CellPress

Signaling hypergraphs

Anna Ritz[1*], Allison N. Tegge[1*], Hyunju Kim[1,†], Christopher L. Poirel[1,‡], and T.M. Murali[1,2]

[1] Department of Computer Science, Virginia Tech, Blacksburg, VA, USA
[2] ICTAS Center for Systems Biology of Engineered Tissues, Virginia Tech, Blacksburg, VA, USA

Signaling pathways function as the information-passing mechanisms of cells. A number of databases with extensive manual curation represent the current knowledge base for signaling pathways. These databases motivate the development of computational approaches for prediction and analysis. Such methods require an accurate and computable representation of signaling pathways. Pathways are often described as sets of proteins or as pairwise interactions between proteins. However, many signaling mechanisms cannot be described using these representations. In this opinion, we highlight a representation of signaling pathways that is underutilized: the hypergraph. We demonstrate the usefulness of hypergraphs in this context and discuss challenges and opportunities for the scientific community.

Signaling pathways and their representations

Signaling pathways mediate the responses of a cell to its environment, starting with recognition of an external stimulus at receptors, proceeding through intracellular protein interactions and activation of transcription factors, and culminating in perturbation of the expression of target genes. Owing to their importance in cellular communication, signaling pathways are often perturbed in diseases. Numerous publicly available and often manually curated databases store information about signaling pathways [1–6]. Despite growing knowledge of signaling pathways gained from experimental data, these databases face a number of obstacles for storing and conveying this information. Databases representing signaling pathways from manual curation of the literature produce high-quality interactions, but are time-consuming to construct, are often incomplete or outdated, and might be biased according to the curators' expertise [7–10]. Databases that use automated methods for literature searches, such as predictive text mining, are relatively easy to maintain but tend to have many erroneous entries [11]. Different databases may represent the same biological event in different ways, making them difficult to standardize for computational use.

In this opinion, we describe the common representations that have been used in computational analyses of signaling pathways. After examining the limitations of these representations, we encourage the use of hypergraphs as models that better capture the complex relationships in underlying biological mechanisms. We describe three applications to motivate more powerful representations of signaling pathways. Pathway enrichment assesses whether discovered proteins are significantly enriched for proteins/interactions in a pathway of interest. Pathway reconstruction explicitly reconstructs and discovers missing proteins and interactions in a pathway of interest. Finally, pathway crosstalk captures how stimulation of one pathway may result in alternative downstream responses.

Current representations of signaling pathways

Signaling pathways as sets of proteins

The simplest representation of a pathway is a list of its members, that is, the set of proteins involved in the pathway (Figure 1). Catalogs such as the Gene Ontology [1] and the Molecular Signatures Database [12] provide signaling pathways in this format. For this representation, pathway enrichment identifies pathways whose members occur surprisingly often in a set of experimentally identified proteins (e.g., from analysis of differential gene expression) [13]. However, such set-based approaches ignore the relationships between proteins within a pathway, and thus provide no clues as to how interactions may alter gene expression [8]. These methods can correct and adjust for proteins shared among multiple pathways [14,15], and thus account for crosstalk to some extent. By definition, purely set-based methods can reconstruct only the proteins in a pathway and not the interactions among them [16].

Signaling pathways as directed graphs

Signaling pathways are also conceptualized as graphs in which nodes represent proteins and edges represent

Glossary

Node: an element (protein or compound).
Undirected edge: an unordered pair of nodes (physical interaction between two proteins).
Directed edge: an ordered pair of nodes (kinase phosphorylates a substrate). In a directed graph, an undirected edge between nodes u and v is replaced by two directed edges (u,v) and (v,u).
Hypernode: a set of node(s) (protein; protein complex).
Directed hyperedge: an ordered pair of sets of hypernodes (complex assembly).
Regulated hyperedge: a directed hyperedge regulated by a hypernode (kinase phosphorylates a protein complex, thereby activating it).

Corresponding author: Murali, T.M. (murali@cs.vt.edu).
Keywords: signaling pathways; graphs; hypergraphs.
[*] These authors contributed equally to this work.
[†] Current address: BEYOND: Center for Fundamental Concepts in Science, Arizona State University, Tempe, AZ, USA.
[‡] Current address: RedOwl Analytics, Baltimore, MD, USA.

0167-7799/
 http://dx.doi.org/10.1016/j.tibtech.2014.04.007

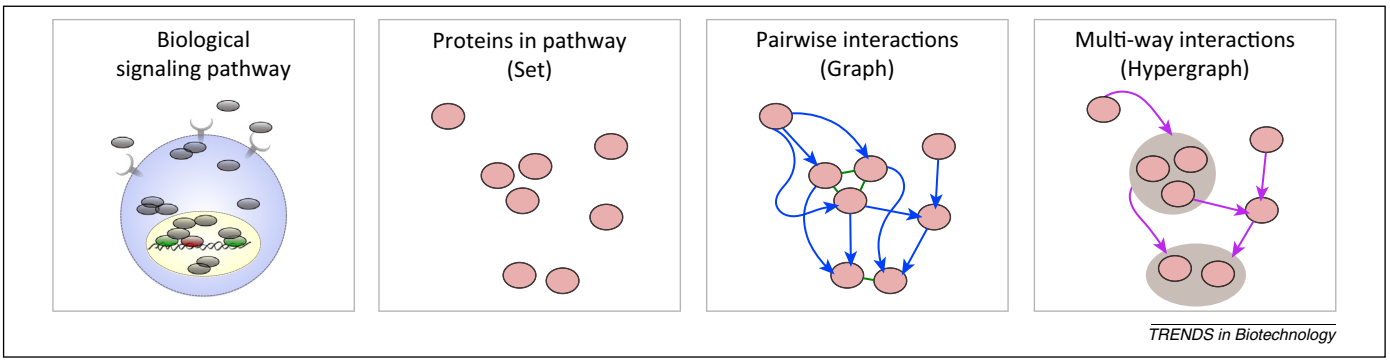

Figure 1. Signaling pathway representations. There are three main ways of representing signaling pathways. A signaling pathway may be simply represented as a set of proteins, with no additional information. Graphs encode pairwise interactions between proteins; these interactions may be undirected (green) or directed (blue). Hypergraphs, the focus of this article, encode multi-way interactions and reactions. Box 1 provides examples of graph and hypergraph representations of reactions in signaling pathways.

pairwise interactions between proteins (Figure 1 and see Glossary). The edges are often directed in signaling pathways, such as when a kinase phosphorylates a substrate. Recent enrichment methods make use of pathway topology in their scoring metrics by taking the interactions among member proteins into account [8]. There is ongoing development of this class of approaches. Pathway reconstruction algorithms for graphs typically use a large background interactome (such as a protein–protein interaction network) and identify pathways as subgraphs of proteins and interactions within the interactome. These approaches often try to find connections between signaling initiators (membrane receptors) and downstream regulators (transcription factors). Pathway reconstruction algorithms use many well-known concepts from graph theory [17–21]. Graph-based approaches to assess pathway crosstalk rely on the notion that two crosstalking pathways (each represented as a set of genes) will have statistically more interactions connecting their members than expected in a random network [22,23]. However, these approaches fail to compute the specific paths of signaling interactions that contribute to such crosstalk.

Graph representations of signaling pathways are an improvement from the 'set of proteins' representation because they capture pairwise relationships between proteins. However, signaling pathways contain more complicated relationships that are problematic for graph representations. For example, graphs often represent a complex by connecting all its members, which can artificially increase the number of edges (Figure IA in Box 1) [24]. More importantly, graphs do not accurately represent several types of molecular reactions, including regulation (e.g., activation and inhibition) or protein complex assembly and disassembly (Figure IB in Box 1). Finally, graphs do not typically distinguish between inactive and active forms of a protein or complex (Figure IC in Box 1).

Other representations of signaling pathways

Although directed graphs have been useful for representing signaling pathways, their limitations are widely recognized. A number of approaches have modified and extended graph representations. Compound graphs [25] and metagraphs [26] represent a complex as a single entity and allow a nested structure among complexes. Factor graphs [27] and Petri nets [28] introduce different types of nodes into a directed graph to represent events involving sets of proteins. Multimodal networks associate four entities with each edge: a head, a tail, a regulator, and a mode [29]. The head, tail, and regulator can each be a set of proteins, and the mode specifies how the regulator controls the transition from head to tail, for example by activation or repression.

These models of signaling pathways seek to address the shortcomings of directed graphs. However, each approach has drawbacks, including an inability to comprehensively model the complexity of signaling pathways, applicability to a limited range of computational problems, and underutilization in systems biology. Nodes in compound graphs and metagraphs focus on protein complexes. Multimodal networks do not support the hierarchical structure of signaling networks. Factor graphs and Petri nets are not ideal for generalizations of common graph-theoretic operations such as paths, connectivity, and random walks. In the next section, we seek to unify these models under the umbrella of signaling hypergraphs.

Signaling pathways as hypergraphs

Hypergraphs are a generalization of graphs that are capable of representing relationships among two or more proteins (Figure 1). Typically, directed hypergraphs consist of a set of nodes and a set of directed hyperedges in which each hyperedge connects two sets of nodes. Directed hypergraphs are an attractive alternative to directed graphs for representing complex facets of cellular processes, especially for metabolic networks [29–33]. They are also advantageous for signaling networks [30,34,35]; however, they remain an underutilized tool.

In our definition, a signaling hypergraph consists of hypernodes, directed hyperedges, and regulated hyperedges. Each hypernode represents an individual protein or a set of proteins, each directed hyperedge connects one set of hypernodes to another, and each regulated hyperedge is a directed hyperedge with one or more hypernodes that act as regulators. Box 1 illustrates this definition of signaling hypergraphs using three biological events in the canonical Wnt signaling pathway. These biological events (protein complexes, assembly of protein complexes, and regulation of proteins and complexes) commonly occur in signaling pathways. Each event may be represented as a graph consisting of multiple edges (Figure I left) or as a