Evaluating Between-Pathway Models with Expression Data

B. J. HESCOTT, M. D. M. LEISERSON, L. J. COWEN, and D. K. SLONIM

July 22, 2009

Abstract

Between-Pathway Models (BPMs) are network motifs consisting of pairs of putative redundant pathways. In this paper, we show how adding another source of high-throughput data, microarray gene expression data from knockout experiments, allows us to identify a compensatory functional relationship between genes from the two BPM pathways. We evaluate the quality of the BPMs from four different studies, and we describe how our methods might be extended to refine pathways.

1 Introduction

In this paper, we use microarray expression data to validate instances of an important class of network motif in the yeast interactome. These motifs, called “Between Pathway Models” or BPMs, consist of putative pairs of redundant pathways and were first described by Kelley and Ideker in a seminal paper in 2005 [KI05]. The basic idea behind the BPM is simple. Protein-protein interaction relationships may be divided into two types: physical interactions and genetic interactions. In particular, the subset of the genetic interactions that are “synthetic lethal” interactions are considered. Synthetic lethality between two yeast genes indicates that strains lacking either individual gene (called “knockout” or “deletion” strains) are viable on rich media, but if the two genes are deleted simultaneously, the yeast is inviable. Kelley and Ideker defined two sets of genes, G1 and G2, to represent a BPM motif if there are few synthetic lethal edges *within* G1 and *within* G2, but many synthetic lethal edges between G1 and G2, and if the opposite holds for the physical interaction edges, i.e., there are many physical interaction edges within G1 and G2, but few between G1 and G2. In their notation and ours, gene sets G1 and G2 are referred to as the “pathways” of this BPM.

The term BPM has been used to describe both purely mathematical substructure within the protein-protein interaction network, but also as a synonym for two sets of genes that function in redundant pathways. It is in the second fashion that we will use the terms “BPM” and “putative or candidate BPM” in this work; that is, independent of the exact definition of the graph theoretic criteria used in each of the papers we consider, the two sets of genes (and their connectivity edges) will be considered a candidate BPM for us, where we will try to validate the assertion that these two sets of genes come from redundant pathways using other high-throughput measures (in the present case, using microarray studies).

The motivation for this terminology, and indeed for the interest in the BPM network motif, is the following. Organisms often evolve a sort of fault-tolerance or redundancy, in that they may have multiple sets of proteins capable of performing certain essential functions. Suppose, for some particular essential function, there are two pathways, each of which can compensate for the other if only one of the pathways is disrupted. Then one would expect to see many physical interactions between the genes in each pathway. One would also expect to see synthetic lethal interactions whenever a gene essential to *each* pathway is simultaneously suppressed or deleted. Thus these pathways will organize into a BPM pattern, with many synthetic lethal edges between G1 and G2 and with many physical interaction edges tending to occur within G1 and within G2.

Compared to Kelley and Ideker [KI05], a subsequent paper of Ulitsky and Shamir [US07b] identified BPMs using somewhat different definitions of "many" and "few." The graph theoretic structures of the Ulitsky and Shamir pathways, and thus of their BPMs, are somewhat different (Ulitsky and Shamir also included "synthetic sick" interactions in their data). Recent work of Ma, Tarone and Li [MTL08] and of Brady et al. [BMDC08] redefined BPMs mainly in terms of the structure and placement of the synthetic lethality edges only, evaluating their putative BPMs after the fact by enrichment or the placement of physical interaction edges within the motifs. All four studies claim that their putative BPMs form possible examples of pairs of compensatory pathways in yeast. It is important to stress that regardless of the *algorithm* used to generate the putative BPM, or the exact graph-theoretic structure that is sought, all four studies generate pairs of sets of genes (G1,G2), where they claim that G1 and G2 make up compensatory pathways.

It is hard to validate the claim that two individual sets of genes (G1,G2) make up compensatory pathways given only the networks of protein-interaction and genetic-interaction data. One method might be to see if a study's putative BPMs *correctly predict* which untested pairs of genes will physically interact or will be synthetically lethal. Note that Brady et al. [BMDC08] employ this approach: the method of [BMDC08] run on data from a prior BioGRID download was shown to be able to predict the location of "new" synthetic lethal edges in a more current version of BioGRID. However, the standard measure of "goodness of BPMs" agreed on by all four studies [KI05, US07b, MTL08, BMDC08] has been to show that a substantial fraction of their BPMs exhibit functional coherence, in the form of GO enrichment for the pathways. While this sort of enrichment result provides evidence that the genes in each pathway have some common function, it does not directly demonstrate a relationship between the two pathways unless the enriched functions happen to be related.

In this paper, we show how adding another source of high-throughput data, microarray gene expression data, allows us to identify a compensatory functional relationship between genes from the two pathways. Expression data gives us a temporal snapshot of the cells' RNA under specific conditions. We can use gene expression data from knockout experiments to help evaluate the quality of the BPMs from all four studies.

In particular, suppose that we had microarray expression profiles of all yeast single-gene deletion strains. Then suppose (G1,G2) is a putative BPM with pathways G1 and G2, and let g be a gene from G1. If the loss of g disables pathway G1, and if pathway G2 is truly compensating for G1's loss as the BPM suggests, then we might see the genes in G2 show a coherent change in gene expression in the g-deletion strain compared to wild-type. We furthermore expect this change to be stronger than what we would see for a random set of genes.

This signal is subtle, but detectable. Our method is similar to that used by the Gene Set Enrichment Analysis (GSEA) method [STM$^+$05] to test for coherent expression of gene sets. We compute a cluster-rank score, analogous to their Enrichment Score, that measures the coherence of expression changes of the genes in the pathway. Because the pathway being validated may include genes involved in different but related roles in the same functional process, we use the absolute value of the log expression change to rank our genes. Thus, a pathway that has strong up-regulation of some genes and strong down-regulation of others in response to a compensatory deletion will still score well.

In practice, we do not currently have expression profiles for all the yeast deletion strains. However, the Rosetta Compendium [HMJ$^+$00] includes expression profiles of 276 deletion mutants. This is enough data for an initial assessment of many of the BPMs, and certainly enough for method development. In all four studies we find example BPMs that we can validate, though the percentage of validated BPMs is highest for the BPMs of Brady et al. [BMDC08]. Since the previous version of this work [HLCS09] we have tested the set of potential BPMs against all deletion strain data from the Rosetta Compendium.

We show how to use expression data to refine pathways within a candidate BPM. Again, suppose (G1, G2) is a putative BPM with pathways G1, and G2. Let g be a gene in G1. Now suppose that when g is deleted or disabled G2 does not show a coherent change, but a subset P of G2 does express a coherent change. This subset P could better represent a compensatory pathway. To see if this is a consistent refinement we test to see if P expresses a coherent change when the other genes in G1 are deleted or disabled.

We note that there have been several successful previous approaches for integrating microarray measurements with protein-protein interaction knowledge. Ideker et al. [IOSS02] searched for connected sets of genes with unexpectedly high levels of differential expression. Segal, Wang and Koller [SWK03] and Ulitsky and Shamir [US07a] used correlated expression data together with protein-protein interaction data to identify sets of genes that putatively act in similar pathways. Liu et al. [LLK$^+$07] showed how such analysis could be done in reference to genes disregulated in type 2 diabetes; and Ulitsky, Karp and Shamir [UKS08] most recently did a high-throughput analysis identifying connected gene subnetworks enriched for genes disregulated in different diseases.

2 Methods

2.1 Computing the ClusterRank score

Our BPM evaluation method relies on measuring the coherence of changes in the complementary pathway's genes' expression levels. We do this by computing what we call the *ClusterRank Score*. Note that this score is similar to the GSEA Enrichment Score [STM$^+$05], except that the weights are the raw ranks of the *absolute value* of the "log 10" ratio of change in expression with regard to wild-type. This allows us to validate pathways whose genes respond strongly, but in different ways, to knockouts in a compensatory pathway.

Let (X, Y) be the two pathways of a candidate BPM. Without loss of generality, let $\bar{x} \in X$ be a gene for which there is expression data for the deletion strain. Furthermore, we require that Y contain at least three genes. For every gene g, we define the "log-10-ratio" (as in the Rosetta Compendium data) to be $Q_{\bar{x}}(g) = \log_{10} \frac{\varepsilon_{\bar{x}}(g)}{\varepsilon^*(g)}$ where, $\varepsilon_{\bar{x}}(g)$ is the expression of gene g in the deletion strain $\bar{x}$ and $\varepsilon^*(g)$ is the expression of gene g in wild-type. Let the yeast genes, $g_1 \dots g_N$ be enumerated such that $|Q_{\bar{x}}(g_1)| \leq |Q_{\bar{x}}(g_2)| \leq \dots \leq |Q_{\bar{x}}(g_N)|$. Similar to the GSEA method define:

$$hit(i) = \frac{\Sigma_{j=i}^{N} j I_j}{\Sigma_{j=1}^{N} j I_j}$$

$$miss(i) = \frac{\Sigma_{j=i}^{N} (1 - I_j)}{\Sigma_{j=1}^{N} (1 - I_j)}$$

where

$$I_j = \begin{cases} 1 \text{ if } g_j \in \text{ pathway } Y \\ 0 \text{ otherwise} \end{cases}$$

We now define the $ClusterRankScore$ as:

$$ClusterRankScore(Y_{\bar{x}}) = \max_{i=N \text{ to } 1} hit(i) - miss(i).$$

Note that genes with the same “log 10” ratio have the same rank.

2.2 Estimating Statistical Significance

How can we tell whether the observed score for a particular pathway is good or not? It is hard to know what a good $ClusterRankScore$ is because, like the GSEA Enrichment Score, it depends in part on the number of genes in the pathway. Therefore, we generate an appropriate null distribution each time, and estimate significance from that.

Specifically, the $ClusterRankScore$ of a pathway Y is converted to a p-value using a permutation test as follows. We generate 99 random subsets of genes the same size as the complement pathway, Y, $G_1, \ldots G_{99}$, with $|G_1| = \ldots |G_{99}| = |Y|$. For each subset, G_i, we calculate the $ClusterRankScore(G_{i\bar{x}})$. The set of all 100 tests are sorted and the p-value is the percentile of $ClusterRankScore(Y_{\bar{x}})$.

We say a BPM pathway is *validated* if its p-value is below 0.10. This value was chosen as an arbitrary cutoff. In fact, we can simply use these p-values to rank the pathways from most- to least-supported by the given knockout data available. In Figure 6 we look instead at cumulative rates of validation as this cutoff value ranges from zero to one.

We also performed a second test we call the *random-knockout validation test* on a subset of the validated pathways. In this case the $ClusterRankScore$ of Pathway Y with knockout gene $\bar{x}$, $ClusterRankScore(Y_{\bar{x}})$, is compared against the scores of random knockout strains of genes not present in either pathway. Specifically, we find 99 random deletion mutants $\bar{z}$, $\bar{z} \notin X, \bar{z} \notin Y$ and compute $ClusterRankScore(Y_{\bar{z}})$. All 100 scores are sorted, and the reported p-value is the percentile of $ClusterRankScore(Y_{\bar{x}})$ in the list. If the p-value of pathway Y is less than 0.10 we say that the BPM passes the random-knockout validation test. More about the relative merits of the two validation methods is said in the Discussion.

2.3 Using Expression Data to Refine BPMs

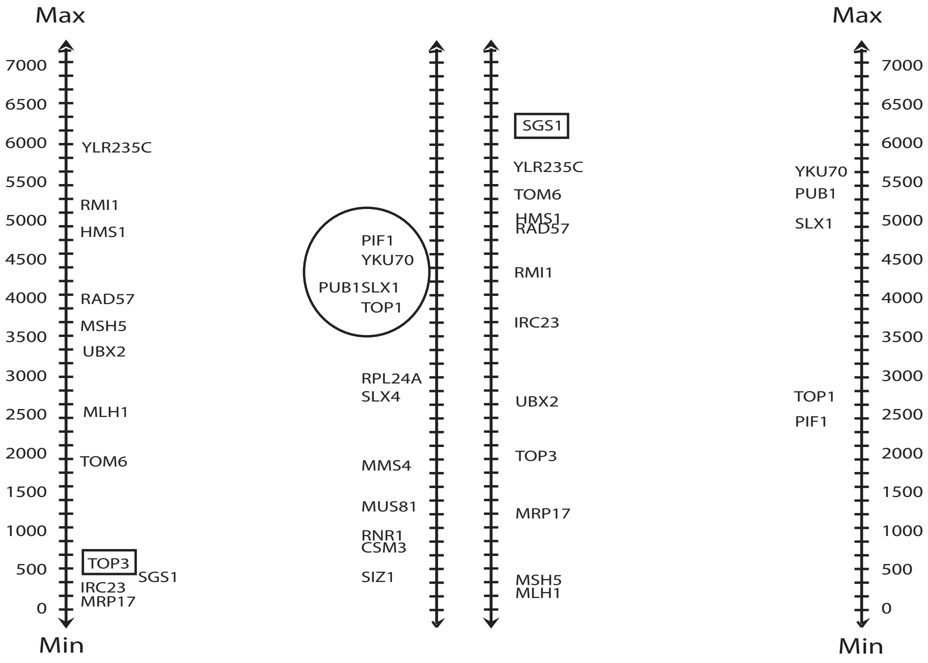


Figure 1: A BPM from Brady plotted with deletion strain data from TOP3. The individual genes in each pathway are represented by their name. The deleted strain's gene is boxed. Each is plotted next to a number line indicating the ranks of the genes when sorted by their differential expression between the wild-type and mutant strains. Here the prune set generated by TOP3 is circled. On the right the prune set is witnessed by SGS1.

As many BPMs are created using differing techniques, it is possible to refine the proposed BPMs using expression data. The refinement we consider is a pruning of a pathway. We devise a method to remove these potentially noisy genes within a BPM. Consider a putative BPM with pathways G1 and G2 as in the left side of figure 1. For a gene g in pathway G1 we define the prune set, $P(G2)_g$ with threshold t, to be the set of genes from G2 whose absolute value of the log-10-ratio of the change in expression for deletion strain g as compared to wild-type is greater than the threshold t. In the example in Figure 1 the gene TOP3 is deleted and we see the genes in the second pathway ranked with regard to their change in expression. If we then consider only the set of genes which rank in the top half, we obtain the prune set circled. We call the gene g the generating gene for the prune set. Note that every pathway G1 in a BPM can generate up to $|G1|$ potential prune subsets of the opposite pathway G2 for a fixed threshold t. We say a gene g' in pathway G1 is a *witness* to the prune $P(G2)_g$ if $g' \neq g$ and its $ClusterRankScore(G2_{\bar{g'}})$ does not decrease. This is illustrated by the gene SGS1 in the right side of Figure 1. We say a gene g'', $g'' \neq g$, g'' in pathway G1, is a *non-witness* if its score decreases. Note there are two possibilities with a non-witness; either the particular prune set is not a valid prune for this BPM, or the gene g'' should be removed from pathway G1 in the model. To avoid this we consider prunes for which there are no non-witnesses. Specifically, we say a prune is *supported* if it has no non-witnesses.

2.4 Data

The Rosetta Compendium[HMJ$^+$00] includes genome-wide expression profiles for 276 yeast deletion mutants. Each mutant is missing exactly one gene, and for each mutant the expression levels of all yeast genes are measured, compared to wild-type. The $\log_{10}$ of the mutant-to-wild-type expression ratio is reported.

We consider between pathway models where at least one gene from the 276 knockout strains in the Rosetta data appears somewhere in the BPM and the number of genes in the compensatory pathway is greater than two. We find 228 of the 404 BPMs reported by Kelley and Ideker [KI05] satisfy these criteria, as do 39 BPMs from Ulitsky and Shamir [US07b], 78 from Ma et al. [MTL08], and 1062 from Brady et al. [BMDC08].

Note, that Ma et al. describe producing over 2,000 possible BPMs, which they believe contain many false positives. Since they do not supply this large set of BPMs, when we refer to the dataset of the Ma et al. study, we are referring to the subset of the BPMs that they make available; namely those of their BPMs that were found to be enriched for the same function in both pathways.

2.4.1 Website

All data and *ClusterRank* code are available online at http://bcb.cs.tufts.edu/validatedBPM/.

3 Results

3.1 BPM Validation

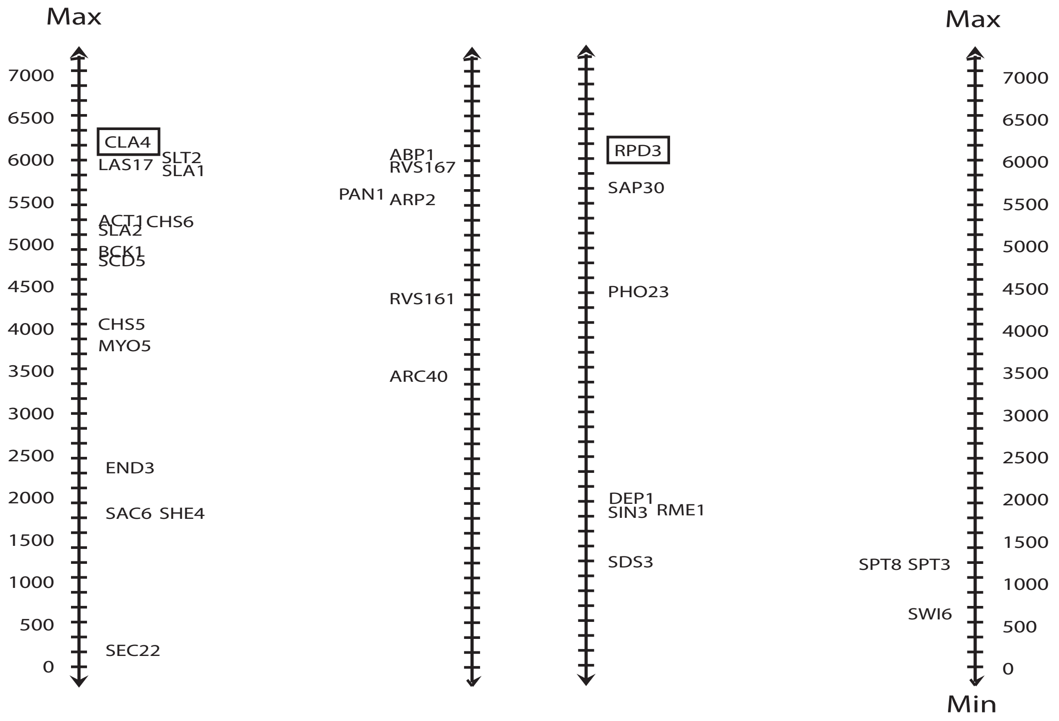


Figure 2: Two BPMs with validation data. The individual genes in each pathway are represented by their name. The deleted strain's gene is boxed. Each is plotted next to a number line indicating the ranks of the genes when sorted by their differential expression between the wild-type and mutant strains. The figure on the left shows the data for a validated BPM from Ulitsky and Shamir. In contrast, the figure on the right shows the genes in a BPM we do not validate. Higher ranks correspond to larger changes in expression value (in either direction) in the deletion strains as compared to wild-type.

In Figure 2 we show both a validated BPM from Ulitsky and Shamir and a second BPM that we do not validate. For the BPM on the left, pathway 1 (by convention, shown on the left-hand side) contains the deleted gene CLA4 . Note that CLA4 has a high rank, as we take the absolute value of its expression ratio. Notice how the genes in pathway 2 are all differentially regulated when CLA4 is deleted, with ranks ranging from 3458 to 6013, yielding a $ClusterRankScore$ of .63 with a p-value $< .05$. In the second BPM in the Figure, in pathway 1 RPD3 is the deleted gene. Note how the differential expression in pathway 2 in response to the loss of gene RPD3 clusters towards the low end, corresponding to an absolute expression change close to zero. The highest ranked expression change, 1263, is below the lowest rank we saw with the previous BPM. These data give the second BPM a $ClusterRankScore$ of -0.79, corresponding to a p-value of 0.97.

In the validated BPM from this figure, both pathways display functional enrichment for the GO Biological Process term "Establishment of Cell Polarity".

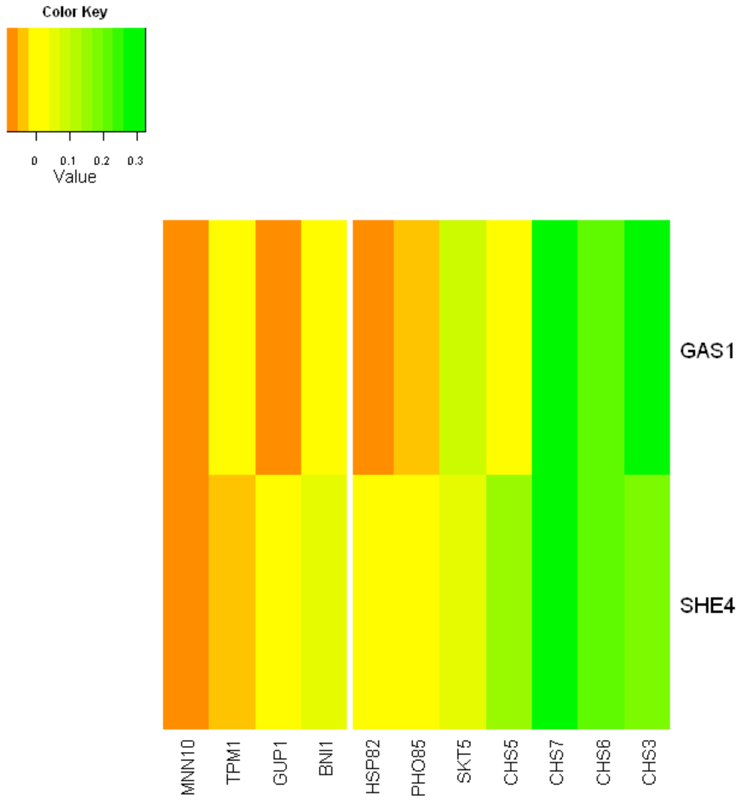
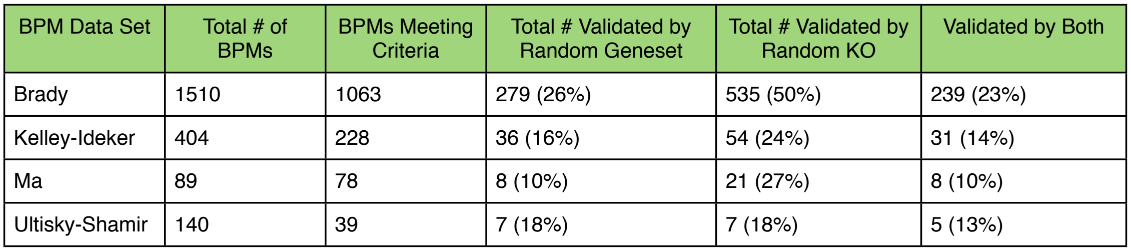


Figure 3: A heat map of the differential expression of yeast genes in response to the deletion of two different genes (SHE4 and GAS1 from Pathway 1) from a validated BPM of Ma et. al. The remaining genes from Pathway 1 are shown on the left and the genes from Pathway 2 are shown on the right. Notice that genes across both pathways are expressed similarly under the GAS1 and SHE4 deletion mutants.

When we have data for two or more deletion mutants in the same pathway, we can do more than simply look at the data from each individually. Figure 3 shows a heat map of a BPM from Ma et al. [MTL08]. Here we have two deletion mutants from pathway 1, SHE4 and GAS1. Both validate the opposite pathway with p-values of 0.02 and 0.09 and ClusterRankScores of 0.64 and 0.56 respectively. With both strains we can compare both pathways and order the genes in terms of up regulation and down regulation. It is clear to see that CHS3, CHS6, and CHS7 are over-expressed within both mutants and that MNN10 is highly down-regulated in both.



BPM Data Set	Total # of BPMs	BPMs Meeting Criteria	Total # Validated by Random Geneset	Total # Validated by Random KO	Validated by Both
Brady	1510	1063	279 (26%)	535 (50%)	239 (23%)
Kelley-Ideker	404	228	36 (16%)	54 (24%)	31 (14%)
Ma	89	78	8 (10%)	21 (27%)	8 (10%)
Ultisky-Shamir	140	39	7 (18%)	7 (18%)	5 (13%)

Figure 4: For each method, we consider all BPMs for which we can compute a $ClusterRankScore$ based on data currently available, i.e. for BPMs where at least one gene intersects the Rosetta Compendium data and its opposite pathway contains at least three genes. Such a pathway is considered validated if its $ClusterRankScore$ has a p-value ≤ 0.10

In Figure 4 we summarize the number of pathways in each of the four studies whose intersection with the knockout data was non-empty and contained at least three genes. A total of 1407 pathways across all four studies met these criteria, of which we found 330 pathways that we were able to validate (ClusterRank test described above with $p < .1$.) The Kelley-Ideker data had 228 pathways meet this criteria with 36 of these pathways validated. Ulitsky-Shamir had 39 pathways tested and 7 of these pathways were validated. Ma, et. al. had 78 pathways meet the criteria with 8 validated pathways. Brady, et. al. had 1062 pathways meet the criteria and had 279 of these validate.

To compare the various techniques we consider the percentage of tested pathways and their $ClusterRankScore$ with regard to their lowest p-value. For each method we plot the p-value against the percentage of pathways that have a smaller or equal p-value. We saw

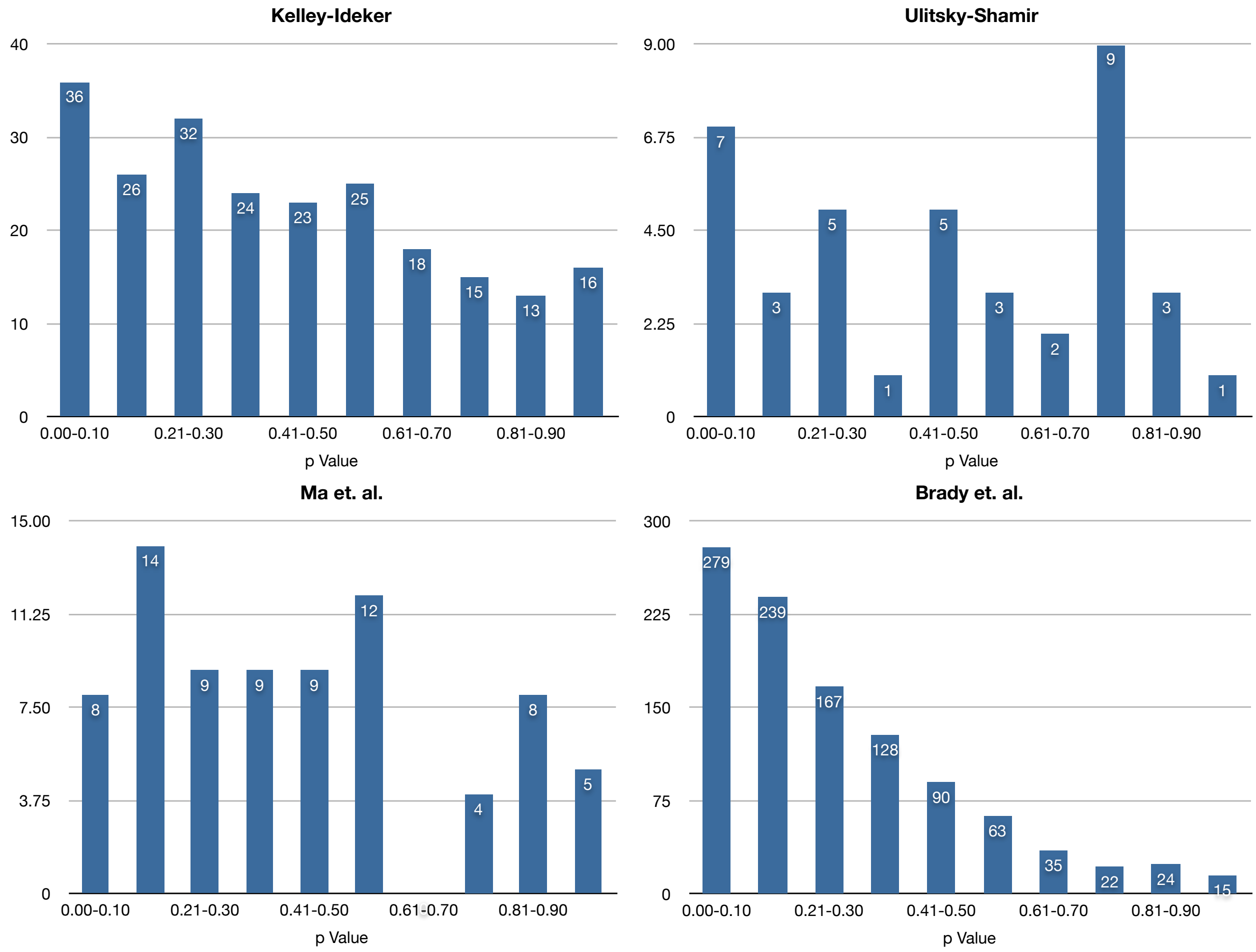


Figure 5: Histograms of the distribution of $ClusterRankScores$ by p-value for each of the four methods.

that 49% of the Brady pathways had values less than 0.30. Similarly we saw 41% for Kelley and Ideker, 38% for Ulitsky and Shamir, and 40% for Ma et. al., see Figures 5 and 6.

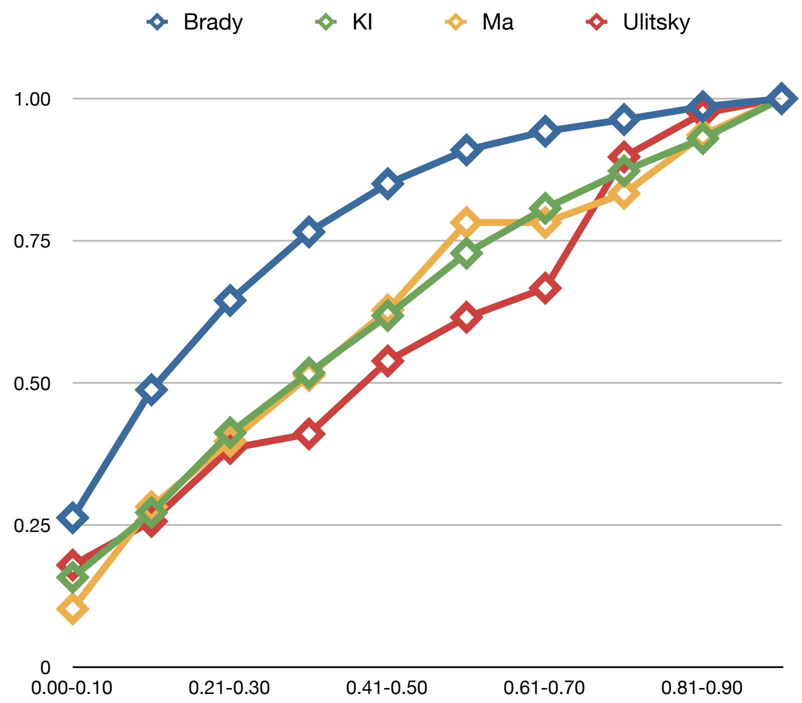


Figure 6: For each p-value, the percentage of pathways tested that have a smaller or equal p-value.

The "validation" described above can be described as a measure for how likely a pathway in a BPM is likely to demonstrate coherent expression changes when a gene in the second pathway is deleted. However, it is possible that a validated pathway would show coherent expression changes when *many* genes (both on and off the second pathway) were deleted, or even just change coherently in the wild-type strain in response to different conditions. If the validated pathway shows *more* coherent expression changes when a gene on its other BPM pathway is deleted than when a random gene is deleted, we say it passes the random-knockout validation test. This more stringent test is again, something that we don't have enough data to apply in a high-throughput fashion; but this stronger test in fact shows that the two pathways are connected to each other, not just individually coherent.

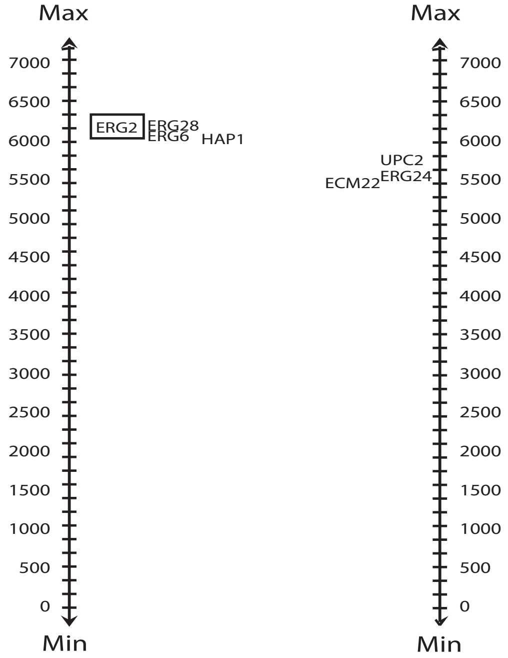


Figure 7: A BPM from Brady et al. that passes both the random-gene validation and the random-knockout validation tests.

In Figure 7 we see a validated BPM from Brady. Pathway 1 contains ERG8, HAP1, ERG6, with knockout ERG2; the complementary pathway contains ERG24, ECM2, and UPC2. Here we have validation that pathway 2 clusters nicely in the ERG2 deletion strain against random gene sets of size 3: we observe a p-value of 0.01 for its $ClusterRankScore$. We also know that this pathway clusters best with this particular deletion mutant as compared to random deletion mutants in the Hughes data set: the BPM passes the random-knockout validation test with a p-value of < 0.01.

In fact, this turns out to involve interesting and important genes. In particular, these genes are all involved in the ergosterol biosynthesis pathway, required for generation of a major constituent of the fungal plasma membrane, ergosterol [PKH92, PC95]. This pathway has been extensively studied as a target of antifungal drugs [LDC$^+$02]. The synthesis of lanosterol is the first step in the pathway dedicated to yeast biosynthesis in wild-type yeast, and ERG24 is required to complete C-14 demethylation of lanosterol. [BEBB07]. Later in the pathway, ERG6 converts zymosterol to fecosterol, which is then converted by ERG2 to episterol [VBB$^+$06]. ERG28 has recently been shown to be a scaffolding protein that physically interacts with ERG6. HAP1, ECM22 and UPC2 are all transcription factors that are involved in the regulation of the ergosterol biosynthesis pathway. Many deletion strains with single mutants of these genes (or double mutants of genes on the same side of the partition) result in accumulation of atypical sterol intermediates, viable, at least in certain genetic conditions and on rich media; for example, single mutants of ERG24 causes the accumulation of the aberrant sterol ignosterol [VBB$^+$06]. Based on the BPM structure and the expression data, we postulate that there is some mechanism of compensation involving the transcription factors ECM22 and UPC2 (perhaps to remove toxic sterol intermediates) when the late stages of the egesterol biosynthetic pathway is disrupted. On the other hand, when ERG24 is deleted, perhaps HAP1 can upregulate other genes to synthesize aberrant sterols that let the yeast hang on to viability in certain favorable conditions. If the two subportions of the pathway are thus compensating, stronger antifungal drugs might result from targeting genes in both subsets simultaneously.

3.2 Using Expression Data to Refine BPMs

In Figure 1 we have a BPM and a pruned BPM from Brady. The putative BPM shown on the left was not validated by either the random knockout test or the random permutation test. Notice in the second pathway that we do see a subset of genes showing a coherent change when TOP3 from Pathway 1 is deleted. On the right we see how SGS1 witnesses this prune. In fact, all 3 genes for which we have deletion strain data in pathway 1 witness this particular prune, making this prune set supported. Furthermore, this pruned BPM passes the random knockout test for this deletion strain. We tested this pruning technique against the potential BPMs with the given deletion strains. We found 26 supported prunes from a potential prune set of 54 for Kelley and Ideker, 3 supported from a potential prune set of size 6 for Ulitsky and Shamir, 39 supported from a total set of 59 for Ma et al. and 1295 supported from a total of 3891 for Brady et al.

4 Discussion

Understanding the functional relationships between proteins is essential for the interpretation of genomic data sets. Manual, experimental construction of pathway models is a slow and painstaking process [GBO^{+}97, DRO^{+}02]. Thus, there is a need for computational methods to predict pathway models and functions. Between-pathway models are an especially advantageous method of pathway-mining, in that the compensatory nature of the two complementary pathways may provide additional clues to function.

Here, we suggest a method for evaluating the results of BPM discovery projects. However, our method is still only a pragmatic approximation of our ultimate goal, which would be reachable if we had gene expression data for all yeast single-gene deletion strains. In that case, we could reasonably expect to find knockouts of most genes on both sides of a BPM. We would then be able to look for consistent ClusterRank scores across most genes in each pathway from a BPM, and for a relationship between the scores of the knockouts on both sides. Even those with only middling p-values might be validated if knocking out every gene in one pathway produces the same set of coherent changes in the other, and vice-versa.

We have instead worked with the Rosetta compendium - the best source of such data available, but still containing knockouts of only about 5% of the yeast genome. Thus, finding BPMs with more than a few deleted genes is difficult, and we have needed to adjust our methods accordingly. We note that, because few BPM pathways contain more than just one deleted gene, we may fail to validate some pathways that are simply noisy rather than incorrect. For example, suppose that a pathway contains 8 pathway- related genes and 2 unrelated ones. If the one deletion strain we have is one of the unrelated genes, then the pathway will not have been validated in our study. However, the addition of more data would address this problem.

Similarly, this lack of data has had an effect on our methods to evaluate statistical significance. If we had the full set of knockout data, then the random-knockout validation method would be the preferred approach. This would be a much stronger way of assessing the significance of coherent expression changes in *the specific gene set* than comparison to a random set of genes. (This is because we know that random sets of genes have low expression correlation, while it is possible that a particular gene set is co-expressed, wildly variable, and completely unrelated to the knockout from the complementary pathway.) However, in this paper, we still prefer random-gene-set permutations over random-mutant-strain permutations, simply because the set of mutants considered in our data set is so small (276), and it is biased heavily towards those genes whose deletion mutants were considered most interesting to investigators. Thus, selecting 100 random mutant strains from this set is almost certainly heavily biased towards specific functions, many of which may well be intentionally correlated with those of the pathways being evaluated.

We point out that the BPMs of Brady et al. represent the work of our own group (LC) or colleagues. We worked independently to design a validation method that made the most sense, and we at first applied it only to the other three data sets (whose combined sizes made their evaluation much more efficient). Only once the methods were fixed did we attempt to evaluate the final Brady data set using our method.

Finally, we note that for specific BPMs, it may be becoming financially feasible for interested investigators to obtain expression profiles of all BPM genes in all or nearly all BPM knockout strains (perhaps by taking advantage of new custom array technologies), and thus having complete data sets may ultimately be a reality. Once such data are available, the pruning techniques can be used to refine these BPM models. Imagine what Figure 3 would look like if it contained data for not just two deletion strains, but deletions for all the genes in one pathway from the BPM. Clustering the rows of the matrix might begin to allow a potential ordering of the genes in the pathway itself.

Acknowledgments

BH and LC were supported in part by NIH grant 1R01GM080330-01A1, ML was supported in part by NSF grant (ASE+NHS)(dms)0428715 and DS was supported in part by NIH grant R21 LM004911.

References

[BEBB07] M. Shah Alam Bhuiyan, J. Eckstein, R. Barbuch, and M. Bard. Synthetically lethal interactions involving loss of the yeast erg24: the sterol c-14 reductase gene. *Lipids*, 42(1):69–76, 2007. PMCID: PMC1847747.

[BMDC08] A. Brady, K. Maxwell, N. Daniels, and L.J. Cowen. Fault tolerance in protein interaction networks: Stable bipartite subgraphs and redundant pathways, 2008. Submitted manuscript.

[DRO$^+$02] EH Davidson, JP Rast, P Oliveri, A. Ransick, C. Calestani, C.H. Yuh, T. Minokawa, G. Amore, V. Hinman, C. Arenas-Mena, O. Otim, C.T. Brown, C.B. Livi, P.Y. Lee, R. Revilla, A.G. Rust, Z. Pan, M.J. Schilstra, P.J. Clarke, M.I. Arnone, L. Rowen, R.A. Cameron, D.R. McClay, L. Hood, and H. Bolouri. A genomic regulatory network for development. *Science*, 295(5560):1669–78, 2002. doi:10.1126/science.1069983 PMID: 11872831.

[GBO$^+$97] S. Goto, H. Bono, H. Ogata, W. Fujibuchi, T. Nishioka, K. Sato, and M. Kanehisa. Organizing and computing metabolic pathway data in terms of binary relations. In *Pac Sympos Biocomp*, 1997. PMID: 9390290.

[HLCS09] Benjamin J Hescott, Mark D.M. Leiserson, Lenore Cowen, and Donna Slonim. Evaluating between-pathway models with expression data. In *RECOMB 2009*, 2009.

[HMJ+00] Timothy R. Hughes, Matthew J. Marton, Allan R. Jones, Christopher J. Roberts, Roland Stoughton, Christopher D. Armour, Holly A. Bennett, Ernest Coffey, Hongyue Dai, Yudong D. He, Matthew J. Kidd, Amy M. King, Michael R. Meyer, David Slade, Pek Y. Lum, Sergey B. Stepaniants, Daniel D. Shoemaker, Daniel Gachotte, Kalpana Chakraburtty, Julian Simon, Martin Bard, and Stephen H. Friend. Functional discovery via a compendium of expression profiles. *Cell*, 102(1):109–126, 2000. doi: 10.1016/S0092-B674(00)00015-5 PMID: 10929718.

[IOSS02] T. Ideker, O. Ozier, B. Schwikowski, and A. Siegel. Discovering regulatory and signaling circuits in molecular interaction networks. *Bioinformatics*, 18:S233, 2002. PMID: 12169552.

[KI05] Ryan Kelley and Trey Ideker. Systematic interpretation of genetic interactions using protein networks. *Nature Biotechnology*, 23(5):561–566, 2005. doi: 10.1038/nbt1096 PMID: 15877074.

[LDC+02] A. Lupetti, R. Danesi, M. Campa, M. Del Tacca, and S. Kelly. Molecular basis of resistance to azole antifungals. *Trends in Molecular Medicine*, 8(2):76–81, 2002. doi: 10.1016/S1471-4914(02)02280-3 PMID: 11815273.

[LLK+07] M. Liu, A. Liberzon, S.W. Kong, W.R. Lai, Peter J Park, Isaac S Kohane, and Simon Kasif. Network-based analysis of affected biological processes in type 2 diabetes models. *PLoS Genetics*, 3(6):e96, 2007. PMCID: PMC1904360.

[MTL08] X. Ma, A.M. Tarone, and W. Li. Mapping genetically compensatory pathways from synthetic lethal interactions in yeast. *PLoS One*, 3(4):e1922, 2008. doi:10.1371/journal.pone.0001922 PMCID: PMC2275788.

[PC95] L.W. Parks and W.M. Casey. Physiological implications of sterol biosynthesis in yeast. *Annual Review of Microbiology*, 49(1):95–116, 1995. doi: 10.1146/annurev.mi.49.100195.000523 PMID: 8561481.

[PKH92] F. Paultauf, S. Kohlwein, and SA Henry. Regulation and compartmentalization of lipid synthesis in yeast. In *The Molecular and Cellular Biology of the yeast Saccharomyces: Gene Expression*, volume 2, pages 415–500, 1992.

[STM$^+$05] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Ladner, and Jill P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, 2005. PMCID: PMC1239896.

[SWK03] E. Segal, H. Wang, and D. Koller. Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics*, 19(Suppl 1):264–271, 2003. PMID: 12855469.

[UKS08] I. Ulitsky, R.M. Karp, and R. Shamir. Detecting disease-specific disregulated pathways via analysis of clinical expression profiles. In *RECOMB 2008*, 2008.

[US07a] Igor Ulitsky and Ron Shamir. Identification of functional modules using network topology and high-throughput data. *BMC Systems Biology*, 1(8), 2007. PMCID: PMC1839897.

[US07b] Igor Ulitsky and Ron Shamir. Pathway redundancy and protein essentiality revealed in the *s. cerevisiae* interaction networks. *Molecular Systems Biology*, 3(104), 2007. PMCID: PMC1865586.

[VBB$^+$06] M. Valachovic, B.M. Bareither, M. Shah Alam Bhuiyan, J. Eckstein, R. Barbuch, D. Balderes, L. Wilcox, S.L. Sturley, R.C. Dickson, and M. Bard. Cumulative mutations affecting sterol biosynthesis in the yeast saccharomyces cerevisiae result in synthetic lethality that is suppressed by alterations in sphingolipid profiles. *Genetics*, 173(4):1893–1908, 2006. PMCID: PMC1569731.