similar biological function as the pathway in KEGG. We found pathway 20 contains one more protein (YKR025W) than listed in the KEGG pathway seco03020. As a subunit of RNA polymerase, YKR025W has been extensively studied recently and it plays an important role in the regulation of RNA polymerase III transcription (Flores *et al.*, 1999; Rosonina *et al.*, 2009). Therefore, it is probable that YKR025W is a missing member of the pathway involved in the function of RNA polymerase.

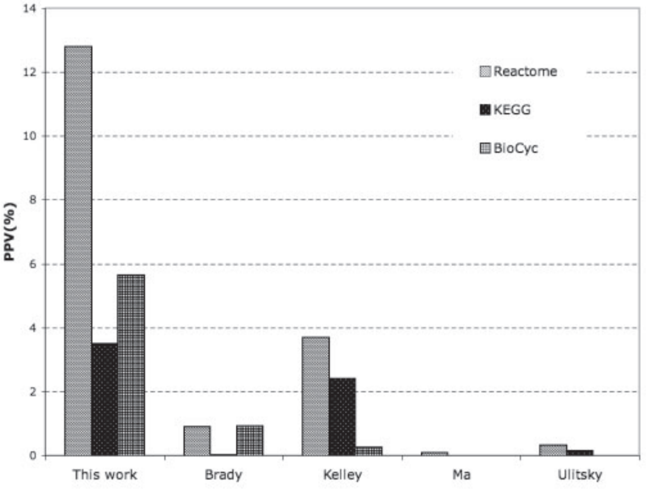

Fig. 3. Comparison between different approaches based on PPV scores tested on Reactome, KEGG and BioCyc pathway annotations. A bar plot demonstrates the performance of each approach tested on three pathway annotations.

3.6 Revealed redundant pathways

Since GIs suggest the existence of parallel pathways, we investigated the possibility of functionally redundant pathway pairs existing in the pathways we identified. To evaluate this, we calculated a

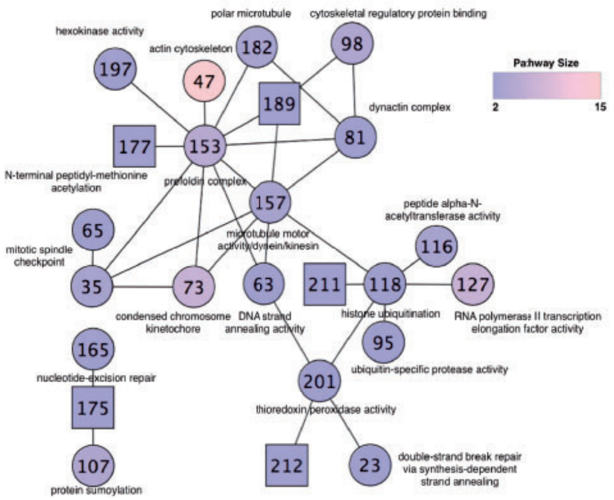
polar microtubule
cytoskeletal regulatory protein binding
hexokinase activity
actin cytoskeleton
182
98
197
47
189
Pathway Size
2
15
dynactin complex
177
153
81
N-terminal peptidyl-methionine acetylation
prefoldin complex
157
peptide alpha-N-acetyltransferase activity
65
microtubule motor activity/dynein/kinesin
116
mitotic spindle checkpoint
35
73
63
211
118
127
condensed chromosome kinetochore
DNA strand annealing activity
histone ubiquitination
RNA polymerase II transcription elongation factor activity
165
95
ubiquitin-specific protease activity
nucleotide-excision repair
201
175
thioredoxin peroxidase activity
107
212
23
double-strand break repair via synthesis-dependent strand annealing
protein sumoylation

Fig. 4. The redundant pathway organization in *S. cerevisiae*. The redundant pathway organization in yeast was generated from discovered pathway pairs. Each node represents a pathway and each edge represents the connection between a pair of redundant pathways. Numbers on nodes are identifiers of our discovered pathways in Supplementary Table S1. The annotation of each pathway was assigned by the GO term with the smallest *P*-value derived from FuncAssociate (Berriz *et al*., 2003). Pathways without GO term annotations were represented as squared nodes. Pathway size was mapped to node color.

Z-score for each possible pathway pair in our identified pathways to show whether or not the difference between the observed number of GIs of our pathway pair and the expected number of GIs of pathway pairs in a random set is statistically significant. We found 31 pathway pairs with *P*-value <0.01 (Fig. 4). A list of these pathway pairs is summarized in Supplementary Table S3. We also found that 58% (18 out of 31) of the pathway pairs contain at least one common functional-enrichment GO term, suggesting the presence of pathway redundancy. For example, pathway 35 and 73 are annotated as the pathways involved in mitotic spindle checkpoint and condensed chromosome kinetochore, respectively. They also share seven function-enriched GO terms (0000777: condensed chromosome kinetochore, 0000778: condensed nuclear chromosome kinetochore, 0000780: condensed nuclear chromosome, pericentric region/condensed nuclear chromosome, centromere, 0000779: condensed chromosome, pericentric region/condensed chromosome, centromere, 0000775: chromosome, pericentric region/centromere, 0000794: condensed nuclear chromosome and 0000793: condensed chromosome) with each other. Pathway 35 shares high similarity with the Reactome pathway 504720 (Amplification of signal from unattached kinetochores via a MAD2 inhibitory signal), with the ARI score of 0.8.

Our predicted pathway pairs represent the redundancy mechanism between a pair of pathways in which proteins can compensate for each other to perform in the same or functionally related biological process. Therefore, we speculated that proteins having similar biological functions might genetically interact with each other if they appear in our identified pathway pairs. For example, pathways 175 and 73 are predicted to be a pair of parallel pathways. We found that there is one enriched GO term (0015630: microtubule cytoskeleton) common to both pathways and there are six synthetic lethal interactions between this pair of pathways, suggesting functional redundancy between them. Due to technical limitations, a large number of GIs in yeast either have been found to be false negatives, or have not yet been tested (Tong *et al.*, 2004). Thus we hypothesized that a pair of proteins found within a pathway pair might genetically interact with each if they share at least one common GO term. We did a 10-fold cross-validation test in which a set of 2371 GIs between pathways that share at least one common GO term and 2371 genetically non-interacting protein pairs tested by Tong *et al*. (2004) was used. Our approach achieved an average sensitivity of 72% and an average specificity of 81%, suggesting good capacity of discovering GIs. For example, ADA2 (YDR448W) in pathway 76 and BRE1 (YDL074C) in pathway 118 share two common GO terms (0016570: histone modification and 0016569: covalent chromatin modification) yet do not genetically interact with each other based on the GI data. By our approach, however, we predict them as a pair of genetically interacting proteins. In a very recent publication (Lin *et al.*, 2008), it was reported that there is a synthetic fitness or lethality defect interaction between ADA2 and BRE1, involved in yeast histone acetylation and deacetylation. This finding provides a good example of the ability of our approach to predict novel GIs. We also generated a network of discovered redundant pathways (Fig. 4 and Supplementary Fig. S7). As expected, most pathways show the 1 : 1 redundant relationship. Interestingly, we found that several pathways, such as pathways 35, 118 and 153, demonstrate the 1 : *N* redundant relationship. By closely examining these pathways, we found them to contain a 3.6-fold enrichment of GO annotations compared to other pathways. Because some of these pathways intersect with multiple pathways, we speculate that these pathways are temporally and spatially multi-tasking.

4 CONCLUSION

In this study, we introduced a systematic multiple evidence-based pathway finding approach in *S.cerevisiae*. In contrast to previous approaches, we examined the pathway organization in yeast in terms of the protein relationship scored by multiple types of biological evidence and discovered 195 biological pathways, which covers 16 685 physical interactions, 890 synthetic lethal interactions and 1407 DDIs involving 1617 yeast genes/proteins. Compared to other predictive approaches, our approach achieved to the best performance when tested against to the Reactome, KEGG and BioCyc pathway databases. We also discovered 31 functionally redundant pathway pairs by a probabilistic test. Analysis of the resulting pathways and pathway pairs provided us with a more comprehensive and reliable view of important pathway organization in yeast. As the size of GI networks in other model organisms grows in the future, our study could ultimately lead us to a more complete identification of the functional interactome interpreted by pathway organization. This could shed light on the overall picture of how subsystems in cells, such as pathways, work together to determine phenotypes and functions.

ACKNOWLEDGEMENTS

The authors are grateful to Michelle D. Brazas, Guanming Wu and Li Zhang for comments on the manuscripts and Joseph Yamada for help with the Pandora web page.

Funding: This work was conducted with the support of the Ontario Institute for Cancer Research through funding provided by the government of Ontario. KXZ is supported by the CIHR/MSFHR Strategic Training Program in Bioinformatics. KXZ is also supported by the CIHR Canada Graduate Scholarships Doctoral Award.

Conflict of Interest: none declared.

REFERENCES

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

Berriz,G.F. *et al*. (2003) Characterizing gene sets with FuncAssociate. *Bioinformatics*, **19**, 2502–2504.

Boutros,M. *et al*. (2004) Genome-wide RNAi analysis of growth and viability in Drosophila cells. *Science*, **303**, 832–835.

Brady,A. *et al*. (2009) Fault tolerance in protein interaction networks: stable bipartite subgraphs and redundant pathways. *PLoS ONE*, **4**, e5364.

Brown,J.A. *et al*. (2006) Global analysis of gene function in yeast by quantitative phenotypic profiling. *Mol. Syst. Biol.*, **2**, 2006 0001.

Carbon,S. *et al*. (2009) AmiGO: online access to ontology and annotation data. *Bioinformatics*, **25**, 288–289.

Curtis,R.K. *et al*. (2005) Pathways to the analysis of microarray data. *Trends Biotechnol.*, **23**, 429–435.

Finn,R.D. *et al*. (2005) iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics*, **21**, 410–412.

Flores,A. *et al*. (1999) A protein-protein interaction map of yeast RNA polymerase III. *Proc. Natl Acad. Sci. USA*, **96**, 7815–7820.

Formstecher,E. *et al*. (2005) Protein interaction mapping: a Drosophila case study. *Genome Res.*, **15**, 376–384.

Gavin,A.C. *et al*. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.

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

Giot,L. *et al*. (2003) A protein interaction map of Drosophila melanogaster. *Science*, **302**, 1727–1736.

Ho,Y. *et al*. (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*, **415**, 180–183.

Hubert,L. and Arabie,P. (1985) Comparing partitions. *J. Classification*, **2**, 193–198.

Husnjak,K. *et al*. (2008) Proteasome subunit Rpn13 is a novel ubiquitin receptor. *Nature*, **453**, 481–488.

Huttenhower,C. *et al*. (2007) Nearest Neighbor Networks: clustering expression data based on gene neighborhoods. *BMC Bioinformatics*, **8**, 250.

Ito,T. *et al*. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA*, **98**, 4569–4574.

Kamath,R.S. *et al*. (2003) Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. *Nature*, **421**, 231–237.

Kanehisa,M. *et al*. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, **34**, D354–D357.

Karp,P.D. *et al*. (2005) Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res.*, **33**, 6083–6089.

Kelley,R. and Ideker,T. (2005) Systematic interpretation of genetic interactions using protein networks. *Nat. Biotechnol.*, **23**, 561–566.

Krogan,N.J. *et al*. (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature*, **440**, 637–643.

Lehner,B. *et al*. (2006) Systematic mapping of genetic interactions in Caenorhabditis elegans identifies common modifiers of diverse signaling pathways, *Nat. Genet.*, **38**, 896–903.

Li,S. *et al*. (2004) A map of the interactome network of the metazoan C. elegans. *Science*, **303**, 540–543.

Lim,W.A. *et al*. (1994) Structural determinants of peptide-binding orientation and of sequence specificity in SH3 domains. *Nature*, **372**, 375–379.

Lin,Y.Y. *et al*. (2008) A comprehensive synthetic genetic interaction network governing yeast histone acetylation and deacetylation. *Genes Dev.*, **22**, 2062–2074.

Ma,X. *et al*. (2008) Mapping genetically compensatory pathways from synthetic lethal interactions in yeast. *PLoS ONE*, **3**, e1922.

Maslov,S. and Sneppen,K. (2002) Specificity and stability in topology of protein networks. *Science*, **296**, 910–913.

Matthews,L. *et al*. (2009) Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.*, **37**, D619–D622.

McGough,A.M. *et al*. (2003) The gelsolin family of actin regulatory proteins: modular structures, versatile functions. *FEBS Lett.*, **552**, 75–81.

Meluh,P.B. *et al*. (2008) Analysis of genetic interactions on a genome-wide scale in budding yeast: diploid-based synthetic lethality analysis by microarray. *Methods Mol. Biol.*, **416**, 221–247.

Mete,M. *et al*. (2008) A structural approach for finding functional modules from large biological networks. *BMC Bioinformatics*, **9** (Suppl. 9), S19.

Nash,R. *et al*. (2007) Expanded protein information at SGD: new pages and proteome browser. *Nucleic Acids Res.*, **35**, D468–D471.

Pitre,S. *et al*. (2008) Global investigation of protein-protein interactions in yeast Saccharomyces cerevisiae using re-occurring short polypeptide sequences. *Nucleic Acids Res.*, **36**, 4286–4294.

Rosonina,E. *et al*. (2009) Sub1 functions in osmoregulation and in transcription by both RNA polymerases II and III. *Mol. Cell Biol.*, **29**, 2308–2321.

Royer,L. *et al*. (2008) Unraveling protein networks with power graph analysis. *PLoS Comput. Biol.*, **4**, e1000108.

Schuldiner,M. *et al*. (2005) Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. *Cell*, **123**, 507–519.

Seong,K.M. *et al*. (2007) Rpn13p and Rpn14p are involved in the recognition of ubiquitinated Gcn4p by the 26S proteasome. *FEBS Lett.*, **581**, 2567–2573.

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

Tong,A.H. *et al*. (2001) Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science*, **294**, 2364–2368.

Tong,A.H. *et al*. (2004) Global mapping of the yeast genetic interaction network. *Science*, **303**, 808–813.

Uetz,P. *et al*. (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, **403**, 623–627.

Ulitsky,I. and Shamir,R. (2007) Pathway redundancy and protein essentiality revealed in the Saccharomyces cerevisiae interaction networks. *Mol. Syst. Biol.*, **3**, 104.

Viswanathan,G.A. *et al*. (2008) Getting started in biological pathway construction and analysis. *PLoS Comput. Biol.*, **4**, e16.

Wang,J.Z. *et al*. (2007) A new method to measure the semantic similarity of GO terms. *Bioinformatics*, **23**, 1274–1281.

Yi,M. and Stephens,R.M. (2008) SLEPR: a sample-level enrichment-based pathway ranking method—seeking biological themes through pathway-level consistency. *PLoS ONE*, **3**, e3288.

Yu,H. *et al*. (2007) The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput. Biol.*, **3**, e59.

Zhang,K.X. and Ouellette,B.F. (2009) GAIA: a gram-based interaction analysis tool—an approach for identifying interacting domains in yeast. *BMC Bioinformatics*, **10**(Suppl. 1), S60.

Zhu,J. *et al*. (2008) Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nat. Genet.*, **40**, 854–861.