Spectral Decomposition of Signaling Networks

B. Parvin, N. Ghosh, L. Heiser, M. Knapp, C. Talcott,
K. Laderoute, J. Gray, and P. Spellman *
Lawrence Berkeley National Laboratory Stanford Research Institute
Berkeley, CA 94720 Stanford, Ca.

Abstract

Many dynamical processes can be represented as directed attributed graphs or Petri nets where relationships between various entities are explicitly expressed. Signaling networks modeled as Petri nets are one class of such graphical modeling and representations. These networks encode how different protein in specific compartments, interact to create new protein products. Initially, the proteins and rules governing their interactions are curated from literature and then refined with experimental data. Variation in these networks occurs in topological structure, size, and weights associated on edges. Collectively, these variations are quite significant for manual and interactive analysis. Furthermore, as new information is added to these networks, the emergence of new computational models becomes paramount. From this perspective, hierarchical spectral methods are proposed and applied for inferring similarities and dissimilarities from an ensemble of graphs that corresponds to reaction networks. The technique has been implemented and tested on curated signaling networks that are derived for breast cancer cell lines.

1 Introduction

Graphical models of complex systems such as biological processes enable semantic representation of complex processes. The advent of high throughput experimental processes and structured curation of knowledge has yielded a more complete picture of these systems. One class of such processes is signaling networks that characterize protein interactions and their byproducts. Modeling of signaling networks can occur at different scales of abstraction. At one of end of spectrum, continuous modeling represents dynamic biochemical interactions. At the other end of the spectrum, modeling can be purely statistical; this is amenable for high throughput experimental design. While detailed continuous modeling enables mechanistic exploration, higher level modeling provides the basis for constructing relationships through statistical correlation. An intermediate step between these end points are the Boolean, deterministic, and Petri net representations. These representations not only provide a simplified view, but also place a potential bound and enforce stability on a more complex model, e.g., ordinary differential equations. The emerging notion is not an integrated and aggregate model, which embodies all cellular functions, but multiscale models that encapsulate different aspects of the model systems at different levels of abstraction. Signaling networks are complex and may contain up to several hundred proteins. Potential variations in a network originate from different model system (e.g., different cell lines), changes in microenvironment, and treatment regime. It is therefore desirable to compare and contrast these graphical structures for similarity, as well as dissimilarities, for large scale databases. While the proposed method and application have been evaluated on signaling networks derived from Pathway logic [16], this technique is extensible to any kind of database of graph structures for post analysis. The primary contribution of this paper is a methodology for hierarchical decomposition of an ensemble of graphs. The actual decomposition is through recursive application of spectral analysis [1, 2, 15] on a directed graph and derived subgraphs. The rationale for hierarchical decomposition originates from the fact that graphical structures are weighted and aggregation of an ensemble of graphs leads to preferred topological motifs with similar weighting. This similarity in weighting can be multiscale and hierarchical; thus, multiscale recovery of these motifs contribute to improved visualization, partitioning, and understanding of an ensemble of networks. Two examples of signaling network for two mammalian cell lines are shown in Figure 1.

Section 2 provides a short summary of the back-

*Research funded by the National Cancer Institute, under contract number DE-AC02-05CH11231 with the University of California. LBNL publication number is LBNL-58866. Point of contact: parvin@media.lbl.gov

ground biology. Section 3 reviews spectral analysis within the context of graph representation. Section 4 outlines the details of approach, examples, and concluding remarks. Section 5 concludes the paper.

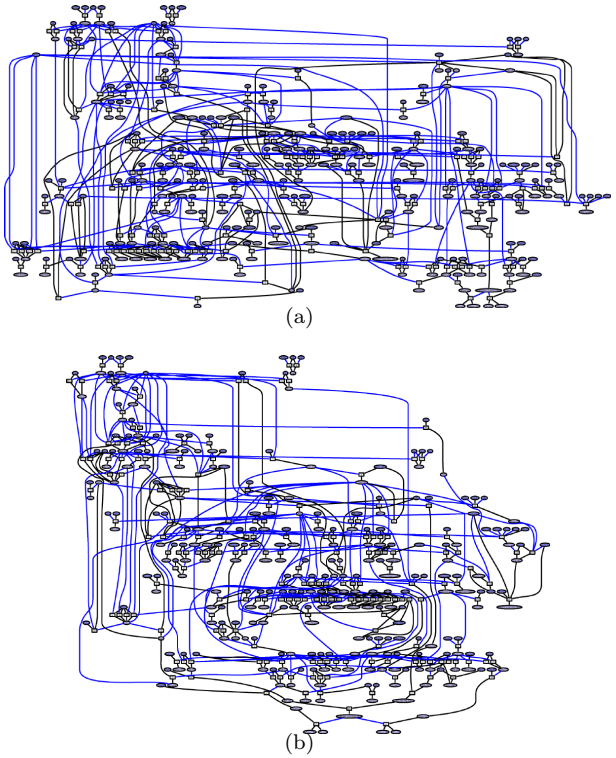


Figure 1: A graphical representation of the signaling networks generated with Pathway Logic. In each graph, the colored circles represent proteins. The white boxes represent rules, or signaling between the proteins. Two cell types are shown: (a) MCF7, a basal cell line and (b) SKBR3, a luminal cell line.

2 Biological driver

In cancer cells, the pathways that control the cell cycle, cell growth, apoptosis (cell death), and cell adhesion become deregulated through mutations [7]. Our goal is to understand the cellular signaling pathways associated with breast cancer. To that end, we have modeled the signaling pathways in a panel of 51 breast cancer cell lines. These cell lines capture both rare mutations, as well as those that frequently occur in breast tumors. One important feature of both breast tumors and cell lines is the site of origin. Tumors and cell lines that originate from basal epithelium tend to be much more invasive than those that originate from luminal epithelium. Furthermore, these two groups show distinct genetic patterns [12]. We were interested in the cellular signaling pathways that distinguish the basal and luminal cell lines.

3 Graph partitioning

Graph partitioning is concerned with the grouping of the vertices of a connected graph into subsets so as to minimize the total cut weight, as shown in Figure 2. One intent of graph decomposition is to simplify the graph matching problems into simpler subgraph matching problems. For example, it has been shown that error-tolerant graph matching [10] can be simplified using decomposition methods and reduced to indexing. Within the context of this paper, spectral analysis enables stable decomposition of graphs for further analysis by global structural properties of eigenvectors corresponding to the Laplacian matrix. Let $q_i = \{1, -1\}$ be a membership function for the assignment of each node i for a two-way decomposition. The optimum cut is given by $J_{min} = \frac{1}{4}\Sigma_{i,j} w_{ij}[q_i - q_j]^2$, which can be rewritten as $\frac{1}{2}q^T(D - W)q$. The solution to J_{min} is given by second-smallest eigenvector of $(D - W)$. Formally, for a weighted graph $G = (V, E, \Omega, W)$, where V is a set of nodes, E is the set of arcs, $\Omega = V_i, i \in V$, and $W = w_{ij}, ij \in E$. The Laplacian matrix is given by

$$L_{ij}(G) = \left\{ \begin{array}{ll} \sum_{ik \in E} w_{ik} & \text{if } i = j \\ -w_{ij} & \text{if } i \neq j \text{ and } (i,j) \in E \\ 0 & \text{otherwise} \end{array} \right\} \quad (1)$$

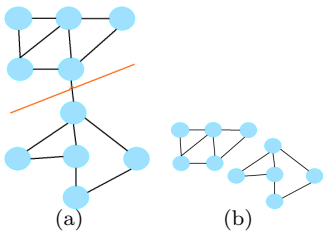


Figure 2: Decomposition of an undirected graph into two subgraphs while maintaining balanced graph size.

The second smallest eigenvector of the Laplacian is also known as the Fiedler vector with many well-known applications and properties. The major applications of the Laplace eigenvalues are the max-cut problem, semidefinite programming, and steady-state random

walks on Markov chains [11]. Furthermore, the interlacing properties of the eigenvalues have been shown to be related with the chromatic number (minimum number of colors such that no two adjacent vertices share the same color), and the diameter and bandwidth of graphs [6]. In addition, the Fiedler vector has been used for recursive partitioning of the image [15] for segmentation, and placing nodes of the graph in a serial order for visualization [4]. For the purpose of image partitioning, grouping is defined as a case of graph partitioning, where partitioning is normalized to inhibit formation of small sets of isolated nodes. An elegant objective function is derived in the form of a Rayleigh quotient, which has a standardized solution as the eigenvector corresponding to the second smallest eigenvalue.

One approach for graph partitioning is to construct non-overlapping super-cliques using the node order by the Fiedler vector [13,14]. A superclique is defined as a center node and all of its immediate neighboring nodes. A preliminary metric (distance measure) can be defined to test the method and for locating non-overlapping super-cliques, e.g., a center node that is not on the perimeter and whose metric exceeds its neighbors. The resulting super-cliques are relatively stable subgraphs, which can be matched to the original graph using discrete relaxation and edit distance algorithms [18].

3.1 Graph clustering

Despite recent progress in measuring similarities between graphs and performing inexact graph-matching, clustering on graphs remains an open and challenging problem. Besides lack of ordering, graphs are often noisy (e.g., contain different numbers of nodes and edges), and, as a result, standard pattern recognition techniques are inadequate (e.g., variable vector size). One approach is to measure pairwise similarities of the graphs and cluster them by searching for sets of graphs with strong mutual affinity [8]. Another approach to overcoming this problem is through spectral representation of the graphs and representing the structures of the graphs as vectors of "fixed" length. Each component of a vector represents a different spectral mode of the graph adjacency matrix. The spectral graph theory suggests a number of unary and binary features [9]. Examples of unary features are (1) leading eigenvalues, (2) eigenmode volume, (3) eigenmode perimeter, and (4) derived features such as Cheeger constants. Binary features correspond to pairwise attributes of the eigenmodes. Examples are the "mode association matrix", which projects the adjacency matrix onto the basis spanned by the eigenvectors, and the "intermode distances," which is the path associated with the minimum number of edges, between the most significant nodes associated with each eigenmode of the adjacency matrix. These modes are then embedded in a pattern space such as principal component analysis (PCA) or independent component analysis (ICA).

4 Approach

Review of spectral methods, as applied to graphs, indicates that previous efforts have focused on either decomposition of a single graph or feature extraction from an ensemble of graphs for subsequent clustering. The focus of this paper is on a distinct biological application and recursive decomposition of an ensemble of graphs through spectral analysis. The main advantages of spectral method are (1) a more stable decomposition, (2) reduction in the number of free parameters, and (3) recursive application of this technique for coarse-to-fine decomposition. The first step of the process is to construct a composite representation from within and between labeled graphs for similarities and dissimilarities, respectively. This is followed by iterative decomposition of the Laplacian of the composite graph for revealing coarse-to-fine motifs from the signaling network. Although the results are limited to the Petri net derived from Pathway Logic, as shown in Figure 1, the technique can be extended to other forms of graphical structures, such as workflow and dynamical processes.

4.1 Network representation

The network models were curated from literature and then refined with experimental data. Signaling motifs have two node types, corresponding to (1) protein abundance and (2) rules. Internally, a typical representation uses Systems Biology Markup Language as follows:

```
<sbml level="2" version="1"
xmlns="http://www.sbml.org/sbml/level2">
<model id="myGraph" name="myGraph" >
  <listOfCompartments>
    <compartment id="CLc" name="Cell cytosol" />
   </listOfCompartments>
  <listOfSpecies>
      <species id="o109" name="Pkcz-act-CLi"
      compartment="CLi" initialConcentration="0"/>
      <species id="o124" name="(Raf1:Rkip)-CLc"
      compartment="CLc" initialConcentration="1"/>
      <species id="o125" name="Raf1-CLc"
      compartment="CLc" initialConcentration="1"/>
      <species id="o126" name="Rkip-phos-CLc"
      compartment="CLc" initialConcentration="0"/>
  </listOfSpecies>
```

Affinity type	$E_i = E_j$	$E_i \times E_j = 0$	$E_i \times E_j \neq 0$
Similarity	Max value	0	$\sum E_i$
Dissimilarity	0	Max value	$\sqrt{\lvert E_i^2 - E_j^2 \rvert}$

Table 1: Construction of the affinity matrix for similarity or dissimilarity analysis: E_i and E_j indicate the weighted edge from initial composition of the graph ensemble.

```
  <reaction id="t127" name="230.Rkip.by.aPkc">
      <listOfReactants>
          <speciesReference species="o124"/>
      </listOfReactants>
      <listOfProducts>
          <speciesReference species="o125"/>
          <speciesReference species="o126"/>
      </listOfProducts>
      <listOfModifiers>
          <modifierSpeciesReference species="o109"/>
      </listOfModifiers>
  </reaction>
</model>
</sbml>
```

Where Reactants and Products are input and output, respectively. Modifiers refer to protein or other components that must be present for the reaction to take place, and it remains unchanged during the reaction. Each signaling node, in every graph, has a distinct ID; thus, it is invariant to motif discovery. For similarity analysis, a composite graph is constructed by aggregating self-similar edges (e.g., edges with identical nodes between multiple graphs). This composite graph is computed by $G_c = \sum_i G_i$ and then normalized for the number of graphs in the database. For dissimilarity analysis, the system is designed to compute differences between two groups of graphical networks. First the composite representation, within each group, is computed through aggregation as before. Next the difference between two composite representations is computed. This difference corresponds to the differences among self-similar edges, e.g., $G_c = |\sum_i G_i^{Group\ A} - \sum_j G_j^{Group\ B}|$. The corresponding affinity matrix has a symmetrical distance property (e.g., identical distance measure when computed from G_i to G_j and vice versa). These aggregation operators for edges in the composite graphs are shown in Table 1. These operators generate weight matrices that are positive and symmetric. Furthermore, a computed composite graph is often disjoint; thus, its connected components are identified. A numerical example of constructing a composite graph that represent dissimilarities between two graphs are shown in Figure 3.

Intuitively, such a representation constructs a weighted graph for capturing corresponding affinities from a set of graphs. The affinity matrix is symmetric with each element encoding a weighted edge between corresponding nodes. Presently, these composite affinity matrices are of the order of 900 nodes.

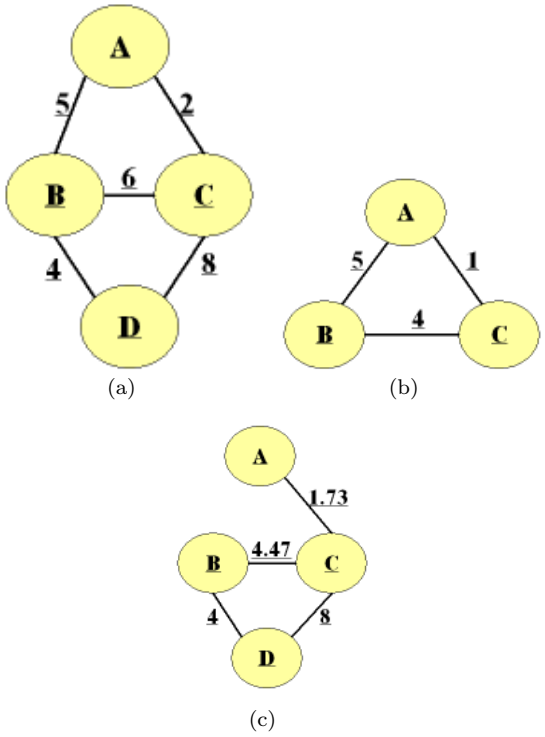


Figure 3: A numerical example for constructing a composite graph corresponding to dissimilarities: (a) graph A; (b) Graph B; and (c) computed composite graph.

4.2 Decomposition algorithm

The above representation leads to the realization that decomposition can be hierarchical, due to the weighted representation of the graph, e.g., higher weights on a set of edges implies stronger grouping. The spectral decomposition is as follows:

1. Compute connected components of the composite graph, G_c,
2. For each connected component compute in G_c do
   (a) Compute Laplacian of the connected component and its corresponding Fiedler vector,
   (b) Partition the composite graph into two subgraphs, based on the sign of each element of Fiedler eigenvector,

(c) Repeat above 2 steps until each element of the Fiedler vector has the same sign.

4.3 Model generation

The Pathway Logic framework was used to model the signaling networks for the panel of 51 breast cancer cell lines, 26 basal and 25 luminal [5, 16]. The construction of a Pathway Logic model requires two key elements: a set of rules and an initial state. The rules represent all the biochemical transitions that can occur. Presently, the model has a set of 880 rules that detail cellular signaling along multiple pathways, including the raf-mek-erk pathway, PI3K cascade, and ErbB-related signaling [3,17,19]. The initial state specifies all of the molecular components present in the cell, as well as the cellular compartment in which they are located. A Petri net for each cell line has been generated, and data about protein function were acquired from the *SOURCE* database (http://genome-www5.stanford.edu/cgi-bin/source/sourceResult).

4.4 Initialization of Petri net

For the similarity analysis, protein abundance data was used to determine the presence or absence of 28 proteins known to be critically involved in cancer cell signaling. For dissimilarity analysis, mRNA expression data was analyzed to determine the initial state for each cell line. Briefly, for each gene, the distribution of mRNA expression values across the panel of cell lines was analyzed. These expression values were grouped using PAM clustering technique. Genes best represented by 2 clusters were considered present in some cell lines and absent in others. Genes that yielded a single cluster were considered present in all cell lines. With this method, 39 out of 640 components in the model were present in the initial state of some cell lines and absent in others; the remainder were present in all cell lines.

4.5 Experimental results

A number of experiments on signaling networks for similarity and dissimilarity analysis were performed, and in all cases spectral analysis has isolated significant motifs with meaningful biological implications and validated by domain expert. One experiment included 25 Petri nets, which were generated by Pathway Logic, and a database of rules were curated to describe the interactions of proteins. These Petri nets represent individual breast cancer cell lines, and for each cell line, Western blot data were used to determine the presence

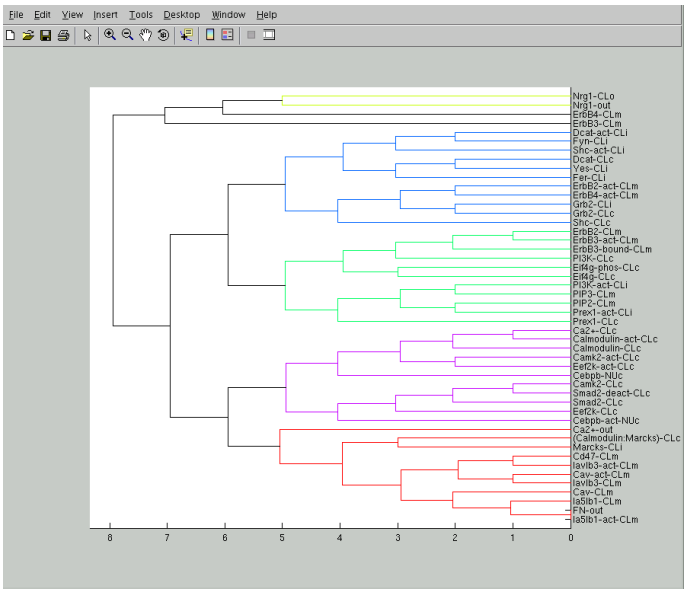


Figure 4: Hierarchical decomposition of one connected component of the composite graph using recursive spectral bisection. Each node in the dendrogram corresponds to a directed graph.

or absence of 28 proteins known to be critically involved in cancer cell signaling. These breast cancer cell lines can be classified by their site of origin as being basal or luminal, where the label serves one of the bases of analysis.

For similarity analysis, a composite graph was constructed as described earlier, and spectral clustering was performed for each connected component. Figure 4 shows the dendrogram that corresponds to one of these connected components. At each node of the dendrogram, two "child" motifs are generated from the "parent" motif. Figure 5 shows an example of the motif structure at a specific node along the dendrogram. The corresponding left and right child decompositions are shown in Figure 6. The two child graphs contain proteins involved in two distinct cellular processes. The left child motif, shown in Figure 6a, contains many proteins involved in cell surface signaling. Integrins (Ia5Ib1 and IavIb3) are the most prominent features in this graph. Integrins are heterodimeric integral membrane proteins composed of an alpha and beta chain. These proteins are involved in cell adhesion and cell-surface mediated signaling. Both of these processes are extremely important for the maintenance of cell integrity. These processes are frequently disrupted in cancer. The proteins in the right child motif, shown in Figure 5b, are involved in the two-step process used to generate new proteins: transcription

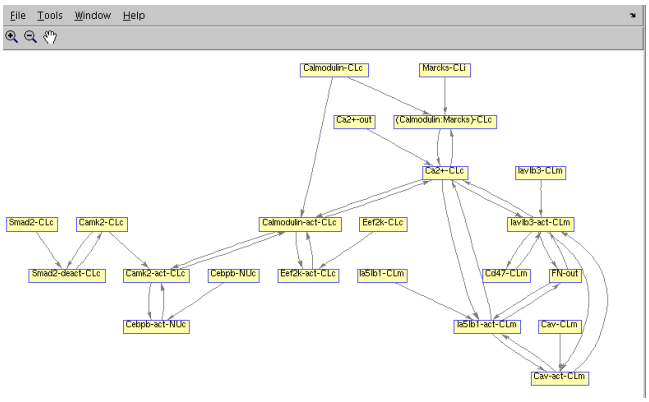
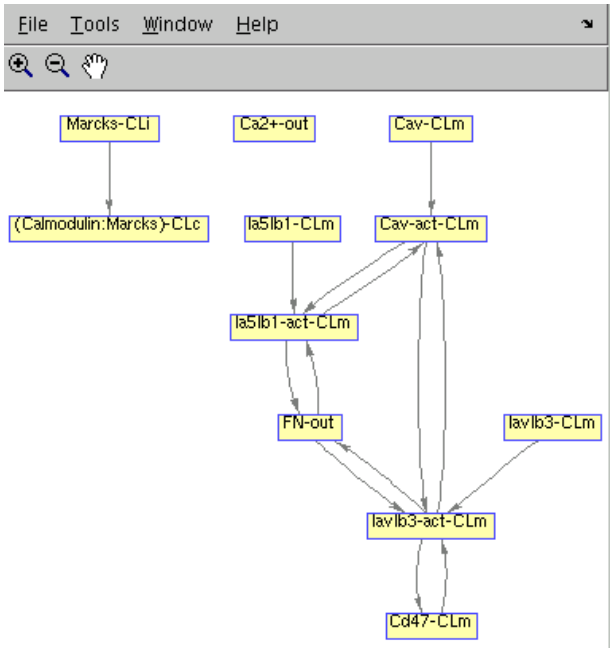


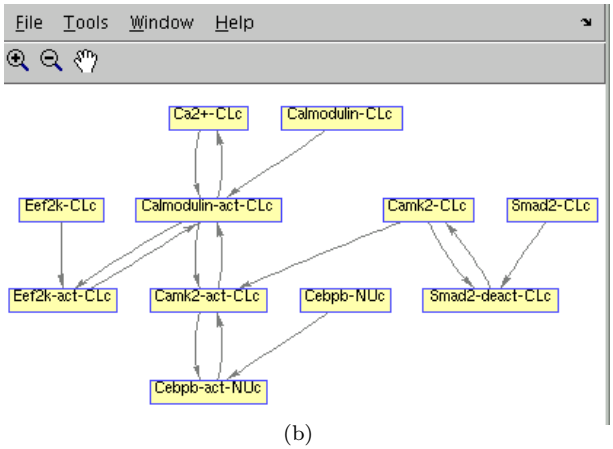
Figure 5: A graphical motif at two levels below the root node of Figure 4.

and translation. Smad2 is a transcriptional regulator, while Eef2k is a regulator of translation. Smad proteins are signal transducers and transcriptional modulators that mediate multiple signaling pathways. It mediates the signal of the transforming growth factor (TGFβ), and therefore regulates multiple cellular processes, including cell proliferation, apoptosis, and differentiation. Eef2k links activation of cell surface receptors to cell division, and is involved in the regulation of translation (protein synthesis). This example demonstrates the utility of spectral decomposition for isolating signaling motifs that represent unique cellular functions.

For dissimilarity analysis, the technique revealed many signaling motifs in our model networks, where some are more frequent than others. This frequency is estimated by the weight of affinity matrix from the composite graph. Here we describe the biological basis for a set of child graphs that are validated by domain expertise. Figure 7 shows two child graphs at the fifth level of a dendrogram computed for dissimilarity analysis. Both of these motifs occur more frequently in luminal than basal cell lines. The left child of Figure 7a contains signaling related to cell structure and motility. Specifically, there is a small network centered on the phosphorylated form of beta-catenin (Bcat-Yphos-CLc). Beta-catenin is an adherens junction protein involved in the regulation of cell adhesion. Interestingly, activating mutations in beta-catenin have oncogenic activity that may result in tumor development. The other subnetwork in this graph involves rac1, elmo1, and dock1. These proteins are involved in regulating changes in cell shape required for cell motility and engulfment of apoptotic cells. All together, the proteins in this signaling motif are important for cellular integrity. The right child graph 7b shows a small

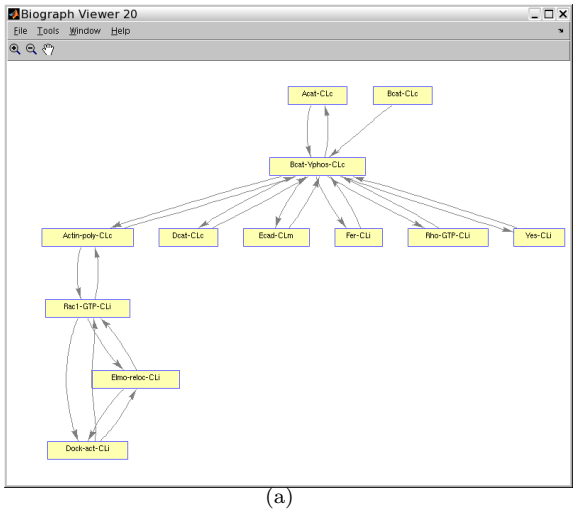


(a)

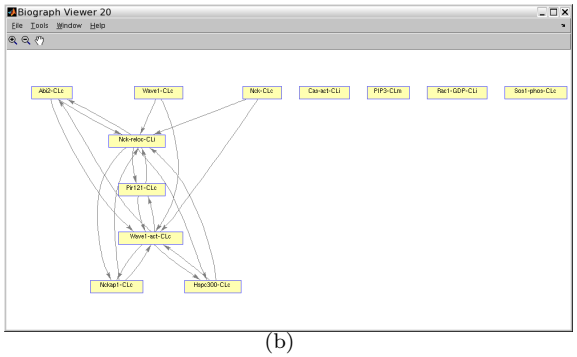


(b)

Figure 6: Decomposition of Figure 5 into left and right child motifs.



(a)



(b)

Figure 7: Two child motifs of a dissimilarity analysis corresponding to luminal and epithelial cells.

network involving wasf1 (Wave1-act-CLc) and nck1. These proteins play a role in signal transduction from small GTPases and receptor tyrosine kinases to downstream targets that include ras. Ras is an oncogene, and mutations in signaling associated with ras have been implicated in many types of cancers. The examination of these two signaling motifs indicate validation of this method for identifying and understanding key regions of a large signaling network.

5 Conclusion

A system has been developed and implemented for iterative decomposition of an ensemble of graphs for similarity or dissimilarity analysis using spectral graph theory. Operators are defined to compute corresponding affinity matrices in both cases. The spectral methods enable a more stable model for decomposition with a reduced number of free parameters. The proposed technique has been applied to signaling networks to reveal coarse-to-fine motifs of significance that either advocate preferred similarities or dissimilarities. These motifs were compared and validated for their biological affinity. Future extensions of this method will focus on improved design of affinity matrices that allow spatio-temporal analysis with the same objectives.

References

[1] J.E. Atkins, E.G. Boman, and B. Hendrickson. A spectral algorithm for seriation and the consecutive ones problem. *SIAM Journal on Computing*, 1998.

[2] F.R.K. Chung. Spectral graph theory. *CBMS Series 92, American mathematical society, edited*, 100, 1997.

[3] M. Cully, H. You, A.J. Levine, and T.W. Mak. Beyond pten mutations: the pi3k pathway as an integrator of multiple inputs during tumorigenesis. *Nat Rev Cancer*, 6:184–192, 2006.

[4] J. Diaz, J. Petit, and M. Serna. A survey on graph layout problems. *ACM Computing Surveys*, 34:313–356, 2002.

[5] S. Eker, M. Knapp, K. Laderoute, P. Lincoln, J. Mesgeuer, and K. Sonmez. Pathway logic: symbolic analysis of biological signaling. In *Pacific Symposium on Biocomputing*, pages 400–12, 2002.

[6] W.H. Haemers. Interlacing eigenvalues and graphs. *Linear Algebra and its applications*, pages (226–228):593–616, 1995.

[7] D. Hanahan and R. A. Weinber. The hallmarks of cancer. *Cell*, 100:57–70, 2000.

[8] T. Hoffmann and J.M. Buhmann. Pairwise data clustering by deterministic annealing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:192–201, 1997.

[9] B. Luo, R.C. Wilson, and E. Hancock. Spectral clustering of graphs. *Graph-based Representation in Pattern Recognition, 4th IAPR International Workshop*, pages 190–201, 2003.

[10] B.T. Messmer and H. Bunke. A new algorithm for error-tolerant subgraph isomorphism detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:493–504, 1998.

[11] B. Mohar. Some applications of laplace eigenvalue of graphs. *Graph Symmetry: algebraic methods and applications*, 497 Nato ASI Series:C:227–275, 1997.

[12] C.M. Perou, T. Solie, M.B. Eisen, M. van de Rijin, S.S. Jeffrey, C.A. Rees, D.T. Pollack, J.R.and Ross, H. Johnsen, L.A. Akslen, O. Fluge, A. Pergamenschikov, C. Williams, S.X. Zhu, P.E. Lonning, A.L. Borresen-Dale, P.O. Brown, and D. Botstein. Molecular portraits of human breast tumours. *Nature*, 406:747–52, 2000.

[13] H. Qiu and E. Hancock. Spectral simplification of graphs. In *8th European conference on computer vision*, 2004.

[14] H. Qui and E. Hancock. Graph partitioning for matching. *Graph-based Representation in Pattern Recognition, 4th IAPR International Workshop*, pages 178–189, 2003.

[15] J. Shi and J. Malik. Normalized cut and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:888–906, 2000.

[16] C. Talcott, S. Eker, M. Knapp, P. Lincoln, and K. Lateroute. Pathway logic modeling of protein functional domains in signal transduction. In *Pacific Symposium on Biocomputing*, pages 568–580, 2004.

[17] I. Vivanco and C.L. Sawyers. The phosphatidylinositol 3-kinase akt pathway in human cancer. *Nat Rev Cancer*, 2:489–501, 2002.

[18] R. Wilson and E.R. Hancock. Structural matching by discrete relaxation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:634–648, 1997.

[19] Y. Yarden and M.X. Sliwkowski. Untangling the erbb signalling network. *Nat Rev Mol Cell Biol*, 2:127–137, 2001.