Mining of conserved motifs in protein interaction networks

Swaroop Jagadish[1] and Nikolay Laptev[2]

[1]Computer Science department, U.C. Santa Barbara
[2]Economics and Computer Science department, U.C. Santa Barbara

Abstract. Inherently, the interpretation of large-scale protein data is a computationally intensive task; however its analysis is an important undertaking that holds the potential of improving our understanding of cellular machinery. In this project we try to answer the following two questions i) how can we discover the most significant simple paths? ii) Can we identify relationships between pathways which were previously unknown? We build on techniques for simple path finding proposed by Scott et al [1] in order to do this. To test our techniques, we conduct experiments and present results for the yeast protein interaction network. We analyze the results for biological significance.

1 Introduction

Finding significant simple paths in a protein interaction network allows us to discover the important class of linear signal transduction pathways. In a protein interaction network, each edge represents an interaction between the two protein nodes. Usually a weight is assigned to the edges representing the probability that two vertices (proteins) interact.

In general, the problem of finding the optimal simple path of length K is a NP-hard problem. Scott et al have proposed a heuristic which is linear in the number of nodes and exponential in the path length. We use this to find simple paths of manageable lengths such as 4, 5 and 6. The heuristic uses a random coloring of the nodes and finds simple paths by requiring that each node in the path be of a different color. Since we are working with an annotated graph i.e. each node has been labeled with GO terms, we can take advantage of this knowledge to discover links between different pathways which were unknown before. We modify this coloring scheme such that nodes with highly similar GO terms have the same color. Essentially, we drop edges between nodes which have similar GO terms.

This paper is organized as follows: Section 2 contains description of Scott's technique which relies on color coding and dynamic programming. In Section 3 we present our modification, Section 4 gives our results and lastly in Section 5 we draw the conclusion.

2 Finding significant simple paths

Scott et al have proposed a heuristic which requires that each node in the path that we find is of a different color for a random coloring of the graph. This ensures that there are no cycles. Number of colors used to color the graph is equal to the maximum length of the path. The technique uses a dynamic programming setup to find the optimal path for a given graph coloring. The score for a path is just the product of the scores of the individual edges. The probabilities are converted into scores by using Score = - log(probability). Since the initial coloring of the graph determines which paths we find, several random colorings are tried out and paths with the best scores are chosen. The dynamic programming setup is explained below.

For each nonempty set $S \subseteq \{1,2,....,k\}$ AND
For each vertex v such that $c(v) \in S$
Let W(v,S) = minimum weight of a simple path of length |S| that starts within I visits a vertex of each color in S, and ends at v. I is the start set

We use the following recurrence to tabulate our dynamic programming matrix

$$W(v,S) = \min_{u:c(u)\in(S-\{c(v)\})} W(u, S-\{c(v)\}) + w(u,v), |S| > 1$$

where $W(v,\{c(v)\}) = 0$ if $v \in I$ and ∞ otherwise.

Thus by considering all pairs of v,S such that |S|=k, the optimal path v is the minimum of W(v,{1,2,….,k})

The running time is $O(2^k km)$ and space requirement is $O(2^k n)$

3 Finding interaction between pathways

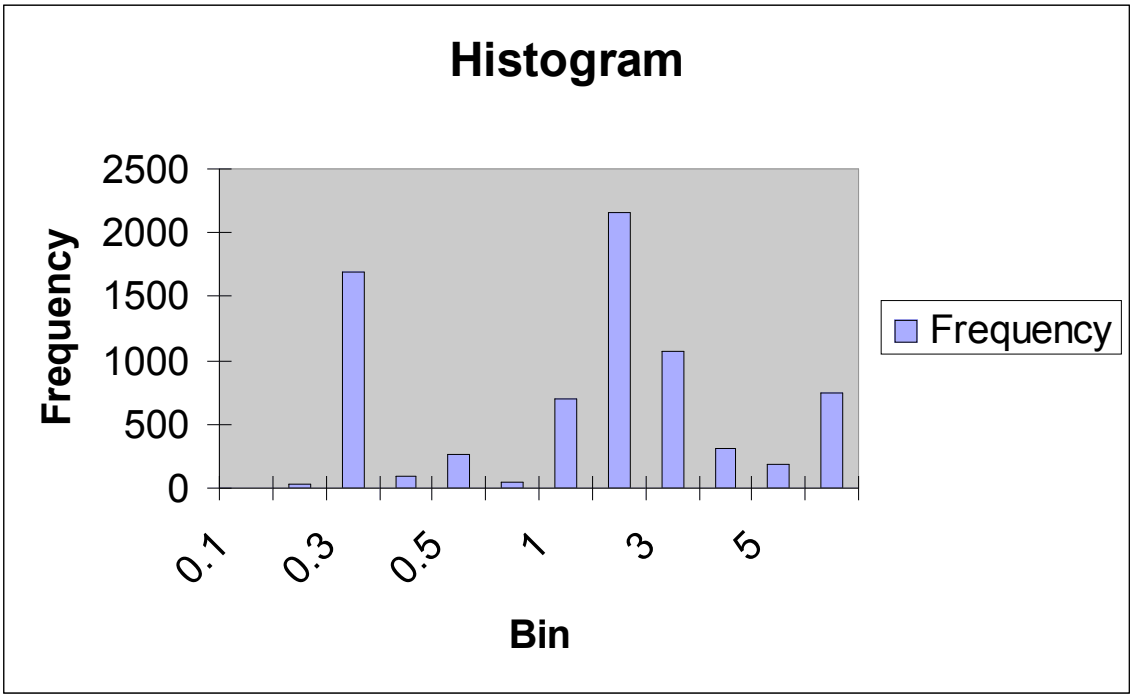
Since we are working with an annotated graph i.e. each node has been labeled with GO terms, we can take advantage of this knowledge to discover links between different pathways which were unknown before. If two nodes have highly similar GO terms, then they are likely to show up together if one of them is part of a significant simple path that we find. We claim that if we eliminate edges which connect nodes with highly similar GO terms and then find significant simple paths, we can discover the less obvious connections. We use the yeast network provided by Asthana et al [5] which has 3027 proteins and 12519 edges. We perform the following steps.

a) Find distance between nodes in the graph which have an edge between them using the GO distance measure developed by Vishwakarma Singh and Sayan Ranu. The distance measure takes into account unequal number of GO terms on the two nodes.

The distance function actually returns three distances – 1) Biological distance 2) Molecular distance 3) Cellular distance. We take a weighted average as follows

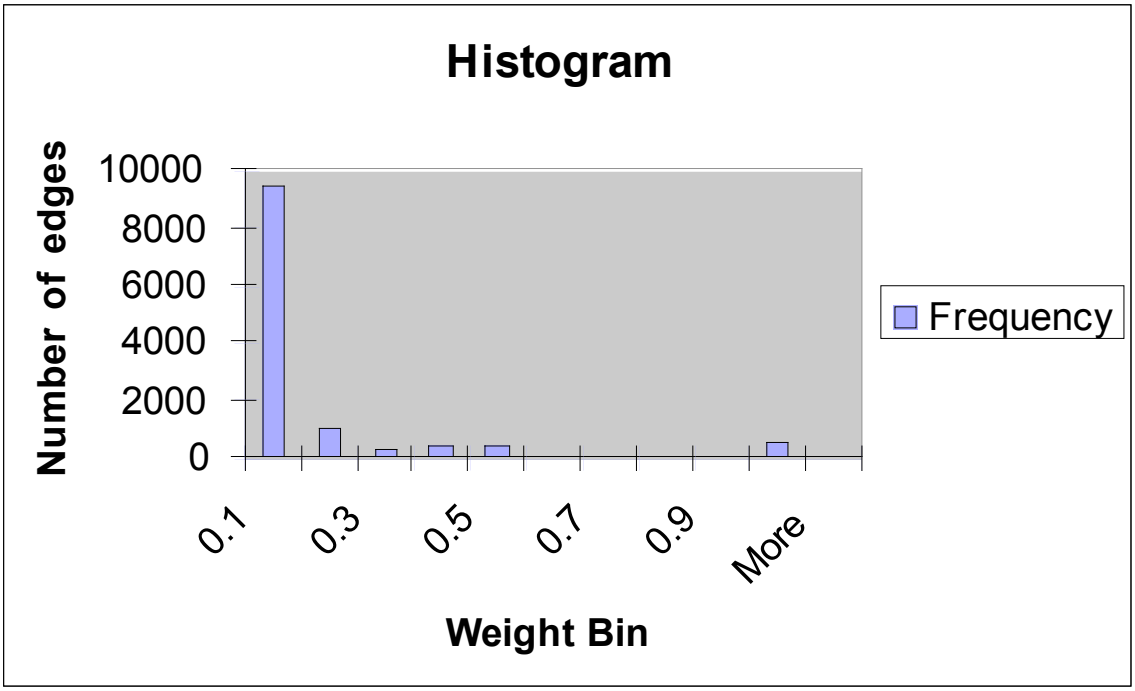
Distance = (0.2*bioscore) + (0.4*cellscore) + (0.4*molscore)

b) The distance distribution using the above distance measure for the graph is shown below.

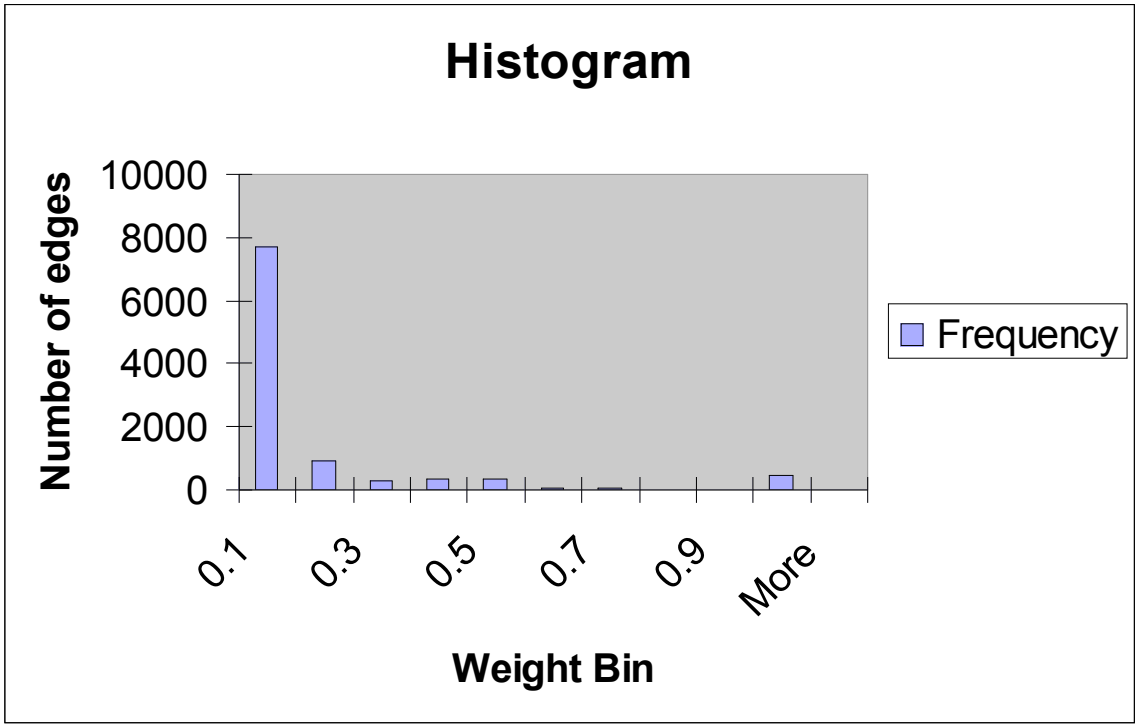


Using this distance function, we tried to cluster the nodes into K groups where K is the length of the simple path we are trying to find. The reasoning was that we'll have maximum GO term diversity in the paths that we find if we assign the same color to nodes belonging to the same cluster. Unfortunately, we found that the best clustering was found only when K = 2. For the other cases, we ended up with some thinly populated and some heavily populated clusters. Therefore, we decided on using a distance threshold in order to determine high-similarity. We analyzed the above distribution and used a threshold of 0.6 for high-similarity. We removed the edges which fall below the threshold.

c) The distribution of the weights on the edges before the qualifying edges were removed is shown below.



We eliminated edges in which the nodes had a distance < 0.6. The number of qualifying edges was 2284. **Therefore, the number of edges in the graph reduced from 12519 to 10235.**
The weight distribution after these edges were removed is shown below.



We find that out of 2284 which were removed, 1715 were from the 0.1 weight bin. This suggests that the graph still has many high-quality links for us to find significant simple paths.

4 Results

10 most significant simple paths are listed for each case. The biological significance of the top two paths is also explained.

With random coloring

Following are the 10 most significant paths of length 4.

- **ynl071w-yml064c-ynl218w-ylr423c**

Protein involved in Pyruvate metabolism in the Pyruvate dehydrogenase complex →
Protein involved in binding regulation →
Protein involved in maintenance of genome →
Protein involved in autophagy(Garbage collection inside the cell).

- **yal032c-ylr423c-ynl218w-yml064c**

Protein required for pre-mRNA splicing →
Protein involved in autophagy(Garbage collection inside the cell) →
Protein involved in maintenance of genome →
Protein involved in cytokinesis (cell division)

- ypr182w-yml064c-ynl218w-ylr423c
- ydl195w-yml028w-ynl189w-yjr133w
- ypl214c-ynl189w-ymr290c-yml064c
- yhr114w-ylr423c-ynl218w-yml064c
- yfr052w-yjr133w-ynl189w-yml028w
- ydr477w-ynl218w-ylr423c-yil150c
- ylr423c-ynl218w-yml064c-yjr056c
- ydl113c-ylr423c-ynl218w-yml064c

Following are the 10 most significant paths of length 5

- **ydl113c-ylr423c-ynl218w-yml064c-ydl126c**

Protein required for transport of aminopeptidase in the cytoplasm-to-vacuole targeting pathway →
Protein involved in autophagy(Garbage collection inside the cell) →
Protein involved in maintenance of genome →
Protein involved in cytokinesis (cell division) →
Protein involved in degradation of ubiquitinated proteins

- **ygl212w-ylr423c-ynl218w-yml064c-yel015w**

Component of the vacuole SNARE complex involved in vacuolar morphogenesis →
Protein involved in autophagy(Garbage collection inside the cell) →
Protein involved in maintenance of genome →
Protein involved in cytokinesis (cell division) →
Non-essential conserved protein of unknown function, plays a role in mRNA decapping

- ynl189w-ymr290c-yml064c-ynl218w-ylr423c
- yal032c-ylr423c-ynl218w-yml064c-yjr056c

- yal032c-ylr423c-ynl218w-yml064c-yjr056c
- ynl218w-yml064c-ymr290c-ynl189w-ypl214c
- yhr114w-ylr423c-ynl218w-yml064c-yjr056c
- yer070w-yml064c-ynl218w-ylr423c-yil150c
- ygl212w-ylr423c-ynl218w-yml064c-ypr182w
- ygl212w-ylr423c-ynl218w-yml064c-yjr056c

Following are the 10 most significant paths of length 6

- ycr057c-ynl064c-yml064c-ynl218w-ylr423c-yol069w

Conserved 90S pre-ribosomal component essential for proper endonucleolytic cleavage →
Protein chaperone involved in regulation of the heat shock proteins →
Protein involved in cytokinesis (cell division) →
Protein involved in autophagy(Garbage collection inside the cell). →
involved in chromosome segregation, spindle checkpoint activity and kinetochore clustering

- **ydl113c-ylr423c-ynl218w-yml064c-ybr020w-ygl237c**

Protein required for transport of aminopeptidase in the cytoplasm-to-vacuole targeting pathway →
Protein involved in autophagy(Garbage collection inside the cell) →
Protein involved in maintenance of genome →
Protein involved in cytokinesis (cell division) →
Involved in GALactose metabolism →
Heme Activator Protein

- ylr291c-yer025w-yml064c-ynl218w-ylr423c-ygl212w
- ylr291c-yer025w-yml064c-ynl218w-ylr423c-ygl212w
- ynl333w-yml064c-ynl218w-ylr423c-ygl212w-ygl161c
- yol069w-ylr423c-ynl218w-yml064c-ynl064c-ydr394w
- ydr357c-ypr182w-yml064c-ynl218w-ylr423c-yil150c
- ydr143c-ynl189w-ymr290c-yml064c-ynl218w-ylr423c
- ynl189w-ymr290c-yml064c-ynl218w-ylr423c-yol069w
- ylr423c-ynl218w-yml064c-ymr290c-ynl189w-yml028w

With coloring such that high-similarity proteins have the same color

Following are the 10 most significant paths of length 4

yel015w-yml064c-ynl218w-ylr423c

Non-essential conserved protein of unknown function, plays a role in mRNA decapping → Protein involved in binding regulation → Protein involved in maintenance of genome → Protein involved in autophagy(Garbage collection inside the cell).

ynl189w-ymr290c-yml064c-yjr056c

Karyopherin alpha homolog, forms a dimer with karyopherin beta Kap95p to mediate import of nuclear proteins i.e. involved in protein carrier activity →
ATP-dependent RNA helicase; localizes to both the nuclear periphery and nucleolus→Protein involved in binding regulation → **Protein of Unknown function**

ylr175w-ynl189w-ymr290c-yml064c
yjr131w-yml064c-ynl218w-ydr477w
ybr252w-yml064c-ynl218w-ylr423c
yal032c-ylr423c-ynl218w-yml064c
ypr182w-yml064c-ynl218w-ylr423c
ydr477w-ynl218w-ylr423c-yil150c
ynl333w-yml064c-ynl218w-ylr423c
yol069w-ylr423c-ynl218w-yml064c

Following are the 10 most significant paths of length 5

ylr175w-ynl189w-ymr290c-yml064c-yjr056c

Pseudouridine synthase catalytic subunit of box H/ACA small nucleolar ribonucleoprotein particles →
Karyopherin alpha homolog, forms a dimer with karyopherin beta Kap95p to mediate→
ATP-dependent RNA helicase; localizes to both the nuclear periphery and nucleolus→
Protein involved in binding regulation
→ **Protein of Unknown function**

ypl235w-yml064c-ynl218w-ylr423c-yil150c
ydr174w-yml064c-ynl218w-ylr423c-yhr143w-a
yel015w-yml064c-ynl218w-ylr423c-yol069w
ynl189w-ymr290c-yml064c-ynl218w-ylr423c
ynl333w-yml064c-ynl218w-ylr423c-yal032c
ynl333w-yml064c-ynl218w-ylr423c-yal032c
yfl037w-yml064c-ynl218w-ylr423c-yhr143w-a
yol069w-ylr423c-ynl218w-yml064c-ypr182w
yhr143w-a-ylr423c-ynl218w-yml064c-yjr131w

Following are the 10 most significant paths of length 6

yhr143w-a-ylr423c-ynl218w-yml064c-yjl034w-ymr049c

Daughter cell-specific secreted protein with similarity to glucanases →
Protein involved in autophagy(Garbage collection inside the cell) →

Protein involved in maintenance of genome →
Protein involved in binding regulation →
ATPase involved in protein import into the ER, also acts as a chaperone to mediate protein folding in the ER and may play a role in ER export of soluble proteins→
Protein required for maturation of the 25S and 5.8S ribosomal RNAs

yol069w-ylr423c-ynl218w-yml064c-yjl034w-ymr049c
involved in chromosome segregation, spindle checkpoint activity and kinetochore clustering →
Protein involved in autophagy(Garbage collection inside the cell) →
Protein involved in maintenance of genome →
Protein involved in binding regulation →
Protein involved in binding regulation →
Protein required for maturation of the 25S and 5.8S ribosomal RNAs

yhr143w-a-ylr423c-ynl218w-yml064c-yjl034w-ypr110c
yhr143w-a-ylr423c-ynl218w-yml064c-ydr190c-ykr026c
ygl237c-yfl037w-yml064c-ynl218w-ylr423c-yil150c
ylr291c-ydr190c-yml064c-ynl218w-ylr423c-yhr143w-a
ylr291c-ydr190c-yml064c-ynl218w-ylr423c-yil150c
ydr357c-ypr182w-yml064c-ynl218w-ylr423c-yil150c
ylr423c-ynl218w-yml064c-ymr290c-ynl189w-yjr068w
ynl189w-ymr290c-yml064c-ynl218w-ylr423c-yol069w

5 Conclusion

We find that, overall; most of the proteins involved in the significant simple paths that we found are involved in housekeeping activities inside the cell – autophagy, genome maintenance etc. This result may not be so interesting
There are a few pathways that we find however in which there is more than one protein not involved in housekeeping activities – for example, yol069w and ymr049c.
We find that occurrence of such proteins is more in the case where we have removed edges between nodes with similar GO terms. Also, most importantly, we find two pathways in which a protein of unknown function(yjr056c) is involved. We found this only after we removed the edges between nodes with similar GO terms. It may be interesting to investigate the function of this protein given these interactions.

References:

1) Efficient algorithms for detecting signaling pathways in protein interaction networks. Jacob Scott et al, Journal of computational Biology, 2006
2) R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskili, and U. Alon. Network motifs: simple building blocks of complex networks. Science, 298:824–828, 2002
3) V. Arnau, S. Mars, and I. Marin. Iterative cluster analysis of protein interaction data. Bioinformatics, 21(3):364–378, 2005.
4) M. Ashburner, C. Ball, J. Blake, D. Botstein, H. Butler, J. Cherry, A. Davis, K. Dolinski, S. Dwight, J. Eppig, M. Harris, D. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. Matese, J. Richardson, M. Ringwald, G. Rubin, and G. Sherlock. Gene

ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet, 25(1):25–29, 2000.
5) S. Asthana, O. D. King, F. D. Gibbons, and F. P. Roth. Predicting protein complex membership using probabilistic network reliability. Genome Research, 14:1170–1175, May 2004.
6) G. D. Bader and C. W. V. Hogue. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics, 4(2), 2003.
7) Hu, X. Yan, Y. Huang, J. Han, and X. J. Zhou. Mining coherent dense subgraphs across massive biological networks for functional discovery. Bioinformatics, 21(Suppl. 1):i213–i221, 2005.
8) J. S. Bader. Greedily building protein networks with confidence. Bioinformatics, 19(15):1869–1874, 2003.