2013

Identifying Pathway Proteins in Networks using Convergence

Kathryn Dempsey Cooper
University of Nebraska at Omaha, kdempsey@unomaha.edu

Hesham Ali
University of Nebraska at Omaha, hali@unomaha.edu

Recommended Citation

Cooper, Kathryn Dempsey and Ali, Hesham, "Identifying Pathway Proteins in Networks using Convergence" (2013). *Interdisciplinary Informatics Faculty Proceedings & Presentations*. 24.
https://digitalcommons.unomaha.edu/interdiscipinformaticsfacproc/24

Identifying Pathway Proteins in Networks using Convergence

Kathryn Dempsey and Hesham Ali

Department of Pathology & Microbiology, University of Nebraska Medical Center

College of IS&T, University of Nebraska at Omaha

Email: hali@unomaha.edu

ABSTRACT

One of the key goals of systems biology concerns the analysis of experimental biological data available to the scientific public. New technologies are rapidly developed to observe and report whole-scale biological phenomena; however, few methods exist with the ability to produce specific, testable hypotheses from this noisy 'big' data. In this work, we propose an approach that combines the power of data-driven network theory along with knowledge-based ontology to tackle this problem. Network models are especially powerful due to their ability to display elements of interest *and* their relationships as internetwork structures. Additionally, ontological data actually supplements the confidence of relationships within the model without clouding critical structure identification. As such, we postulate that given a (gene/protein) marker set of interest, we can systematically identify the core of their interactions (if they are indeed working together toward a biological function), via elimination of original markers and addition of additional necessary markers. This concept, which we refer to as "convergence," harnesses the idea of "guilt-by-association" and recursion to identify whether a core of relationships exists between markers. In this study, we test graph theoretic concepts such as shortest-path, *k*-Nearest-Neighbor and clustering) to identify cores iteratively in data- and knowledge-based networks in the canonical yeast Pheromone Mating Response pathway. Additionally, we provide results for convergence application in virus infection, hearing loss, and Parkinson's disease. Our results indicate that if a marker set has common discrete function, this approach is able to identify that function, its interacting markers, and any new elements necessary to complete the structural core of that function. The result below find that the shortest path function is the best approach of those used, finding small target sets that contain a majority or all of the markers in the gold standard pathway. The power of this approach lies in its ability to be used in investigative studies to inform decisions concerning target selection.

General Terms

Algorithms, Theory, Networks, Pathways, Validation.

Keywords

Graph theory, biological networks, convergence, ontology.

1. INTRODUCTION

In systems biology, a high-throughput experiment generally is initiated as an investigatory study or to examine a specific cellular response. Although there exists a wealth of data currently available through open sourcing, it is often possible to lose the best targets for study from a holistic experiment in the "noise" generated by the study. This "noise" rises from the complexity of the majority of biological systems, and can impede the selection of optimal biological targets by offering multiple 'interesting' results from a cellular genomic survey. This is further complicated by the pleiotropic nature of genes; for example, one study found that almost half of all known genes can be found in multiple pathways[1]. This complexity, combined with multiple processes undertaken by a cell at a given time (housekeeping gene activity, metabolism, and other homeostatic functions) can cloud systems biology experimental analyses either as *noise* or by revealing themselves as functionally enriched (and therefore interesting) results. This is particularly prevalent in investigatory or "fishing" studies – a systematic approach using biological networks, for example, may reveal 'interesting' network substructures such as hub nodes[2] and clusters[3], but these results could be largely an artifact of the holistic nature of the experiment. As such, when performing these studies, it can be often helpful to have a set of "seed" genes, proteins, or gene products that are able to inform the selection of targets from the analysis.

By contrast, if an experiment returns a list of gene products or proteins with potential impact in the domain at hand, the connection between these "markers" – biological or otherwise- is not always readily apparent. It is in these cases where systems biology can be particularly useful, particularly network systems biology. We have developed a method that, given an input set of seeds or "markers," will return a set of target nodes *T* that describe the core function of those markers (if it exists). Further, using this approach, we can identify which original markers to exclude or include from the target set, and we can also identify which targets are best to include via recursion, based on graph theory. Particularly, it is known that the majority of proteins perform their functions as complexes[4]; In protein-protein interaction networks, protein complexes are likely to be found as cliques (complete subgraphs, where for some group of nodes *n*, all possible interactions between all nodes in the network exist) or as semi-cliques, where almost all possible interactions between all nodes in the group exist. In this way, density can be used to identify proteins that work together for some function[5,6]. Further, it stands to reason that if a group of proteins exist together in a typical pathway, there is going to be interaction between those proteins that result in high density subgroups when represented in a protein protein interaction network, as shown in the example in Figure 1. In this example, there are three complete cliques in the hypothetical protein-protein interaction network, a K_4 (Ghi, Mno, Blok1, Blok2), a K_3 (Ghi, Jkl, and Mno), and another K_3 (Abc, Def, and Ghi). These all have edge densities of 100%. Further, the hypothetical proteins in these clusters are shared between cliques – Ghi in all three cliques, and Mno in two. These three cliques

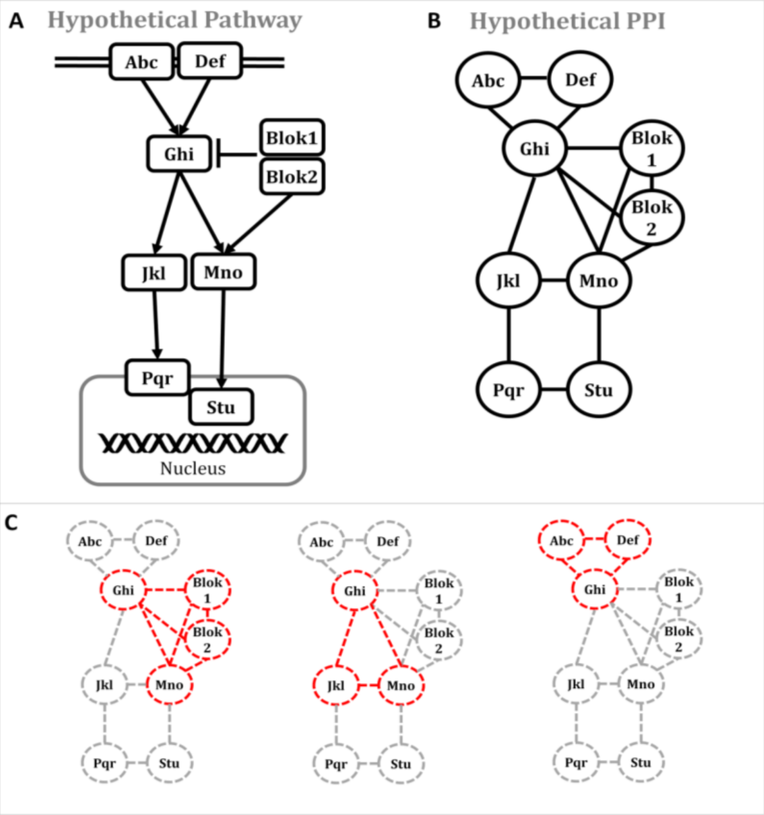


Figure 1. A hypothetical pathway and its hypothetical protein-protein interaction network. (A) The hypothetical signaling pathway, which begins at membrane receptors and signals transcription in the nucleus. (B) The corresponding PPI displaying hypothetical protein names and their binary interactions (if they interact at all, there is an edge between them. If they do not, there is no edge). (C) The three cliques formed by the protein-protein interaction network, a K_4 (left) and two K_3's (center, right). In reality, we expect these proteins to have higher intraconnection (all nodes in the network are more connected than in this example) but lower overall density (not all clusters will be 100% complete graphs).

combined contain 7 nodes and 11 edges, for an edge density of 52.38%. Other combinations of these cliques, for example, the K_4 and the K_3 containing Ghi, Jkl, and Mno, contain 5 nodes and 8 edges, for an edge density of 80%. Thus, density can be an indicator of nodes working together toward a common function in a pathway in a protein-protein interaction network[7,8].

In this study, we present our method that uses a graph theoretic method to identify new targets to add to the input markers. The graph theoretic methods used are *k*-Nearest-Neighbor, All Pairs Shortest Paths, and clustering. These methods are used identify new targets are briefly described here and explained in detail in the Model section. Previous work using shortest paths to identify new nodes from a set of input markers has shown promising results in Alzheimer's disease[9]. For example, if we have a pair of markers *i* and *j*, we can identify the shortest path between them. If the shortest path between them is of length 1, this indicates that *i* and *j* are already neighbors. If the shortest path length is greater than 1, we add the nodes on the path between *i* and *j* as new targets. By adding these nodes, we improve the overall shortest path length of our original marker set. Adding targets via the *k*-Nearest-Neighbor is a straightforward approach. For example, if *k* = 1, only the direct neighbors are added to the new target set. If *k* = 2, the neighbors of the original set are added, and then the neighbors of those nodes are added in, and so on and so forth. This method is a straightforward way to add the closest "associates" of the original markers. Finally, the cluster approach is more traditional: after clustering the network, if any or all of the markers are contained in one or more small, dense clusters, the nodes contained in that cluster that are not in the original marker set become new targets for the set.

To measure the impact of adding new targets to the original marker some global parameter of the subnetwork induced by the markers, targets, or markers and targets combined is measured. In this study, the average shortest path between markers/ targets/ markers+targets and edge density of the subgraph induced by the markers/ targets/ markers+targets is used to measure the effectiveness of adding new targets. In the case of edge density, if adding new targets results in a dense network, this is considered an improvement on the network. As such, we use edge density to determine which set (markers, targets, or markers+targets) defines the optimal subgraph connecting the original markers.

1.1 Proof of Concept

One of the best understood pathways in yeast, the mating pheromone response pathway[10], is employed in this study. The main players in the pathway are essentially all known[11], making this pathway and its components an ideal test case for the proof of concept of our application. The 25 main players in this pathway are listed in Table 1. The induced subgraph of the yeast protein-protein interaction network is shown in Figure 2. 22 of the 25 original markers were present in the network (missing: MEK, MEKK, MAPK). This network contains 100 edges, with a total possible number of edges coming to 462; this gives the induced subgraph an edge density of 43.29%, an average clustering coefficient of 71.8% and a characteristic shortest path length of 1.632.

2. Model

We present the following model to describe and test the proposed convergence approach in the yeast Pheromone Mating Response Pathway.

2.1 Data Origin and Network Creation

The 25 constant proteins named in the *Saccharomyces cerevisiae* Yeast Mating Pheromone Response pathway are listed in Table 1. The known protein-protein interaction network of the yeast proteome was downloaded from BioGrid (Release 3.2.99) on April 17, 2013. Duplicate edges and self-loops were immediately removed. 22 of the 25 original markers were present in the network (missing: MEK, MEKK, MAPK). This network contains 100 edges, with a total possible number of edges coming to 462, this gives the induced subgraph an edge density of 43.29%, an average clustering coefficient of 71.8% and a characteristic shortest path length of 1.632. The induced subgraph of these 25 yeast proteins in the known interaction network is shown in Figure 2.

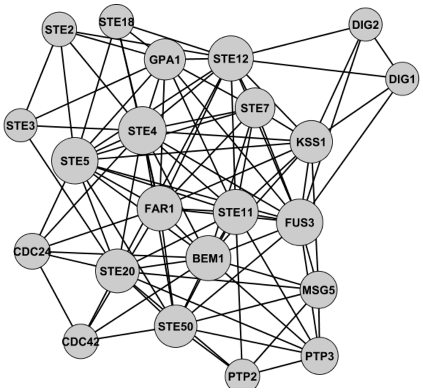


Figure 2. Induced subgraph by the *S. cerevisiae* Pheromone Response Pathway genes (listed in Table 1). Initial network parameters are: edge density = 43.29%, average clustering coefficient = 71.8%, and average shortest path length = 1.632.

2.2 Marker Set Definition

For each experiment, we define a set of markers *M* that includes the gene symbol of the protein name. If no gene symbol for a given protein exists, it is not included in the set. The set of targets *T* is the set of targets that result from the convergence for that iteration. For i teration 1, *M* is the marker set and *T* is the target set. The exit parameter (edge density or average shortest path) is defined for marker set *M* and then measured again for target set *T,* and additionally for the union of *M* and *T,* the markers+targets set. If the exit parameter improves from *M* to *T* or from *M* to *M+T*, the process iterates again. Then in iteration 2, the marker set becomes *T* or *M+T*. This iteration continues until the target set is an empty set or until the exit parameter does not improve (convergence).

For the yeast pheromone mating response case-study, we have defined three simulated datasets (listed in Table 2):

1. ***Ideal-case*:** The markers for this dataset are drawn randomly from any of the 25 original markers known to play a part in the pathway (Table 1). Markers/proteins outside this list of 25* were not used. Markers were randomly chosen using the Perl rand() function in groups of 100% (all markers in the list), 75%, 50%, 25%, and 15%, or until the minimum required amount of markers (3) was met. For each set of markers in 100%, 75%, 50%, 25%, and 15%, the number of chosen markers was rounded down. For example, using the 25 markers described here, the 75% group would technically contain 18.75 of the original 25 markers; in each case, this percentage was rounded down (in this example to 18 markers). This is ideal-case because it assumes in the input marker set, there is complete coverage of the entire pathway. *It is known that only 22 of the 25 markers are contained in the yeast protein-protein interaction network used. When selecting which markers would be included, we allowed all 25 markers to remain as candidates, as this best reflects the real-world possibility that complete studies on the interactions of some proteins will not be complete or even yet studied.
2. ***Semi-realistic:*** The markers for this data were chosen as such: ~50% (12) of the markers were chosen from the original list of markers in the pathway per grouping, and ~50% (13) of the markers were chosen from the proteins in the Yeast PPI with original marker proteins removed per grouping. Markers were randomly chosen using the Perl rand() function. This is semi-realistic because it assumes that some of the markers are valid and related to the desired studied function and some are not related.
3. ***Random:*** The markers for this data were chosen by randomly choosing proteins from the yeast PPI network. No restrictions were made in determining where the nodes came from. This set highlights the performance of the convergence method on a set of random markers from the yeast PPI.

Table 1. List of genes in the *S. cerevisiae* Pheromone Response Mating Pathway

Ste2^{6,7}	*Ste5*^{6,7}	*Dig1*^{6,7}	*Cdc24*^{6,7}	*Bem1*^{6,7}
Ste3^{6,7}	*Ste11*^{6,7}	*Dig2*^{6,7}	*Cdc42*^{6,7}	*Ptp3*^{6,7}
Ste4^{6,7}	*Ste7*^{6,7}	*Ste12*^{6,7}	*Far1*^{6,7}	*Ste20*^{6,7}
Ste18^{6,7}	*Fus3*^{6,7}	*MEKK*^{6,7}	*Ste50*^{6,7}	*Ptp2*^{6,7}
Gpa1^{6,7}	*Msg5*^{6,7}	*MEK*^{6,7}	*Kss1*^{6,7}	*MAPK*^{6,7}

Table 2. Markers for the ideal, semi-realistic, and random datasets.

	Original Markers	% Markers	# Markers
Ideal – Markers	Ste2, Ste3, Ste4, Ste18, Gpa1, Ste20, Bem1, Cdc24, Cdc42, Ste5, Ste11, Ste7, Fus3, Msg5, Ptp2, Ptp3, Far1, Dig1, Dig2, Ste12, MEKK, MEK, MAPK, Kss1, Ste50	100%	26
	Cdc42, Ste3, Far1, MEK, Gpa1, Ste4, Ptp2, Ste4, Ste20, Ste12, Ste3, Far1, Dig1, MAPK, Ste12, Ste20, Cdc24, Dig2	75%	18
	Cdc24, Ste2, Ste11, Gpa1, Ptp2, Cdc42, MEKK, Ste3, Ste12, Dig2, Ste5, Far1	50%	12
	Dig2, MAPK, Ste3, MAPK, Ste50, Fus3	25%	6
	Far1, Bem1, Cdc24	15%	3
Semi-Real – Markers	MEK, STE2, DIG1, PTP2, STE3, GPA1, MEKK, FUS3, MAPK, KSS1, FAR1, STE12, FRK1, YCL021W-A, RPL6B, PFK26, NOP9, PHB2, RPS7B, UBP8, ENA2, YPS6, YET2, RAD6, YOR214C	100%	25 (12 orig, 13 rand)
	STE11, STE5, FAR1, STE5, STE2, PTP3, STE50, STE18, MEK, ACF4, PKP2, ARB1, GEP5, TRI1, SWD1, ECM30, YKL151C, AVT6	75%	18 (9 orig, 9 rand)
	STE12, CDC42, CDC24, GPA1, FUS3, STE50, SOM1, MPS2, TOS3, RPS27A, HEH2, LAT1	50%	12 (6 orig, 6 rand
	STE12, STE50, BEM1, RPN4, FET4, MNN4	25%	6 (3 orig, 3 rand
	MSG5, QNS1, DAL81	15%	3 (1 orig, 2 rand)
Random – Markers	YPR013C, SPC25, HEM1, YLR125W, RXT3, MCD4, SHY1, XKS1, BIR1, SMD1, ATP8, AAH1, VPS30, VTC2,	100%	25

	MED8, SPT3, RTT101, YBR096W, PRP19, CDS1, ORM2, YBR053C, CAT8, FAS1, SPP382		
	SGF29, CNOT1, NCS2, DCP1, SGT2, SRB2, YKL091C, TRM8, YHR009C, CIA1, FIR1, SNN1, STE13, DFG5, AAT1, PUT2, GAP1, SUR1	75%	18
	STO1, ETS1-1, DAL82, PSP2, GCN3, RPN4, KAT2A, PHB1, ESS1, VPS13, MMS21, CAF40	50%	12
	OPT1, RPE1, PCL8, AFT1, FET4, SOG2	25%	6
	YGR130C, CCT3, RRP3	15%	3

2.3 Convergence Model

Our convergence algorithm uses recursion to identify group of relationships that link the original marker set proteins in *M*. Convergence can be achieved in two ways: by setting a stop parameter threshold, where some graph theoretic measure (such as the density of the subgraph induced by the marker or target set) defines when to stop recursion, or by setting a stop parameter condition, such as only continuing to iterate if the convergence algorithm applied to the target set results a new target set containing some or all of the original markers.

2.3.1 Algorithm with Stop Parameter Definition

For a set of markers *M* in some network N_a, identify the set of targets *T* in some network N_b using graph function *f* that satisfies the condition set by parameter *p*. We assume that *M* = the original Marker set, *m* is equal to the |*M*|, *f* is equal to (Shortest path approach | *k*NN approach | clustering approach| …), *p* is equal to (Average shortest path | Clustering coefficient |…), N_a is equal to the Network 1 (Data driven network), N_b is equal to Network 2 (Data driven or ontological network), where N_a can be equal to or disparate from N_b. *T* is the unknown.

```
G   = the subnetwork induced by M in $N_a$
p   = p(G) where p = ASP() or ED()
T   = converge(M, $N_a$, $N_b$)
function converge(M,$N_1$,$N_2$)
    T = f(M,$N_1$) where
    f = shortest_path(), knn(), or cluster()
    $G_{tmp}$ = the subnetwork induced by T in $N_2$
  $p_{tmp}$ = p($G_{tmp}$) where p = ED()
  if $p_{tmp}$ > p
    return T;
    end;
    if $p_{tmp}$ <= p
      converge(T,$N_2$,$N_1$)
 }
```

Stop Parameter. Parameter definitions given a graph *G*(*V*,*E*) where $V = (v_1, v_2, ..., v_n)$ and $E = (e_1,e_2,...,e_m)$. Thus, *n* = the number of nodes in *V* and *m* = the number of edges in *E*:

Edge density:

$$\frac{n*(n-1)}{2}*100 \qquad \text{(Equation 1)}$$

where *n* is equal to the number of nodes in *V*.

2.3.2 Convergence Function Definitions

Function definitions assume that given includes a graph *G*(*V*,*E*) where $V = (v_1, v_2, ..., v_n)$ and $E = (e_1,e_2,...,e_m)$ and a set of marker nodes *M*. Each function returns a set of targets *T*.

Shortest_path:

```
Target set T = ()
For each pair (i,j) of nodes in M where i != j
  For each possible shortest path between i,j
    sp(i,j) = the shortest path(s) between i,j
    If sp(i,j) > 1
      Add nodes on sp(i,j) to target set T
T = T - M    # Remove original markers from T
Return T
```

k-Nearest-Neighbor:

```
Target set T = M
  For (i = 1 to k)
    For each node v in M
    neighbors = all direct neighbors of v
    T = T + neighbors
T = T - M    # Remove original markers from T
Return T
```

Clustering:

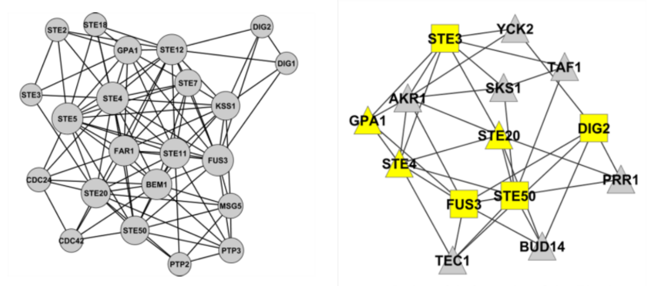
```
Target set T = ()
  C = clusters in the network
  For each cluster c in C
  If cluster c contains at least 2 nodes in M
    T = nodes in c
T = T - M    # Remove original markers from T
Return T
```

Clustering in this case was performed by MCODE v1.2 using the following parameters: Degree cutoff of 5, Haircut (ON), Node Score cutoff of 0.2, K-Core of 4, and Max. Depth of 10. Clusters were exported if they had a density cutoff of 50% or more.

3. Hypothesis

Using the ideal 25% dataset and shortest path convergence approach described above as an example, a preliminary example the ability of the convergence is presented. The ideal 25% dataset including markers and targets contains 10 targets, 3 of which are in the yeast MPR pathway, and 7 of which are not (as shown in Figure 3). The original marker set contained 6 markers from the MPR pathway, 4 of which were in the actual network. In total in the marker+target dataset, 11 proteins of the 22 possible identifiable proteins from the yeast MPR pathway were found. This from a original dataset containing only 3 proteins; highlighting the potential power of the of the convergence method.

Additional targets not in the yeast MPR pathway were found: YCK2, TAF1, SKS1, AKR1, PRR1, BUD14, and TEC1. AKR1 is associated with the yeast MPR pathway in 2 articles via PubMed search: a 2011 study from Hemsley and Grierson (which also mentions YCK2)[12], and a 1996 study from Pryciak and Hartwell[13]. BUD14 is the focus of a 2002 study in the yeast MPR pathway[14], and TEC1 is associated with 13 articles related in the yeast MPR pathway via PubMed search with the terms protein_name + “yeast mating pheromone response”. So while all of the proteins are not directly involved in the pathway, at least 4 of the 7 targets identified have been associated with the pathway in literature. This phenomenon is mentioned by Li *et al.*; that within a network, often it is not *only the complexing proteins* that are captured by a network but also the entire cohort of proteins involved in that function, informally termed a “module.[15]”



Induced Subgraph of the 22 yeast Mating Response Pathway markers

Induced Subgraph of the Final Converged network of the Ideal 25% markers

Figure 3. (Left) The induced subgraph in the *Saccharomyces cerevisiae* protein protein interaction network by the 22 existing nodes of the 25 in the yeast MPR pathway. (Right) The induced subgraph of the 14 markers and targets (4 markers, 10 targets) ideal 25% dataset found by the shortest path convergence function. Triangle nodes represent targets and square nodes represent markers. Yellow nodes represent those that are in the yeast Mating Pheromone Response Pathway. Seven additional non-pathway targets were found: YCK2, TAF1, SKS1, AKR1, PRR1, BUD14, TEC1.

Based on the concepts described above, we propose our hypothesis H_0: *If a group of biological elements are part of a pathway or functional biological module, then beginning with a large subset of these proteins/gene products, the proposed convergence approach will leads to the identification of the other members of the pathway or module.* To test this hypothesis in ideal and real world settings, we use the datasets and functions described above to test this hypothesis. The experiments will also be used to specify what a "large subset" is, or how big a component of the group is needed to identify the entire set.

4. Experiments & Results

4.1 Experimental Study

To test the hypothesis described above, we performed an array of experiments that reveal the effectiveness of the convergence approach: 1. Comparing converged versus non converged networks to determine if the stop parameter is the best measure of a target set, 2. Analyzing the number/percentage of targets found by each method to determine the effectiveness of each method, and 3. Analyzing the number of targets found by each method that are *not* part of the 25-component yeast MPR pathway. We compare these results in the ideal case and also under real world conditions.

4.2 Ideal case

All results in this section describe the "Ideal" case dataset.

4.2.1 Markers versus Markers+Target Set

Table 3 describes the number/percentage of targets found that were in the yeast MPR pathway in the ideal case for each described count of markers using the shortest path convergence approach. Examining only the targets does not offer a full point of view on the performance of the convergence approach as some of the proteins are contained in the marker set. Combining the marker and target sets, we find that using as few as 50% of the markers in the original marker set will yield at least 80% of the total proteins in the pathway; even using 25% of the pathway markers finds at least half of the proteins in the yeast MPR pathway. This reflects the power of the convergence approach.

Table 3. Target set: Targets only or Markers+Target Set. % Total Markers/Markers: The number of markers used in the original marker set. # Targets in MPR pathway: The number of total markers (of 22 possible) found in the target set in the MPR pathway. % Targets in MPR Pathway: The percentage of targets found in the yeast MPR pathway (out of possible 22).

Target Set	% Total Markers	Markers	# Targets in MPR Pathway	% Targets in MPR Pathway
Targets Only	**100%**	**26**	**0**	**0.00%**
	75%	18	7	31.82%
	50%	12	6	27.27%
	25%	6	3	13.64%
	15%	3	0	0.00%
Markers +Targets	100%	26	22	100.00%
	75%	18	20	90.91%
	50%	12	18	81.82%
	25%	6	11	50.00%
	15%	3	3	13.64%

4.3 Real-world applications

All results in this section compare Ideal vs. Semi-Real vs. Random cases.

4.3.1 Converged vs. non-converged networks

To determine if there was a difference between the accuracy of converged versus non-converged networks, we compare the percentage of yeast Mating Response Pathway genes found in converged or final networks versus non-converged, or non-final networks. For example, if an experiment had 4 iterations before converging, this means that there are 4 sets of markers and 4 sets of targets. In this example, this indicates that the subgraph induced by the target set of iteration 1 had a better stop parameter (e.g. edge density) than the subgraph induced by the marker set of iteration 1, and so on. The last iteration would then occur by the subgraph induced by target 4 set having a worse stop parameter (e.g. edge density) than the subgraph induced by marker set 4, which is the same as target set 3. Thus, the converged network in this case uses induced subgraph of the proteins in target set 3, and the non-converged networks use the induced subgraphs of the proteins in target sets 1 and 2. Target set 4 is not included because it is not an improvement on target set 3 and thus is not part of the converged network. In Figure 4 we show the distribution of the percentages of markers found in each converged or non-converged network. The x-axis represents the percent of pathway markers found, or, for all converged or non-converged networks, each network is counted as containing 0% of the total pathway markers, 1-10%, and so on. Counts were then normalized. The percent of pathway markers found represents the total pathway markers found out of 22, not 25, pathway markers, as only 22 of the original pathway markers were present in the protein-protein interaction network used. The y-axis represents the percentage of the converged or non-converged networks containing a specific range of pathway markers found; for example, 30% of the converged networks contained none of the pathway markers (bar 1, red).

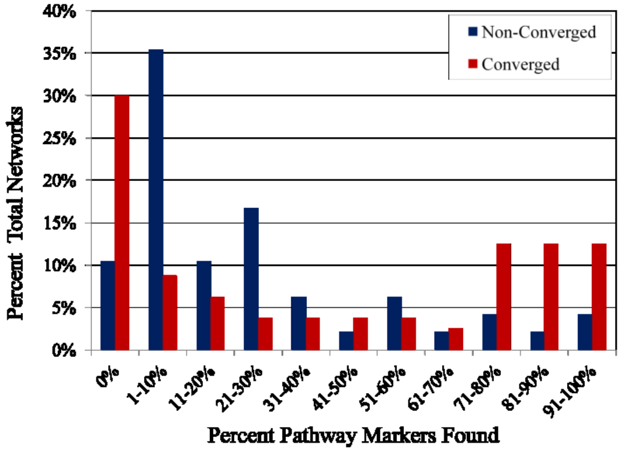


Figure 4. Pathway Markers found in Converged versus Non-converged Networks. Shown below is the distribution of the percentages of markers found in each converged or non-converged network. The x-axis represents the percent of pathway markers found, or, for all converged or non-converged networks, each network is counted as containing 0% of the total pathway markers, 1-10%, and so on. Counts were then normalized. The percent of pathway markers found represents the total pathway markers found out of 22, not 25, pathway markers, as only 22 of the original pathway markers were present in the protein-protein interaction network used. The y-axis represents the percentage of the converged or non-converged networks containing a specific range of pathway markers found; for example, 30% of the converged networks contained none of the pathway markers (bar 1, red).

The results of this comparison are interesting. The distinction between random, semi-real, and ideal cases in this chart is not made, so converged networks with no original pathway proteins in their marker sets (random case) are included, which accounts for the 30% of converged networks containing 0% of pathway markers. However, another 37.5% of the converged networks contain 71-100% of original pathway markers as compared to 10.4% of non-converged networks containing 71-100% of original pathway markers. Using a similar comparison, 79.2% of non-converged networks find 1-70% of original pathway markers compared to 32.5% of converged networks. This indicates that the convergence method may be key in allowing us to discern whether or not a set of proteins are involved in a similar pathway. For example, there were 24 converged networks where the converged network found 0% of the original pathway proteins. Of these 24, 11 found no new targets, and only 1 of these was the ideal case using 15% of the original markers. There were 9 cases where the ideal case found no yeast MPR pathway targets, and 7 of these were using the clustering approach. Because clustering is not re-run every time an iteration occurs (the input network does not change so neither does the clusters) there are often no more than 1 iteration of the cluster function convergence, and thus, no new markers are found.

4.4 Markers found vs. total targets

Percent of pathway markers found versus percent of total targets in converged networks. Figure 5 shows the comparison of performance for the three functions evaluated in the yeast MPR pathway: clusters, kNN, and shortest path. The y-axis represents the percentage of yeast MPR pathway markers found in the final converged target set (out of 22 total) and the x-axis represents the percentage of yeast MPR pathway proteins in the final converged target set. For example, if a target set is found to have 100 total proteins and 11 of those proteins are in the yeast MPR pathway, it would be located at (11%,50%). The optimal result would be for the method to identify any missing yeast MPR pathway markers not in the original dataset; this result would be located in the top right corner of the figure (most of the markers found, with those markers representing most or all of the total target set). The clustering function is the worst performer, never finding more than 10% of the pathway markers and the markers found always representing less than 5% of the total markers. The k-Nearest Neighbor approach performs well in terms of identifying pathway markers, but not by identifying non-pathway targets. Pathway markers never represent more than 5% of the total target set. The shortest path approach is varied in terms of pathway marker identification, finding more targets than the clustering function and having those markers represent up to 30% of the total targets. This however does not reflect the inclusion of markers in the original set. For example, if the marker set used is the ideal case at 100%, the target set will not contain any new targets (they are all in the marker set) and the targets will represent 0% of the total target set. Thus, the same performance combining the marker and target sets is also evaluated in Figure 6.

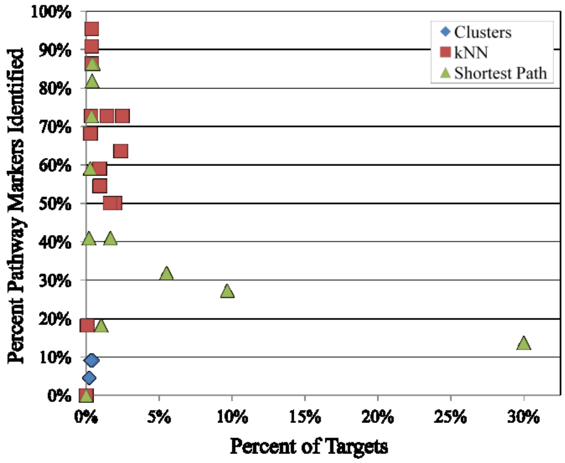


Figure 5. Percent of pathway markers found versus percent of total targets in converged networks. Figure 4 shows the comparison of performance for the three functions evaluated in the yeast MPR pathway: clusters, kNN, and shortest path. The y-axis represents the percentage of yeast MPR pathway markers found in the final converged target set (out of 22 total) and the x-axis represents the percentage of yeast MPR pathway proteins in the final converged target set. For example, if a target set is found to have 100 total proteins and 11 of those proteins are in the yeast MPR pathway, it would be located at (11%,50%). The optimal result would be for the method to identify any missing yeast MPR pathway markers not in the original dataset; this result would be located in the top right corner of the figure (most of the markers found, with those markers representing most or all of the total target set).

4.4.1 k-Nearest Neighbor function vs. Shortest Path

In real world application, generation of connections between markers would not exclude just the new targets; it could be assumed that a marker set of proteins includes interesting targets due to the inquiring scientists intimate knowledge of the topic or an experiment designed to extrapolate markers related to the subject at hand. As such, the markers should be considered with

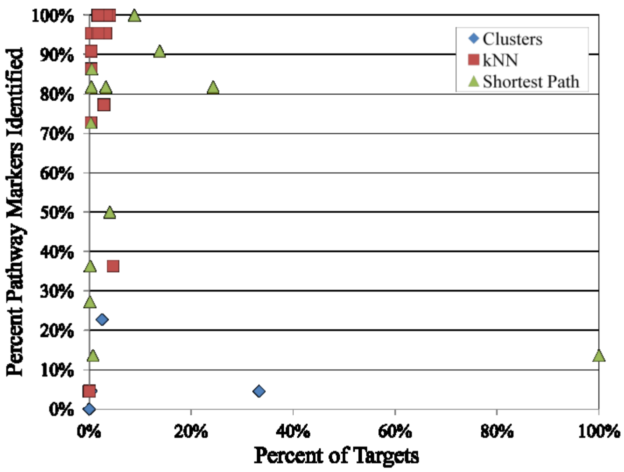


Figure 5. Percent of pathway markers found plus targets versus percent of total targets in converged networks. Figure 5 shows the comparison of performance for the three functions evaluated in the yeast MPR pathway: clusters, kNN, and shortest path. The y-axis represents the percentage of yeast MPR pathway markers found in the combined final target set and final marker set (out of 22 total) and the x-axis represents the percentage of yeast MPR pathway proteins in the final converged target set. For example, a target having 100 total proteins, 11 of which are in the yeast MPR pathway and the marker set containing 9 original markers, it would be located at (11%,90.1%). The optimal result would be for the method to identify any missing yeast MPR pathway markers not in the original dataset; this result would be located in the top right corner of the figure (most of the markers found, with those markers representing most or the total target set).

the targets when determining how well the convergence function has performed. Figure 6 shows the comparison of performance for the three functions evaluated in the yeast MPR pathway: clusters, kNN, and shortest path. The y-axis represents the percentage of yeast MPR pathway markers found in the combined final target set and final marker set (out of 22 total) and the x-axis represents the percentage of yeast MPR pathway proteins in the final converged target set. For example, a target having 100 total proteins, 11 of which are in the yeast MPR pathway and the marker set containing 9 original markers, it would be located at (11%,90.1%). The optimal result would be for the method to identify any missing yeast MPR pathway markers not in the original dataset; this result would be located in the top right corner of the figure (most of the markers found, with those markers representing most or the total target set).

While the clustering approach can be modified to include parameterization that could improve its performance, the clear winners between convergent functions are the k-Nearest-Neighbor and Shortest Path functions. Previously mentioned, Figures 5 and 6 suggest that the k-Nearest-Neighbor approach identifies the majority of yeast MPR pathway markers but identifies many other targets, while the Shortest Path approach indentifies fewer overall targets but has varied performance in terms of yeast MPR pathway identification. The percent of total targets represented by target set pathway markers for kNN (left) and Shortest Path (right) is shown in Figure 6. The k-Nearest-Neighbor approach performs poorly, where the found markers never rise above 5% of the total targets found. With target set sizes reaching up to 5,365 proteins, this would not reduce the search space for new targets at all. Even with poor performance, it becomes readily apparent that the ideal case marker set is the best performer, followed by the semi-real marker sets. Additionally, marker plus target sets perform better than target sets only for the ideal and semi-real cases, which indicates that if a marker set is indeed believed to reflect the biological markers of the function at hand, the markers should be included and considered with the new targets identified.

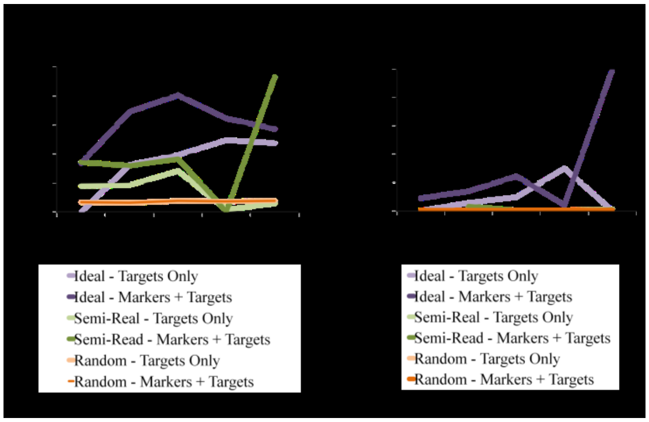


Figure 6. The percent of total targets represented by target set pathway markers for kNN (left) and Shortest Path (right). The x-axis represents the marker size used set as described above, and the y-axis represents the percentage of the total final converged target set represented by the yeast MPR pathway targets. For example, if a target set contained 100 proteins and 10 of them are yeast MPR pathway genes, it would be represented at 10%. If a target set contained 100 proteins and 10 of them are yeast MPR pathway genes and addition of a marker set with 10 yeast MPR pathway proteins was performed, it would be represented at (10+10)/(100+10) = 18.18%.

The shortest path approach performs better than the k-Nearest-Neighbor approach, but does not perform optimally. In the ideal datasets, it is the best performer, particularly when combining markers plus targets so all markers are hit. Unfortunately, the targets identified still represent only around 20% of the total targets. For the Semi-Real case, the results plummet to kNN levels, as they do with the random case. A comparison of the Semi-Real cases in kNN and Shortest Path functions appear in Figure 7; there were no targets found for the Semi-Real SP cases at 100 and 75%, and the rest of the results in general are poor performers in terms of narrowing search space.

4.4.2 Target Network Size

Figure 8 shows the sizes of the target sets of the final converged networks for the Cluster, kNN, and Shortest Path functions. Clusters clearly have the smallest, fewest target sets due to their poor performance. The shortest path methods have either very high or very low target counts, generally within the 0-1,000 range and 3,500-5,000 range. The kNN method has slightly more than the shortest path function targets, with target sets ranging in between 500-1,500 targets and 4,500 to 5,500 targets.

5. Discussion

The novel convergence approach described in this work investigates how to identify the relationships between a set of marker gene products or proteins, particularly when provided by experimental studies. Particularly, given a set of markers, the goal of the proposed approach is how to identify the relationship between them, and which additional markers or proteins need to be added to complete the picture describing their common functions, if they exist. Using a protein-protein interaction network, it is possible to find relationships between models and determine if those relationships constitute the framework for a

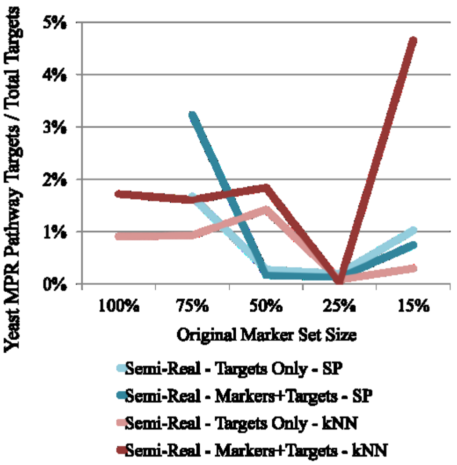


Figure 7. The percent of total targets represented by target set pathway markers for kNN (left) and Shortest Path (right) at Semi-Real only. The x-axis represents the marker size used set as described above, and the y-axis represents the percentage of the total final converged target set represented by the yeast MPR pathway targets. For example, if a target set contained 100 proteins and 10 of them are yeast MPR pathway genes, it would be represented at 10%. If a target set contained 100 proteins and 10 of them are yeast MPR pathway genes and addition of a marker set with 10 yeast MPR pathway proteins was performed, it would be represented at (10+10)/(100+10) = 18.18%.

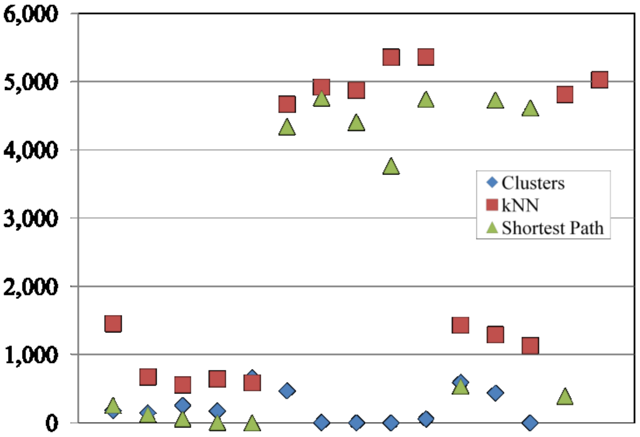


Figure 8. Target set sizes for Clusters, kNN, and Shortest Path. The y-axis represents the number of targets in the final target set of converged networks.

working cellular subsystem, or otherwise, if the markers are related originally via chance or error. Specifically, this work explores three major facets of the convergence approach: 1. defining the method of identifying targets, 2. defining the method of evaluating a final target subset, and 3. defining a stop condition or parameter for the convergence approach. The three methods used to identify new targets are basic graph theory concepts, first, the clustering approach, which adds new targets if they are found in the same cluster; second, the k-Nearest-Neighbor approach, which adds new targets that are *k*-step neighbors of the markers, and thirdly, the shortest path approach, which adds new targets on the shortest paths between markers if they are not directly connected. Our conducted experiments show that in terms of finding the most markers in a pathway while finding the least amount of incorrect proteins, the shortest path approach is optimal. Secondly, the method for identifying the constitution of the network induced by the final set of markers and targets or targets only was investigated using edge density. In this way, it has been shown that edge density can be an indicator of how well a target set predicts convergence; typically, a decrease in edge density is an indicator of the first non-appropriate iteration of the convergence approach. Also discussed was the shortest path measure, which takes the average of all shortest paths between markers and targets. This method shows theoretical promise and is planned for implementation in future work.

This convergence approach is a novel concept, and is a promising first step in using network analysis to better guide decision support for “at the bench” scientists. This work has shown that it is a viable approach to identifying new targets relating to the observed phenomenon behind designed high-throughput analyses. Indeed, as network interaction repositories continue to grow, it is hoped that so should the ability of approaches such as these to predict improved targets and cellular responses.

6. ACKNOWLEDGMENTS

The authors would like to acknowledge Ishwor Thapa for his contribution of code and insights on the shortest path convergence approach.

7. REFERENCES

1. Gu Z, Wang J. CePa: An R package for finding significant pathways weighted by multiple network centralities. *Bioinformatics*. 2013.
2. Jeong H, Mason SP, Barabasi AL, Oltvai ZN. Lethality and centrality in protein networks. *Nature*. 2001;411(6833):41-42.
3. Langfelder P, Horvath S. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics*. 2008;9:559-2105-9-559.
4. Sali A, Glaeser R, Earnest T, Baumeister W. From words to literature in structural proteomics. *Nature*. 2003 ; 422(6928):216-225.
5. Sarac OS, Pancaldi V, Bahler J, Beyer A. Topology of functional networks predicts physical binding of proteins. *Bioinformatics*. 2012;28(16):2137-2145.
6. Qi Y, Balem F, Faloutsos C, Klein-Seetharaman J, Bar-Joseph Z. Protein complex identification by supervised graph local clustering. *Bioinformatics*. 2008;24(13):i250-8.
7. Rhrissorrakrai K, Gunsalus KC. MINE: Module identification in networks. *BMC Bioinformatics*. 2011;12:192-2105-12-192.
8. Bader GD, Hogue CW. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*. 2003;4:2.
9. Managbanag JR, Witten TM, Bonchev D, et al. Shortest-path network analysis is a useful approach toward identifying genetic determinants of longevity. *PLoS One*. 2008;3(11):e3802.
10. Gustin MC, Albertyn J, Alexander M, Davenport K. MAP kinase pathways in the yeast saccharomyces cerevisiae. *Microbiol Mol Biol Rev*. 1998;62(4):1264-1300.
11. Bardwell L. A walk-through of the yeast mating pheromone response pathway. *Peptides*. 2005;26(2):339-350.
12. Hemsley PA, Grierson CS. The ankyrin repeats and DHHC S-acyl transferase domain of AKR1 act independently to regulate switching from vegetative to mating states in yeast. *PLoS One*. 2011;6(12):e28799.
13. Pryciak PM, Hartwell LH. AKR1 encodes a candidate effector of the G beta gamma complex in the saccharomyces cerevisiae pheromone response pathway and contributes to control of both cell shape and signal transduction. *Mol Cell Biol*. 1996;16(6):2614-2626.
14. Cullen PJ, Sprague GF,Jr. The Glc7p-interacting protein Bud14p attenuates polarized growth, pheromone response, and filamentous growth in saccharomyces cerevisiae. *Eukaryot Cell*. 2002;1(6):884-894.
15. Li X, Wu M, Kwoh CK, Ng SK. Computational approaches for detecting protein complexes from protein interaction networks: A survey. *BMC Genomics*. 2010;11 Suppl 1:S3-2164-11-S1-S3.