

Detecting Between-Pathway Model conservation across *S. cerevisiae* and *S. pombe* yeast species

Daniel Malmer

Senior Honors Thesis

Tufts University Department of Computer Science
Advisors: Benjamin Hescott, Lenore Cowen

1 Abstract

The Between-Pathway Model (BPM) motif identifies pairs of fault-tolerant gene pathways within the yeast interactome. In BPMs, many gene pairs between the two pathways contain synthetic lethal interactions, meaning if you knock out one gene or the other only, the yeast lives, while knocking out both is lethal. This suggests that each pathway in a BPM compensates when an opposite pathway is suppressed, defective, or absent. Algorithms have been used to find suggested BPMs in the *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* species of yeast. Here we identify and apply a method for finding conservation of BPMs across these two species. By mapping orthologous genes between the two yeast species, we are able to identify BPM pathways which are highly conserved. Furthermore, with Gene Ontology (GO) enrichment, we can infer functional conservation as well. Given the large evolutionary distance between *S. cerevisiae* and *S. pombe*, a high rate of rewiring within the interactome is generally expected. However, in certain BPMs we found a large percentage of gene conservation and GO enrichment. Some of the pathways we identified were even found in the literature to be conserved across mouse and human. With this research, we hope to identify pathways which are evolutionarily conserved across *S. cerevisiae* and *S. pombe*, and encourage continued study of Between-Pathway Models.

2 Introduction

In 2005, A. Kelly and T. Ideker[1] published their seminal paper first describing the structure of Between-Pathway Models (BPMs). The BPM motif in the *Saccharomyces cerevisiae* yeast interactome is a pair of redundant gene pathways based on protein-protein interaction data. More specifically, BPMs are a graph-theoretic interpretation of genetic interactions, thought to indicate fault-tolerance (see the Background sections below for more information). The work of Kelly and Ideker has sparked substantial research on Between-Pathway Models. Three subsequent papers have been published[2, 3, 4] in which the authors developed their own algorithms to find BPMs in *S. cerevisiae*.

Recently, A. Roguev *et al* published a method for finding nearly genome-wide genetic interaction data in *Schizosaccharomyces pombe* yeast[5]. A group at Tufts University modified the algorithm developed by A. Brady *et al*[4] on the new genetic interaction data to predict novel BPMs in the *S. pombe* genome[6]. The results from this algorithm are still unpublished, but can be found at <http://bpms.herokuapp.com/bpms>. With this new set of BPMs, we can begin to compare the BPMs between *S. cerevisiae* and *S. pombe*. In this study, we present a method for detecting BPM conservation using gene relationships across the two yeast species.

3 Background

3.1 Protein-Protein Interactions

The study of Between-Pathway Models (BPMs) involves examining the interactions of genes in yeast to find putative redundant pathways. One method employed by biologists to study the yeast genome is to perform gene knockout studies on every gene in a particular yeast species. The gene knockout study deletes (or suppresses) individual genes and subsequently determines if the yeast is still viable on rich media. If the yeast is not viable after the deletion of a particular gene, it is said that the gene is *essential*. If the yeast is still viable after the deletion of a gene, the gene is *non-essential*. Gene knockout studies of the *Saccharomyces cerevisiae* genome has shown that around 18% of the genes are found to be essential genes[8]. Further gene knockout studies determine the viability of yeast after knocking out not one, but two non-essential genes from the genome. The resulting effects show an interaction between the two genes called a *genetic interaction*, a measure of epistasis[9].

The *protein-protein interaction* (PPI) network is the system of all known connections between proteins in a given species. These connections come in the form of *physical interactions* and *genetic interactions*. Using the double-gene knockout studies described above, biologists have found many varieties of genetic interactions. One case is the interaction of *synthetic rescue* in which knocking out one gene or the other makes the yeast less healthy while knocking out both genes restores the yeast to a wildtype phenotype. Another genetic interaction, *synthetic growth defect* is found when knocking out either gene individually produces a viable yeast, while knocking out both leaves the yeast unhealthy. The final category of genetic interaction is the *synthetic lethal* interaction. This is the case when deleting one gene or the other results in a viable yeast, while deleting both genes results in an inviable

yeast[10].

Figure 1 shows a theoretical example of a synthetic lethal interaction. The nodes are genes, while the solid lines represent physical interactions between genes and the dotted line represents a synthetic lethal interaction. The pathway in (a) shows a healthy pathway where all the genes are active. These genes produce a functional pathway as all the genes are able to physically interact normally. In (b), gene B is suppressed, or knocked out, however the interactions still form a viable pathway through genes A, C, D, and E. The same is true in (c) where gene C is knocked out and the pathway is formed through genes A, B, D, and E. However, in (d), both genes B and C are knocked out, leaving no connections between gene A and genes D and E. The pathway is now inviable.

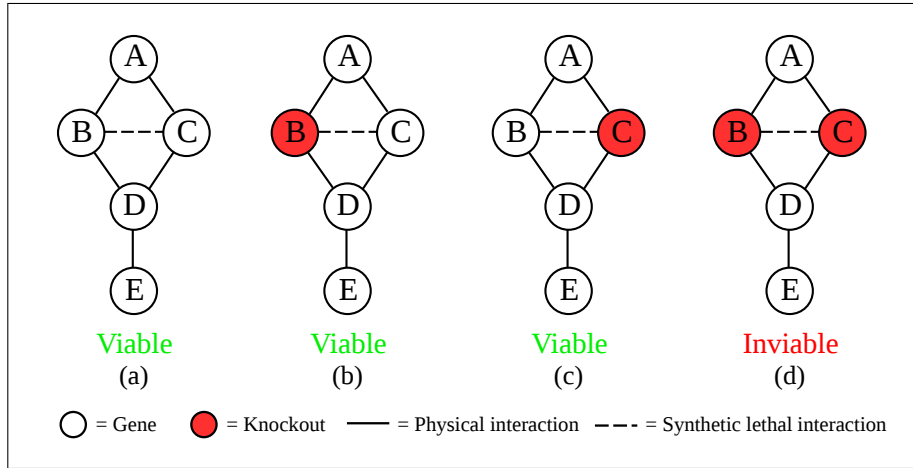


Figure 1: Viability of a theoretical pathway containing a synthetic lethal interaction.

Commonly used resources for genetic interactions data, such as the BioGRID[10] database, classify protein-protein interactions in distinct categories- “Synthetic Lethality”, “Synthetic Growth Defect”, “Synthetic Rescue”, etc. Recently, a scalar representation of the genetic interaction data was published which provides a gradient for the level of synthetic lethality and rescue that results from double gene knockouts[5, 11]. The method used to find this data is called epistatic miniarray profiling, or E-MAP. Instead of a categorical classification, the E-MAP data gives a positive or negative score based on the level of genetic interaction taking place. Positive scores indicate a synthetic rescue interaction. Negative scores indicate synthetic lethal or synthetic growth defect interactions. Thresholds can be set to determine how positive or negative each score needs to be to classify an interaction as rescue, growth defect, or lethal. E-MAP data has been generated for both *S. cerevisiae*[11] and *S. pombe*[5] species.

As opposed to the genetic interaction, a physical interaction within the PPI network is defined as one in which two proteins are connected in one of three ways: one protein physically binds to the other protein, one protein binds upstream on the DNA strand from the other protein, or the two proteins are enzymes that operate on at least one metabolite in common[1]. These physical interactions are what define a gene pathway.

With this protein-protein interaction data within the yeast interactome, we can detect a motif called the Between-Pathway Model, which is the basis for this study.

3.2 Between-Pathway Models

The Between-Pathway Model (BPM) is a description of two sets of genes thought to exist as functionally redundant pathways[1]. By redundant, we mean that the pathways provide the same or similar function so as to acquire a fault-tolerant mechanism. In other words, if one pathway is defective or destroyed, the other pathway will be able to perform a related function to compensate. This sort of mechanism makes sense evolutionarily. If genes within one pathway mutate or are damaged, the redundant pathway is able to take over the first pathway’s function. Thus, the pathways in BPMs are thought to make the interactome more robust and increase the yeast’s fitness.

Using the protein-protein interaction network, we can deduce areas of fault-tolerance within the yeast interactome by finding the Between-Pathway Model motif. Think of the PPI network in terms of a mathematical graph. In this graph, each node represents a protein and the edges between nodes represent the protein-protein interactions. Because there are two types of interactions, there are two types of edges: physical and genetic. The Between-Pathway Model motif in this graph of interactions is defined as pairs of gene pathways containing many physical interactions *within* each pathway and many synthetic lethal interactions *between* the two pathways. Conversely, the two pairs of pathways have very few synthetic lethal interactions *within* each pathway and very few physical interactions *between* each pathway. The idea behind this structure is that synthetic lethal interactions show a measure of compensation between two gene pathways.

Suppressing a gene in one pathway may reduce or destroy that pathway’s functionality[1]. If this functionality is essential to the yeast’s survival, another pathway with redundant functionality is needed to compensate or the yeast will be inviable. If genes in this compensatory pathway are also suppressed, then the yeast will again be inviable. Synthetic lethal interactions show pairs of genes where knocking out one gene or the other produces a viable yeast, but knocking out both leaves the yeast inviable. One explanation for this phenomenon is that these two genes exist in compensatory pathways like the ones just described. Thus, by finding BPM motifs in the interactome with many synthetic lethal interactions between pairs of pathways, we are finding pathways that are likely to be compensatory.

Figure 2 shows a theoretical BPM. As described before, genes within the two pathways are thought of as nodes on a graph and the two types of interactions are edges between nodes. In this case, pathway 1 and pathway 2 form a connected subgraph of physical interactions within each pathway and a clique (every node is connected to every other node) of synthetic lethal interactions between each pathway. These pathways are thought to be functionally redundant. The mutation of a gene in one pathway or the other may destroy that pathways functionality, but the opposite pathway is able to compensate for it.

R. Kelly and T. Ideker, and I. Ulitsky and R. Shamir first detected BPMs using both physical and synthetic lethal interaction data[1, 2]. As described above, the goal of this algorithm is to find pairs of gene sets where many synthetic lethal interactions exist between the two sets and many physical interactions exist within each set. Subsequent papers use only genetic interaction data so as to reduce pathway bias[3, 4]. As a method of verifying the BPMs, physical interactions within each pathway are checked after the BPMs were produced.

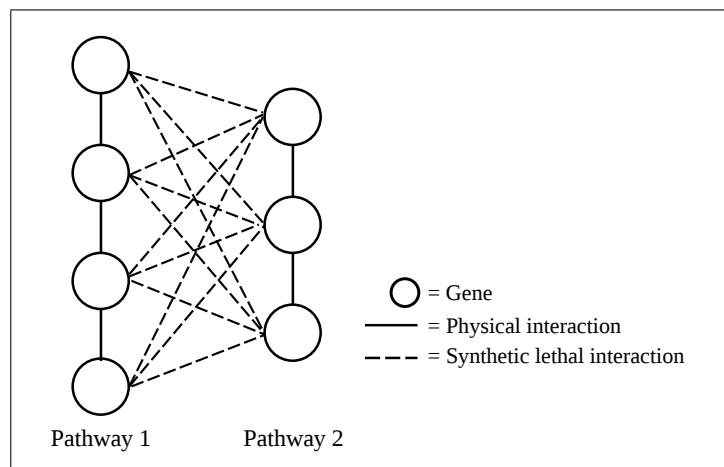


Figure 2: A theoretical Between-Pathway Model. Pathway 1 and Pathway 2 are thought to be fault-tolerant, or redundant, based on the physical interactions within each pathway and the many synthetic lethal interactions between the two pathways.

3.3 GO enrichment

To verify their findings, each paper that produced BPMs had methods of validating their Between-Pathway Models[1, 2, 3, 4]. As mentioned above, papers following Kelly and Ideker, and Ulitsky and Shamir predicted physical interactions in pathways using their BPM identifications[3, 4]. Brady *et al* also ran algorithms on old versions of PPI data to see if the BPMs would correctly predict interactions in the newer version. As well, Hescott *et al* developed a method to use gene expression data to validate previously identified BPMs[12]. What all the papers had in common, though, is they check the Gene Ontology (GO) enrichment to determine if many genes in BPM pathways perform the same biological function. Gene Ontology is structured vocabulary used by biologists to annotate gene function[13]. GO works by creating universal terms for all manually curated, high-throughput, and computational annotations of gene functions. GO is structured as a hierarchy. Well studied genes with known functionality are assigned more specific GO terms that exist farther down the hierarchy. Genes with little-known functionality, on the other hand, will have very general GO terms in high levels of the hierarchy.

With the structured vocabulary, GO terms can be precisely compared to determine functional similarity between genes. The farther down in the GO hierarchy similar GO terms are, the closer the functionality of the compared genes. If a pathway is *GO enriched*, that pathway is said to contain many genes that perform the same or similar function. The GO enriched pathway is given a p-value to indicate the confidence of the enrichment classification. The closer the p-value is to 0, the higher the confidence of the GO enriched pathway[14].

BPMs from all four studies above use GO enrichment as a measure of validating their BPM sets. Showing that a substantial fraction of their BPMs are GO enriched verifies that genes in BPM pathways perform similar functions. It does not, however, verify that the pathways in the BPM are redundant or fault-tolerant. Still, by verifying functional coherence of individual gene pathways, we can say with more confidence that a BPM exists in nature.

GO enrichment is still today the most accepted method of BPM validation[12].

3.4 Homology

In this research, we are attempting to find conservation of BPMs across two species of yeast. To do so, we use information on the relationships between genes in the two species. This data can tell us how related the genes are between the two species' BPMs.

Two genes that derived from a common ancestor gene are called *homologs*. Homologous genes are classified as one of two types: *paralogs* and *orthologs*.

Paralogs are defined as two genes separated by a gene duplication event. Gene duplication occurs when a gene is copied, usually through a transcription error, onto two sections of the species' genome. If the species with the duplicated genes survives, then both copies of the gene will continue to evolve as they undergo evolutionary pressure. The two duplicated genes will evolve separately and attain unique genetic sequences, but will still be evolutionarily related.

Orthologs are defined as two genes separated by a speciation event. Take for instance part (a) of Figure 3, where gene A exists in some ancient species. At some point, this species undergoes a speciation event where the ancestor split into two new species. Now gene A still exists in both of these species, but each gene is undergoing separate evolutionary pressures. These evolutionary pressures change the two genes uniquely, causing their sequences to diverge and become genes B and C. The genes may be different now, but they are still evolutionary related. Genes B and C are classified as orthologs.

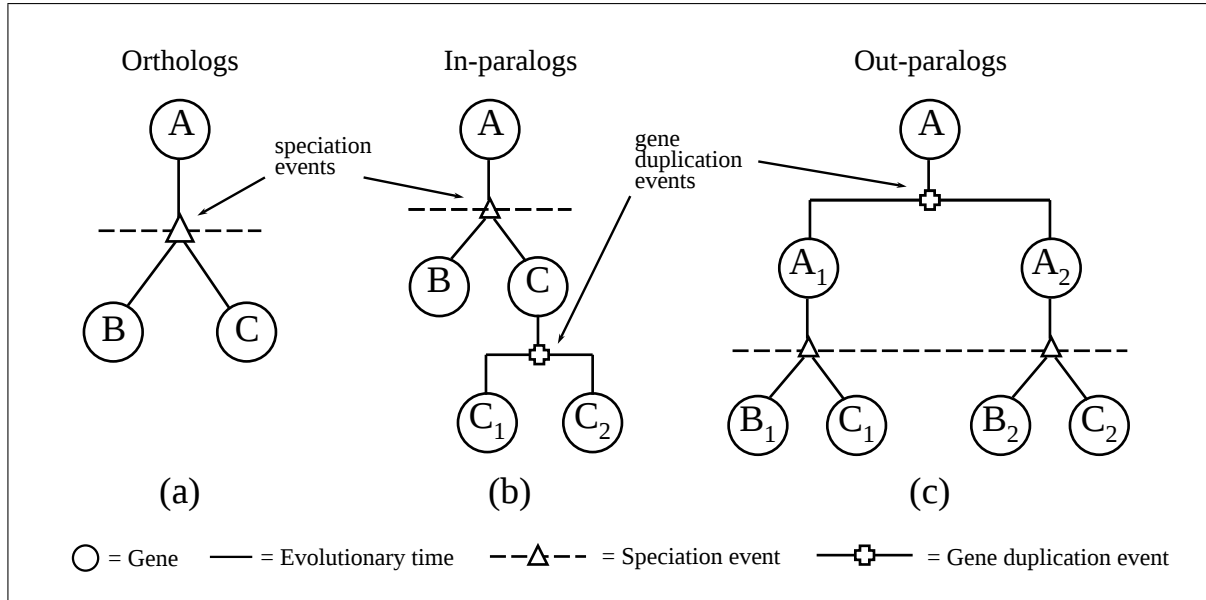


Figure 3: Homologous relationships between genes. (a) Gene B is a normal ortholog of C. (b) Gene B is an in-paralog of genes C₁ and C₂. This is still an orthologous relationship. (c) Genes B₁ and C₁ are out-paralogs of genes B₂ and C₂. This is not an orthologous relationship.

Things get more complicated when gene duplication and speciation events are combined.

To describe the result of these evolutionary events, Tong *et al*[15] coined the terms *in-paralog* and *out-paralog*. An in-paralog occurs when a gene splits at a speciation event, and then one or both of the new genes undergo gene duplication. To better describe this we will refer to part (b) of Figure 3. Gene A diverges at the speciation event to form genes B and C. Then, gene C is duplicated into genes C_1 and C_2 . Genes C_1 and C_2 are then in-paralogs of gene B. Because gene B and genes C_1 and C_2 share the same ancestor gene at the speciation event, the gene B is orthologous to C_1 and C_2 . An out-paralog occurs from a gene that duplicated before a speciation event. Part (c) of Figure 3 shows the relation of an out-paralog. The gene duplication of gene A creates genes A_1 and A_2 . Afterwards, a speciation event occurs, splitting both genes to species B and C. In the case of out-paralogs, the genes of B_1 and C_1 are not orthologous to B_2 and C_2 . This is because the genes of each subscript set have different ancestor genes. From these definitions we can say that we want to find normal orthologs as well as in-paralogs to determine orthologous relationships between two species.

4 Data

Four papers produced original BPMs based on PPI data for *S. cerevisiae*, all of which are used in this research. The first was the seminal paper in 2005 by R. Kelly and T. Ideker[1] that defined the BPM motif. Subsequent papers by I. Ulitsky and R. Shamir[2], X. Ma, A.M. Tarone, and W. Li[3], and A. Brady, K. Maxwell, N. Daniels, L.J. Cowen[4] produced novel BPMs as well. Using *S. cerevisiae* E-MAP data from Collins *et al*[11] and *S. pombe* E-MAP data from Roguev *et al*[5], M. Leiserson, D. Tatar, L. Cowen, and B. Hescott[6] used the algorithm from Brady *et al*[4] to produce more BPMs in *cerevisiae* and *S. pombe*. These two sets of BPMs have not yet been published, so we refer to them as Collins and Roguev in reference to the origins of the E-MAP data.

S. cerevisiae to *S. pombe* ortholog data is from two sources: InParanoid[15] and PHOG[16]. InParanoid ortholog data was retrieved from the InParanoid7 website, <http://inparanoid.sbc.su.se>. PHOG ortholog data was graciously uploaded at our request to their website at <http://phylofacts.berkeley.edu/orthologs/downloads/>.

Gene Ontology enrichment information was retrieved from a site built by Max Leiserson of Tufts University, <http://bpms.herokuapp.com/bpms>. This site allows the user to check the GO enrichment of any BPM in the datasets listed above. The GO information on the website was retrieved from the Gene Ontology Consortium[13].

5 Methods

To detect the conservation of Between-Pathway Models across yeast we perform the following procedure on every BPM of every dataset:

- Map each gene in the BPM to its ortholog, creating an “ortholog BPM” with “ortholog pathways”
- Calculate similarity scores of ortholog pathways when compared against every previously investigated BPM pathway in the opposite yeast species

- Validate highly conserved pathways by checking GO enrichment

All of the code was scripted in the python programming language. The source code is attached as Supplement 1.

5.1 Reading in BPM and ortholog data

Ortholog data between *S. cerevisiae* and *S. pombe* is read in from one of two datasets, InParanoid[15] and PHOG[16]. The naming conventions on the PHOG data varied from the rest of the data, so we used the BioMART on Fungi Ensembl[17] to translate gene names. Genes are stored as pairs in an array, `allOrthologs`. As stated above, there can exist a many-to-many orthologous relationship between genes. Therefore, each gene with more than one ortholog is added multiple times for each unique orthologous relationship.

Between-Pathway Model data is then read from one of the six available *S. cerevisiae* datasets described in the Datasets section. Each individual dataset of BPMs is given in a text file containing pathways of comma-separated genes on each line. Every two pathways make up a unique BPM within the dataset. The program reads the input file sequentially two lines at a time for every BPM. BPMs are stored in a large array called `cereBPMs`, with the index of the array correlating to the BPM number given in the dataset. Each BPM index contains two arrays of genes representing the two gene pathways. Every gene is stored as a `Gene` object containing the name of the gene and a list of its orthologs. As the gene is read in, the `allOrthologs` list is searched to find all pairs containing the gene name. For every gene found in a pair in the list, the orthologous gene is added to the `Gene` object's ortholog list. This is done for every gene in the current dataset of BPMs until every BPM is contained in the `cereBPMs` array. This process is then repeated for the *S. pombe* BPM dataset and read into an array called `pombeBPMs`.

5.2 Mapping BPMs to their ortholog genes

Each BPM is mapped to an orthologous BPM in the opposite species. Without loss of generality, we will refer only to mapping *S. cerevisiae* to *S. pombe*, but the process is performed in the reverse direction as well. Only BPMs with pathways of length 3 or greater are considered when detecting for conservation. This is to avoid pathways of size 1 and 2 that may be hubs of synthetic lethal interactions rather than actual redundant pathways[18].

To map a BPM to its ortholog, the orthologs of each gene within the BPM are placed into another theoretical BPM. So for every BPM in the current *S. cerevisiae* dataset, both pathways are mapped to an "ortholog pathway" in *S. pombe*. The ortholog pathway is made up of genes orthologous to the originals. The two ortholog pathways therefore comprise an "ortholog BPM". Figure 4 shows a visual interpretation of this mapping. You can see that some genes map to multiple orthologs in *S. pombe*, while some do not map to any. This ortholog BPM is entirely theoretical and was not produced by previous researchers. The thought is, however, that if this ortholog BPM of *S. cerevisiae* is similar to an actual BPM of *S. pombe*, then it has been conserved through evolutionary time.

With this ortholog BPM, we can compare it to every *S. pombe* BPM and find those with high similarity. In practice, there is no actual ortholog BPM being created in the program.

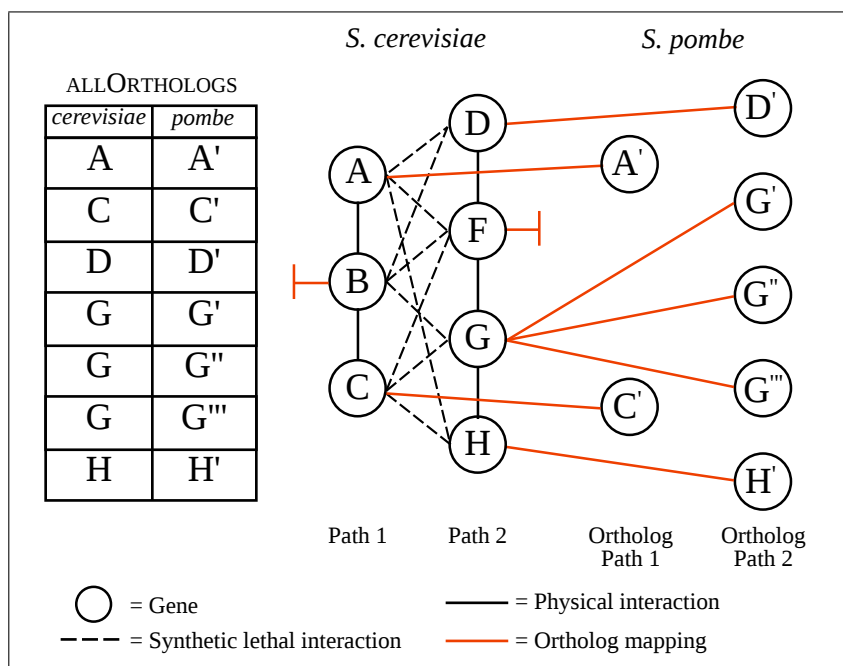


Figure 4: Theoretical ortholog mapping from *S. cerevisiae* to *S. pombe*. Orthologs are listed in the `allOrthologs` array. Note that genes B and F do not have orthologs in *S. pombe*, while gene G has three.

The similarity of one BPM to another using orthologs happens in real time, so no new memory is used up. The theoretical ortholog BPM is just a way to visualize the comparison taking place.

5.3 Determining similarity using Jaccard indexes

To determine the similarity between an original BPM and an ortholog BPM, we used a Jaccard index. A Jaccard index is a statistical tool for measuring similarity between two sets. In this case, the genes within the original and the ortholog BPM comprise each set. The Jaccard index is found by taking the intersection of the two sets and dividing it by the union of the two sets as shown below. In other words, it is the number of genes that occur in both sets, divided by the total number of genes not counting duplicates twice.

$$\text{Jaccard index} = \frac{\cap}{\cup}$$

Using a Jaccard index we can treat genes from either the entire BPM as a set, or just an individual pathway as a set. This allows us to find various similarity scores based on which pathways we choose to map over and which pathways we choose to compare the mapping against. For this research, we found Jaccard indexes for every combination of mapping and comparing between *S. cerevisiae* and *S. pombe*. This includes mapping from both pathways in one species and comparing against both pathways the other species (**Both→Both**), mapping from one pathway and comparing against both pathways (**Pathway→Both**), mapping from

both pathways and comparing against one pathway (**Both**→**Pathway**), and mapping from one pathway and comparing against one pathway (**Pathway**→**Pathway**). This is done for every BPM of the current dataset against every BPM the opposite species' dataset.

The Jaccard indexes are stored in an array named **jaccardIndexes** that is of equivalent size to the current BPM set being mapped. Again, for the purposes of explaining this concept we will assume the BPM dataset being mapped is the *S. cerevisiae*, **cereBPMs**, and it is being compared to the *S. pombe*, **pombeBPMs**. Each index in the **jaccardIndexes** array indicates the *S. cerevisiae* BPM number for which the Jaccard indexes are found. At each index is another large array of size **pombeBPMs**. The index of the subarray represents the *S. pombe* BPM number that the *S. cerevisiae* BPM is being compared against. Finally, at each index of the subarray is a final array containing all the Jaccard indexes found between the two BPMs. The following shows the organization of the final array. The each box represents a unique Jaccard index value, where the first letter stands for the set of genes mapped from the *S. cerevisiae* BPM and the second letter stands for the set of genes compared against in the from *S. pombe* BPM. B stands for the set of both pathways in the BPM, P1 stands for a set of just pathway 1, and P2 stands for a set of just pathway 2. For example, B→B is the Jaccard index found when mapping both *S. cerevisiae* BPM pathways and comparing against both *S. pombe* BPM pathways, P1→P2 is the Jaccard index found when mapping pathway 1 of the *S. cerevisiae* BPM and comparing to pathway 2 of the *S. pombe* BPM pathway, etc.

```
finalArray =
    { [B→B], [B→P1], [B→P2], [P1→B], [P2→B], [P1→P1], [P1→P2], [P2→P1], [P2→P2] }
```

The top results for each of these scores are found and displayed with the correlating *S. cerevisiae* and *S. pombe* BPM numbers. These results can be found in Supplement 2-a and Supplement 2-b.

5.4 Validating conserved BPMs with GO enrichment

Those BPMs shown to have high Jaccard indexes are thought to be conserved across species. As with the original BPM papers, we can validate the conserved BPMs using GO enrichment[1, 2, 3, 4]. If compared pathways receive high Jaccard indexes, then it is known that many genes within each pathway are evolutionarily related. However, it does not necessarily show *functional* conservation. Checking the GO enrichment of pathways with high Jaccard indexes allow us to see if these pathways are found in literature to be functionally conserved as well.

On the site <http://bpms.herokuapp.com/bpms>, you can navigate to a particular BPM and view its gene pathways and protein-protein interaction information. Underneath each pathway, if the pathway is GO enriched, it will say so with a link to the GO enriched terms for the set. A pathway is considered GO enriched if the GO term has a depth of at least three in the hierarchy and the pathway has a p-value of ≤ 0.01 . After running the program to find pairs of BPMs with high Jaccard indexes, we went to the BPM website and checked the GO enrichment for each. If conserved pathways are GO enriched for similar functions with low p-values, this is a more clear indication of conservation.

6 Results

6.1 Similarity scores

Running the program on every BPM dataset found many pathways to be conserved across *S. cerevisiae* and *S. pombe*. Supplement 2-a and 2-b contain the results for running the program on InParanoid and PHOG ortholog data respectively. The results provide the top similarity scores for each type of pathway mapping and comparing from *S. cerevisiae* to *S. pombe* (the types of mapping-comparisons being **Both**→**Both**, **Pathway1**→**Pathway2**, etc, as described in part 3 of the Methods section). Next to the Jaccard index score are the BPM numbers for the *S. cerevisiae* and *S. pombe* BPMs. These BPMs can be found on the website, <http://bpms.herokuapp.com/bpms>.

For the purposes of this study, a Jaccard index ≥ 0.10 is considered to indicate conservation. Studies have shown the *S. cerevisiae* and *S. pombe* PPI networks data to conserve only 17.3% of synthetic lethal interactions[19]. Furthermore, the *S. cerevisiae* and *S. pombe* genomes diverged between 300 and 400 million years ago[20] and PPI networks have been shown to have a high rate of rewiring in eukaryotic species[21]. These numbers suggest that a high percentage of the physical and synthetic lethal interactions that form the structure of the BPMs will not be found in both *S. cerevisiae* and *S. pombe*. Therefore a relatively low threshold of 0.10 Jaccard index score is enough to indicate conservation for the purposes of this study. Even so, many BPMs were shown to be conserved with Jaccard indexes much higher than the 0.10 cutoff. The example in Figure 5 shows the I. Ulitsky and R. Shamir[2] BPM 22 mapping to Roguev *et al*[5] BPM 80. Mapping from both pathways of *S. cerevisiae* and comparing to both pathways of *S. pombe*, the BPMs were shown to be conserved with a Jaccard index of 0.2727. Mapping from pathway 1 of *S. cerevisiae* and comparing to both pathways of *S. pombe*, the Jaccard index produced a score of 0.375.

6.2 Conservation of BPM pathways across *S. cerevisiae* and *S. pombe*

A total of 1279 pathways were found to be conserved (Jaccard index ≥ 0.10) when mapping pathways from *S. cerevisiae* and comparing to *S. pombe* using the InParanoid[15] and PHOG[16] ortholog data sets. 707 pathways were conserved using the InParanoid ortholog data and 572 pathway were conserved using the PHOG ortholog data. These numbers can not be made into a percentage of the total number of BPM pathways because they include *S. cerevisiae* pathways that are conserved in many *S. pombe* pathways and vice-versa. For example, the full BPM of R. Kelly and T. Ideker[1] BPM 305 maps with a high Jaccard index to pathway 2 of A. Roguev *et al*[5] BPMs 11, 13, 55, 69, 86, 116, and 138. This increases the number of **Both**→**Pathway2** mappings by 7 even though there is only 1 *S. cerevisiae* BPM being mapped- in order to find the percentage, we would have to divide by all of the potential mappings, the size of the *S. cerevisiae* dataset times the size of the *S. pombe* dataset. This data is not useful to us as it is comparing the results to *every* pathway in *S. cerevisiae* being conserved in *every* pathway in *S. pombe*, which is impossible in nature. More useful than this is the percentage of *unique* pathways that are conserved across species and out of those pathways, how many are GO enriched.

Table 1 shows the number of unique *S. cerevisiae* pathways that are found to be conserved

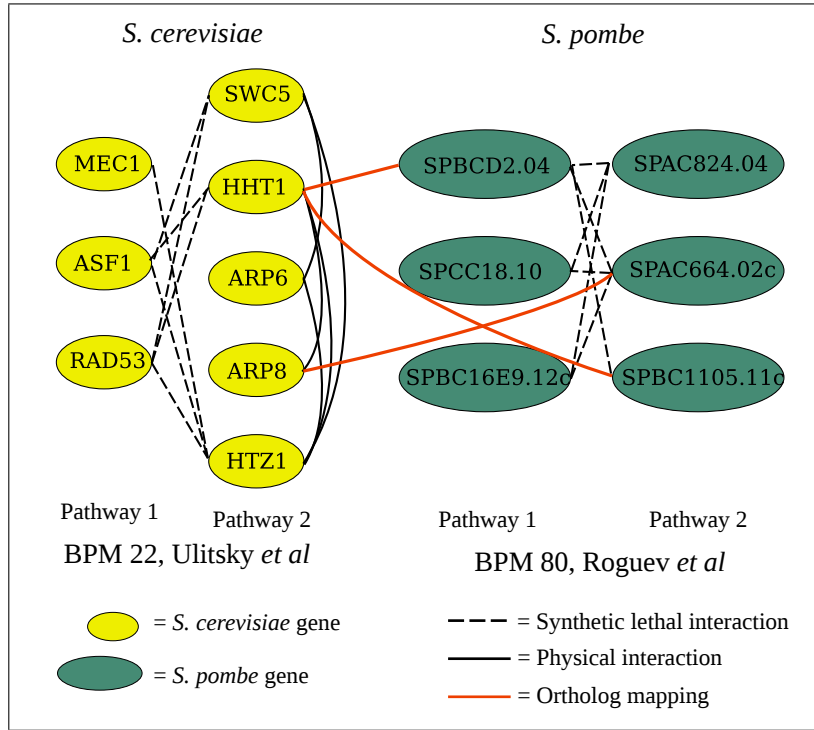


Figure 5: Example BPMs found to be conserved across *S. cerevisiae* and *S. pombe*. Mapping both pathways of Ulitsky 22 and comparing to both pathways of Roguev 80 found a Jaccard index of $3/12 = 0.25$. Mapping pathway 2 of Ulitsky 22 and comparing to both pathways of Roguev 80 found a Jaccard index of $3/8 = 0.375$.

in *S. pombe*. There are a total of 4706 unique pathways in all of the BPM datasets used in this research. This is found by taking the total number of BPMs and multiplying by two because each BPM contains two pathways. Of these unique pathways, 182 were found to be conserved when mapped and compared in some way to *S. pombe*. This shows that only 3.87% of the *S. cerevisiae* BPM pathways are conserved in the *S. pombe* BPMs.

The low percentage of conservation can be attributed to many factors. As mentioned above, studies have shown that only 17.3% of synthetic lethal interactions are conserved between *S. cerevisiae* and *S. pombe*[19]. Additionally, *S. cerevisiae* and *S. pombe* are evolutionarily distant[20] and the interactome has been shown to have a fast rate of rewiring[21]. Furthermore, the BPM datasets are bloated and noisy. This can be seen in the sheer number of BPMs found in the *S. cerevisiae* interactome. In the entire genome, *S. cerevisiae* contains 6275 genes[22]. The number of BPMs used from the datasets in this study totals 2353. If we treat each BPM as a true biological structure, that is saying that there are a total of 4706 *pathways* of genes, each performing a redundant function, in a genome of only 6275. This simply can not be true. Considering all of these factors, the low rate of BPM pathway conservation found in this study is to be expected.

	InParanoid data			PHOG data		
	Unique pathways conserved	Total pathways	Percent conserved	Unique pathways conserved	Total pathways	Percent conserved
Brady	99	3020	3.28	55	3020	1.82
Collins	3	720	0.42	7	720	0.97
KI	41	178	23.03	50	178	28.09
Ma	27	280	9.64	64	280	22.86
Ulitsky	12	508	2.36	8	508	1.57
Total	182	4706	3.87	184	4706	3.91

Table 1: Unique pathways found to be conserved when mapping from *S. cerevisiae* and comparing to *S. pombe*.

6.3 Validation of BPMs found to be conserved

To validate the existence of their BPMs, all four BPM-finding algorithm papers used GO enrichment to measure what fraction of their BPMs exhibit functional coherence[1, 2, 3, 4]. Here, we apply the same approach to validate the BPMs we found to be conserved using our method. Table 2 shows the number of unique *S. cerevisiae* BPMs found to be conserved when mapping the full *S. cerevisiae* BPM and comparing to the full *S. pombe* BPM (a B→B mapping, as described in part 3 of the Methods section). Using the InParanoid[15] ortholog data, a total of 26 unique BPMs were found to be conserved from a B→B mapping. With the PHOG[16] ortholog data, a total of 16 unique BPMs were found to be conserved. From those unique *S. cerevisiae* BPMs, the table also shows the number of BPMs who are doubly GO enriched (GO enriched on both pathways), singly GO enriched (GO enriched on one pathway) and not GO enriched (GO enriched on neither pathway). Using the InParanoid data, 96.2% of conserved B→B mappings were found to be GO enriched. 73.1% were doubly GO enriched, 23.1% were singly GO enriched, and 3.8% were not GO enriched for either pathway. Of all the individual pathways found, 84.6% were GO enriched using InParanoid data. The PHOG data found 100% of B→B mappings to be GO enriched. 68.8% of these were doubly GO enriched and 31.3% were singly GO enriched. Of all the pathways found conserved using PHOG ortholog data, 84.3% were GO enriched

84% GO enrichment is an extremely strong indicator that the BPMs found to be conserved are functionally coherent. To put these numbers in perspective, we can compare them to the GO enrichment validation statistics of BPMs found by the original papers. Table 3 shows the GO enrichment results for BPMs identified by Kelly and Ideker[1], Ulitsky and Shamir[2], and Brady *et al*[4]. These results were compiled by Brady *et al* to use an equal definition of GO enrichment. As stated in Brady *et al*[4], the BPM generation of Ma *et al*[3] bias the pathway samples and not enough of their data is available for comparison, so they were omitted from the table.

The GO enrichment results show Kelley and Ideker BPMs and Ulitsky and Shamir BPMs to receive GO enrichment on about 35% of their pathways. Brady *et al* BPMs found 50.6% of their pathways to be GO enriched. Of course, these statistics are not meant for a direct

InParanoid data					
	Unique BPMS conserved B->B	Enriched (Doubly, Singly)	Unique pathways conserved B->B	Enriched pathways	Percent enriched
Brady	7	7 (4, 3)	14	11	78.57
Collins	3	3 (2, 1)	6	5	83.33
KI	10	10 (9, 1)	20	19	95.00
Ma	2	2 (2, 0)	4	4	100.00
Ulitsky	4	3 (2, 1)	8	5	62.50
Total	26	25 (19, 6)	52	44	84.62

PHOG data					
	Unique BPMS conserved B->B	Enriched (Doubly, Singly)	Unique pathways conserved B->B	Enriched pathways	Percent enriched
Brady	0	0 (0, 0)	0	0	0.00
Collins	7	7 (3, 4)	14	10	71.43
KI	8	8 (7, 1)	16	15	93.75
Ma	0	0 (0, 0)	0	0	0.00
Ulitsky	1	1 (1, 0)	2	2	100.00
Total	16	16 (11, 5)	32	27	84.38

Table 2: Unique BPMS conserved when mapping both pathways in *S. cerevisiae* and comparing to both pathways in *S. pombe*. The results show the number of BPMS that are doubly enriched (enriched on both pathways) and singly enriched (enriched on one pathway). The final column shows the percentage of all the pathways in the BPMS that are found to be enriched.

	Enriched pathways	Total pathways	Percent enriched
Kelly and Ideker	251	720	34.86
Ulitsky and Shamir	100	280	35.71
Brady <i>et al</i>	1528	3020	50.60

Table 3: GO enrichment statistics for BPMS identified in previous papers.

comparison between the conserved BPMs and the set of BPMs generated by the original papers. The numbers represent the GO enrichment statistics of the source BPMs used for detection. From these sets of BPMs, we detected conservation on pathways with over 84% GO enrichment.

Beyond GO enrichment, further validation of BPM conservation can be found by checking the literature for conserved biological pathways. For example, pathway 2 of Brady BPM 1051 was found to map to pathway 1 of Roguev BPM 1 with a Jaccard index score of 0.1739. Each pathway is GO enriched with very low p-values ($< 1.9^{-5}$ for both) as a DNA damage checkpoint. The enriched GO term is defined as: “A signal transduction pathway, induced by DNA damage, that blocks cell cycle progression (in G1, G2 or metaphase) or slows the rate at which S phase proceeds”[13]. Lab studies researching genes in this pathway have found them to be conserved not only across *S. cerevisiae* and *S. pombe*, but all the way across the mouse and human genomes as well[23, 24, 25]. Bluysen *et al*[23] states that for the function described in the GO enriched term, *S. pombe* requires six genes, $rad1^+$, $rad3^+$, $rad9^+$, $rad17^+$, $rad26^+$, and $hus1^+$. Three of these genes $-rad1^+$, $rad17^+$, and $rad26^+$ are found in pathway 1 of Roguev BPM 1. Bluysen *et al* also states that *S. cerevisiae* requires the seven genes RAD9, RAD17, RAD24, DDC1, MEC1, MEC3, and RAD5 for the same function. The five genes RAD9, RAD17, RAD24, DDC1, and RAD5 are all found in pathway 2 of Brady BPM 1051. Furthermore, genes from these complexes contain orthologs in mice and humans[23, 24, 25]. Just as in *S. cerevisiae* and *S. pombe*, three of these genes have been found in humans to be central components of a DNA damage-response complex[25]. The ability of this study to detect conservation of BPMs found to be conserved all the way across mouse and human indicates significant validation of our result.

7 Discussion

The purpose of this research was to develop a method to detect BPM pathway conservation across *S. cerevisiae* and *S. pombe* yeast. Using the technique described here, we found conservation of BPMs based on similarity scores when mapping gene orthologs from one species to another. The method detected conservation in many pathways from *S. cerevisiae* BPMs in the given datasets to BPMs found in *S. pombe*.

Checking the GO enrichment of BPMs gives us a method of validating the functional coherence of pathways found to be conserved using our method. The fact that we are able to pull out a higher percentage of GO enriched pathways than the BPM-identifying papers shows the ability of this conservation detection method to identify functionally coherent pathways. Examples such as the Brady 1051 and Roguev 1 BPM pathways being found in literature to be conserved across mouse and human genomes further validate this claim.

Still, continued study of BPM conservation across *S. cerevisiae* and *S. pombe* is required. As BPM detection algorithms are refined and new interaction data is produced, more accurate BPMs can be identified. With this new data, we expect to see better conservation of BPMs and higher similarity scores across species. Furthermore, extended coverage of E-MAP data to additional species would provide the means to detect novel BPMs. This new dataset would allow comparisons across more species and provide stronger evidence of BPM conservation.

8 Acknowledgments

I would like to thank my thesis advisers, Ben Hescott and Lenore Cowen, who introduced me to computational biology research and first encouraged me to pursue a Senior Honors Thesis. Thanks to Donna Slonim, Anselm Blumer, and the rest of the Tufts Bioinformatics and Computational Biology group for hosting weekly computational biology talks throughout the year. Thanks to Ruchira Datta from PHOG who uploaded the specific *S. cerevisiae* to *S. pombe* ortholog data at my request. I would also like to extend a special thanks to Max Leiserson, who not only built the BPM website that proved instrumental to this study, but continually gave me advice through my research and thesis writing.

References

- [1] Kelly, R., Eked, T., *Systematic interpretation of genetic interactions using protein networks*, Nature Biotechnology, **23** (2005), 561-566.
- [2] Alida's I., Chair, R., *Pathway redundancy and protein essentiality revealed in the Saccharomyces cerevisiae interaction networks*, Molecular Systems Biology, **3** (104), (2007).
- [3] Ma, X., Tarone, A., Li, W., *Mapping Genetically Compensatory Pathways from Synthetic Lethal Interactions in Yeast*, PLoS ONE, **3** (4), (2008).
- [4] Brady, A., Maxwell, K., Daniels, N., Cowen, L.J., *Fault Tolerance in Protein Interaction Networks: Stable Bipartite Subgraphs and Redundant Pathways*, PLoS ONE, **4** (4), (2009).
- [5] Roguev, A., Wiren, M., Weissman, J.S., Krogan, N., *High-throughput genetic interaction mapping in the fission yeast Schizosaccharomyces pombe*, Nature Methods, **4** (10), (2007), 861-866.
- [6] Leiserson, M., Tatar, D., Cowen, L., Hescott, B., *Inferring Fault Tolerance from E-MAP Data*, Abstract in: Proceedings of the 2009 RECOMB-MIT Conference, (2009), 298
- [7] Jordan, K., Rogozin, I.B., Wolf, Y.I., Koonin, E.V., *Essential Genes Are More Evolutionarily Conserved Than Are Nonessential Genes in Bacteria*, Genome Research, **12** (2002), 962-968.
- [8] Giaever, G., *et al*, *Functional profiling of the Saccharomyces cerevisiae genome*, Nature, **418** (2002), 387-391.
- [9] Tong, A.H.Y. *et al*, *Systematic Genetic Analysis with Ordered Arrays of Yeast Deletion Mutants*, Science, **294** (2001), 2364-2368.
- [10] Stark, B., Beitzkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., Tyers, M., *BioGRID: a general repository for interaction datasets*, Nucleic Acids Research, **34** (2006), D535-D539.
- [11] Collins, S.R., *et al*, *Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map*, Nature, **446** (2007), 806-810.

- [12] Hescott, B.J., Leiserson, M.D.M., Cowen, L.J., Slonim, D.K., *Evaluating Between-Pathway Models with Expression Data*, Proceedings of the 13th Annual International Conference on Research in Computational Molecular Biology, (2009), 372-385.
- [13] The Gene Ontology Consortium, *Gene Ontology: tool for the unification of biology*, Nature Genetics, **25** (2000), 25-29.
- [14] Berriz, G.F., King, O.D., Bryant, B., Sander, C., Roth, F.P., *Characterizing gene sets with FuncAssociate*, Bioinformatics, **19** (2003), 2502-2504.
- [15] Ostlund, G., Schmitt, T., Forslund, K., Kostler, T., Messina, D., Roopra, S., Frings, O., Sonnhammer, E., *InParanoid 7: new algorithms and tools for eukaryotic orthology analysis*, Nucleic Acids Research, **38** (2010), D196-D203
- [16] Datta, R.S., Meacham, C., Samad, B., Christoph Neyer, Kimmen Sjlander, *Berkeley PHOG: PhyloFacts orthology group prediction web server*, Nucleic Acids Research, **37** (2009), W84-W89.
- [17] EnsemblGenomes, *Ensembl Fungi release 4*, EMBL-EBI, <http://fungi.ensembl.org>, (Feb 2010).
- [18] Jeong, H., Mason, S.P., Barabasi, A.-L., Oltvai, Z.N., *Lethality and centrality in protein networks*, Nature, **441** (2001), 41-42.
- [19] Roguev, A. *et al*, *Conservation and Rewiring of Functional Modules Revealed by an Epistasis Map in Fission Yeast*, Science, **322** (2008), 405-410.
- [20] Sipiczki, M., *Where does fission yeast sit on the tree of life?*, Genome Biology, **1** (2), reviews1011.1-1011.4.
- [21] Beltrao, P., Serrano, L., *Specificity and Evolvability in Eukaryotic Protein Interaction Networks*, PLoS Computational Biology, **3** (2), (2007), 258-267.
- [22] Cherry, J.M., Ball, C., Weng, S., Juvik, G., Schmidt, R., Adler, C., Dunn, B., Dwight, S., Riles, L., Mortimer, R.K., Botstrin D., *Genetic and physical maps of Saccharomyces cerevisiae*, Nature, **387** (1997), 67-73.
- [23] Bluysen, H.A.R., Os, R., Naus, N.C., Jaspers, I., Heoijmakers, J.H.J., Klein, A., *A Human and Mouse Homolog of the Schizosaccharomyces pombe rad1⁺ Cell Checkpoint Control Gene*, Genomics, **54** (1998), 331-337.
- [24] Freire, R., Murguia, J.R., Tarsounas, M., Lowndes, N.F., Moens, P.B., Jackson, S.P., *Human and mouse homologs of Schizosaccharomyces pombe rad1⁺ and Saccharomyces cerevisiae RAD17 linkage to checkpoint control and mammalian meiosis*, Genes & Development, **12** (1998), 2560-2573.
- [25] Volkmer, E., Karnitz, M.E., *Human Homologs of Schizosaccharomyces pombe Rad1, Hus1, and Rad 9 Form a DNA Damage-responsive Protein Complex*, The Journal of Biological Chemistry, **274** (2), (1999), 567-570.