Pathway Analysis with Signaling Hypergraphs

Anna Ritz
Department of Computer Science
Virginia Tech, Blacksburg, VA
annaritz@vt.edu

T. M. Murali
Department of Computer Science
ICTAS Center for Systems Biology of Engineered Tissues
Virginia Tech, Blacksburg, VA
murali@cs.vt.edu

ABSTRACT

Signaling pathways play an important role in the cell's response to its environment. Signaling pathways are often represented as directed graphs, which are not adequate for modeling reactions such as complex assembly and dissociation, combinatorial regulation, and protein activation/inactivation. More accurate representations such as directed hypergraphs remain underutilized. In this paper, we present an extension of a directed hypergraph that we call a signaling hypergraph. We formulate a problem that asks what proteins and interactions must be involved in order to stimulate a specific response downstream of a signaling pathway. We relate this problem to computing the shortest acyclic B-hyperpath in a signaling hypergraph — an NP-hard problem — and present a mixed integer linear program to solve it. We demonstrate that the shortest hyperpaths computed in signaling hypergraphs are far more informative than shortest paths found in corresponding graph representations. Our results illustrate the potential of signaling hypergraphs as an improved representation of signaling pathways and motivate the development of novel hypergraph algorithms.

Categories and Subject Descriptors

J.3 [**Life and Medical Sciences**]: Biology and Genetics; E.1 [**Data Structures**]: Graphs and Networks; G.1.6 [**Optimization**]: Linear Programming

General Terms

ALGORITHMS,THEORY

Keywords

Hypergraphs, Mixed integer linear programming, Signaling pathways, Wnt signaling

1. INTRODUCTION

Cells respond to signals from their environment through signaling pathways composed of molecular reactions that start at membrane-bound receptors and terminate at transcription factors (TFs) that regulate downstream gene expression. Many types of reactions occur in signaling pathways, e.g., complex assembly and disassembly, activation or deactivation of one protein or complex by another protein or complex, and regulation of reactions by proteins/complexes, etc. Computational methods for reasoning about signaling pathways must be faithful to the complexity of reactions within them. Directed and undirected graphs are the most pervasive representations of the structure of signaling pathways. However, graphs can only model interactions between pairs of molecules; thus they cannot accurately represent the manifold aspects of signaling pathways that involve coordinated activity of assemblages of more than two molecules [13, 18]. Directed hypergraphs and their relatives (reviewed in Section 2) are emerging as attractive alternatives to graphs. Unfortunately, directed hypergraphs continue to remain an underutilized representation for signaling pathways, despite the fact that hypergraph theory has been a well-established area of mathematics since the 1960s [3].

Recently we highlighted the potential and power of hypergraphs to address questions such as pathway reconstruction, enrichment, and crosstalk [23]. Until now, methods to solve these problems have represented pathways simply as set of proteins or as directed or undirected graphs. In this paper, we formally define the *signaling hypergraph* as a powerful representation of signaling pathway structure. Signaling hypergraphs capture several aspects of signaling reactions and the connections among them. We use signaling hypergraphs to address two general classes of questions that may be posed on pathways:

1. Is there a set of reactions that begins at protein A and terminate at protein B?
2. If we are given a set of reactions annotated to a specific pathway P and a comprehensive signaling network $\mathcal{H}$, can we identify un-annotated reactions in $\mathcal{H}$ that are likely to be in P?

We reduce these problems to that of computing hyperpaths in a signaling hypergraph. We consider B-hyperpaths, which generalize paths in a directed graph by accounting for the fact that a reaction can occur only if all its reactants are present. A B-hyperpath from node s to node t has a natural interpretation in signaling pathways: the hyperpath contains all the intermediate reactants and products needed to "reach" t from s. Unlike shortest paths in graphs, shortest B-hyperpaths may contain cycles (see Section 3). We restrict our attention to acyclic B-hyperpaths in analogy to shortest path and related algorithms (e.g., Steiner trees) on

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
BCB September 20-23, 2014, Newport Beach, CA, USA
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-2894-4/14/09 ...$15.00.
http://dx.doi.org/10.1145/2649387.2649450.

graphs, which return acyclic networks. We acknowledge that feedback loops are important aspects of signaling pathways, and we expect to study cyclic B-hyperpaths in the future.

We make four major contributions in this work. First, we describe signaling hypergraphs, which represent complexes, complex assembly/disassembly, and positive regulation more accurately than corresponding graph representations. Second, we present several properties of B-hyperpaths and formulate the NP-complete problem of computing the acyclic B-hyperpath with the smallest number of hyperedges. Third, we define a mixed integer linear program (MILP) to solve this problem. Finally, we find optimal B-hyperpaths in signaling hypergraphs constructed from a signaling pathway database (summarized below). Although notions such as B-hyperpaths have been available since the early 1980s, our work appears to be the first to modify and apply these ideas to answer very natural questions on signaling pathways.

We compute B-hyperpaths in signaling hypergraphs of varying sizes from the National Cancer Institute's Pathway Interaction Database (NCI-PID) [25]. We focus on the Wnt signaling pathway, a well-studied pathway involved in development and often perturbed in cancer. Starting with the canonical Wnt signaling pathway, we identify acyclic B-hyperpaths that end in different forms of β-catenin that correspond to the absence and presence of Wnt signaling, answering Question 1 above. We then explore Question 1 on a more comprehensive Wnt signaling pathway by finding acyclic B-hyperpaths that connect membrane-bound complexes to downstream target genes (TCF1 and LEF1). We show that the resulting B-hyperpaths are much more informative than paths and Steiner trees on corresponding graph representations of the Wnt signaling pathway. Finally, we consider the annotated Wnt signaling pathway in the context of the entire NCI-PID dataset. To answer Question 2, we identify reactions that are not annotated to the Wnt pathway that connect it to the Androgen Receptor pathway.

2. RELATED RESEARCH

We discuss generalizations of signaling pathway graph representations and emphasize their strengths and limitations.

Representing complexes. Compound graphs permit a compound node to contain a set of nodes [9], e.g., a set of proteins in a complex. Similarly, metagraphs support scalable network structure by allowing metanodes to have a nested structure [15]. Compound nodes and metanodes may themselves be connected by edges. Similarly, undirected hypergraphs allow interactions among two or more entities [10]. Software that computes paths, loops, and motifs on compound graphs [8] and visualizes metagraphs [15] have accelerated the adoption of these representations.

Representing pathway directionality. A factor graph [12, 27] is a bipartite graph with partitions corresponding to molecules and to factors, which represent (potentially directed) reactions in a pathway. Factors are connected to the molecules that participate in the reaction. The PARADIGM software [27] uses probabilistic inference on factor graphs to estimate a pathway's activities from high-throughput data on molecular changes in cancer tissues. A Petri net [21, 22] is a directed bipartite graph with two types of nodes – places and transitions – and tokens on the places. In Petri net models of signaling pathways, places represent proteins, transitions represent reactions among proteins, and the number of tokens in a place represent the concentration of proteins. "Firing" a transition corresponds to redistributing the tokens based on certain rules.

Representing regulation. Influence graphs [24] are graphs where each edge has a sign describing one molecule's affect on the other. More generally, logic models [24] define logic functions on hyperedges with potentially multiple nodes in the tail but a single node in the head. Multi modal networks [13] are generalizations of hypergraphs that include a single regulator for each hyperedge. Finally, dynamic models (often based on ordinary differential equations) can describe the control mechanisms within signaling pathways faithfully [17].

Limitations of related work. A major limiting factor of compound graphs and metagraphs is that they connect pairs of entities, making interactions consisting of more than two entities (such as complex assembly and disassembly) difficult to model. Factor graphs and Petri nets are not ideal for generalizations of common graph-theoretic operations such as connectivity and paths, which is the focus of this paper. Influence graphs and logic networks represent protein regulation, but they operate only on the "active" forms of proteins. Moreover, it is unclear how they represent complex assembly/disassembly. Finally, the large amounts of experimental data needed by dynamic models in order to fit and tune parameters limits their scalability.

3. DEFINITIONS

3.1 Signaling Hypergraphs

Let V be a finite set of nodes. A *directed hyperedge* e is a pair $(T(e), H(e))$ where both the *tail* $T(e)$ and the *head* $H(e)$ are non-empty subsets of V. A *directed hypergraph* $\mathcal{H} = (V, E)$ consists of a finite set V of nodes and a finite set E of directed hyperedges. $\mathcal{H}$ is a directed graph in the special case where $|T(e)| = |H(e)| = 1$.

At first glance, directed hypergraphs seem sufficient for representing signaling reactions: each hyperedge consists of a set of reactants in the head and a set of products in the tail. However, many signaling reactions involve protein complexes, where a set of proteins act as a single unit in a reaction. Further, directed hypergraphs do not represent molecules that regulate a reaction (e.g., a kinase that phosphorylates, and subsequently activates, a substrate).

To model complexes, we define a *hypernode*[1] as a set of nodes $U \subseteq V$ that act together as a single unit. U may contain a single node, e.g., to represent a protein that acts on its own. We use $\mathcal{V}$ to denote the set of hypernodes, assuming that each node in V appears in some hypernode in $\mathcal{V}$. We define a *signaling hyperedge* e to be a pair $(T(e), H(e))$ where both the *tail* $T(e)$ and the *head* $H(e)$ are non-empty subsets of $\mathcal{V}$, i.e., each member of the tail or the head is a hypernode.

To model positive regulation, we represent each positive regulator as a hypernode. If a hypernode U is a positive regulator for a reaction, we add U to the tail of the signaling hyperedge representing that reaction. Signaling hyperedges can represent the logic of multiple positive regulators, e.g., if all positive regulators must be present for the reaction to occur, we add all the regulators to the tail of the signaling hyperedge. Alternatively, if any of the positive regulators

[1]Hypernodes may be referred to as undirected hyperedges, compound nodes [9, 8], or metanodes [15] in the literature.

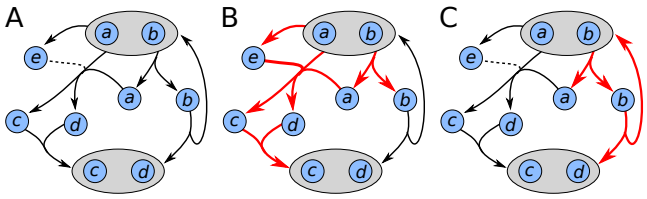

Figure 1: (A) A signaling hypergraph. Dashed portions of hyperedges denote positive regulation and are for visualization purposes only. (B) An acyclic B-hyperpath and (C) a cyclic B-hyperpath from {a,b} to {c,d} (red hyperedges).

can trigger the reaction, we can make copies of the signaling hyperedge, one for each regulator.

We define a *signaling hypergraph* $\mathcal{H} = (V, \mathcal{V}, \mathcal{E})$, where V is a finite set of nodes, $\mathcal{V} \subseteq 2^V$ is a set of hypernodes and $\mathcal{E}$ is a finite set of signaling hyperedges. Figure 1A illustrates a signaling hypergraph with five nodes (a,b,c,d,e), seven hypernodes, and five hyperedges. The gray circles denote hypernodes containing more than one element (e.g. {a,b}). When it is clear from the context, we will refer to signaling hyperedges and signaling hypergraphs simply as hyperedges and hypergraphs, respectively.

Scope of signaling hypergraph representations. Signaling hypergraphs generalize earlier research by simultaneously representing reactions among more than two molecules, complexes, and combinatorial positive regulation. They also model complex rearrangement and post-translational modifications. Signaling hypergraphs, as they are defined here, do not yet represent negative regulation or more complex regulatory logic. We expect to address these important aspects of signaling pathways in the future.

3.2 B-Hyperpaths

There are numerous ways to define paths in directed hypergraphs [1, 26]. In this section, we describe how to extend these ideas to signaling hypergraphs. One intuitive notion is a straightforward generalization of a path in a directed graph. An *s-t path* $P(s,t)$ is an alternating sequence of hypernodes and hyperedges starting at hypernode $s \in \mathcal{V}$ and terminating a hypernode $t \in \mathcal{V}$, i.e.,

$$P(s,t) = (U_1, e_1, U_2, \ldots, U_{k-1}, e_{k-1}, U_k)$$

where $s = U_1$, $t = U_k$, and for every $1 \leq i \leq k$, $U_i \in T(e_i)$ and $U_{i+1} \in H(e_i)$ [1]. We say that a path $P(s,t)$ is *simple* if it contains no repeated hypernodes or hyperedges and that $P(s,t)$ is a *simple cycle* if U_1 and U_k are both in the tail of e_1. We say that $\mathcal{H}$ is *acyclic* if it does not contain any simple cycles for any pair of hypernodes $s, t \in \mathcal{V}$.

Since simple paths report an alternating sequence of hypernodes and hyperedges, they do not capture all the hypernodes associated with each hyperedge in the path. Thus, they are not useful for representing sequences of signaling reactions that involve multiple reactants and/or products. We use formalisms developed in the hypergraph literature [1, 26] to describe the notion that in order for all products of a signaling reaction to be present, *all* reactants must be present.

For a hypernode $U \in \mathcal{V}$, the *backward star* $BS(U)$ of U is the set of hyperedges e for which $U \in H(e)$. Given a hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ and a hypernode $s \in \mathcal{V}$, we say that hypernode $U \in \mathcal{V}$ is *B-connected* to s in $\mathcal{H}$ if either (i) $U = s$ or (ii) there exists a hyperedge $e \in BS(U)$ such that, for all $W \in T(e)$, W is B-connected to s. We use the notation $\mathcal{B}_{\mathcal{H}}(s)$ to denote the set of hypernodes that are B-connected to s in $\mathcal{H}$. Note that positive regulators fit into the definition of B-connected in a biologically meaningful way: in order for all products of a signaling reaction to be present, *all* the reactants and *all* the positive regulators in the tail of the reaction must be present.

A *sub-hypergraph* $\mathcal{H}' = (\mathcal{V}_{\mathcal{H}'}, \mathcal{E}_{\mathcal{H}'})$ of $\mathcal{H}$ consists of a subset of hypernodes $\mathcal{V}_{\mathcal{H}'} \subseteq \mathcal{V}$ and a subset of hyperedges $\mathcal{E}_{\mathcal{H}'} \subseteq \mathcal{E}$ of $\mathcal{H}$, with the property that for every hyperedge $e \in \mathcal{E}_{\mathcal{H}'}$, the hypernodes in $T(e)$ and $H(e)$ are members of $\mathcal{V}_{\mathcal{H}'}$. Given $\mathcal{H}$ and two hypernodes $s, t \in \mathcal{V}$, an *s-t B-hyperpath* $\Pi(s,t)$ is a sub-hypergraph of $\mathcal{H}$ such that $t \in \mathcal{B}_{\Pi(s,t)}(s)$ and $\Pi(s,t)$ is minimal with respect to the deletion of hypernodes and hyperedges. Note that we require that t be B-connected to s using only the hypernodes and hyperedges in $\Pi(s,t)$ itself. We say that $\Pi(s,t)$ is *acyclic* if it contains no simple cycles. Figure 1B depicts an acyclic B-hyperpath and Figure 1C depicts a B-hyperpath that contains a simple cycle.

We conclude our definitions with a concept that is analogous to a topological ordering of a directed acyclic graph. Given a hypergraph $\mathcal{H}$, an *ordering* $o : \mathcal{V} \mapsto \mathbb{R}$ of the hypernodes in $\mathcal{H}$ is function that maps each hypernode in $\mathcal{V}$ to a real number. We say that o is a *valid ordering* with respect to $\mathcal{H}$, if for every $e \in \mathcal{E}$ and for every pair of hypernodes $U \in T(e)$ and $W \in H(e)$, $o(U) < o(W)$ [4].

3.3 Problem Statement

There may be many s-t B-hyperpaths in a hypergraph, as Figure 1 illustrates. We wish to find hyperpaths that represent a minimal set of reactions that lead from s to t. In other words, we seek to compute a B-hyperpath $\Pi(s,t)$ of $\mathcal{H}$ with the smallest number of hyperedges:

$$\Pi(s,t) = \underset{\Pi : t \in \mathcal{B}_{\Pi}(s)}{\arg\min} |\mathcal{E}_{\Pi}| \quad (1)$$

We can assign costs to the hyperedges and compute the B-hyperpath with the smallest cost. In this work, we simply count hyperedges since NCI-PID is manually curated.

Finding the hyperpath $\Pi(s,t)$ that minimizes Eq. (1) is NP-hard by reduction from Minimum Set Cover [1], even when $\mathcal{H}$ is a directed hypergraph (i.e., each hypernode contains exactly one node), and we seek only acyclic hyperpaths. Given a universe of n elements and k sets, this reduction involves the construction of a hypergraph where one of the hyperedges contains all n elements in its tail. Since reactions in signaling pathways are unlikely to involve a very large number of proteins, we are interested in computing minimum acyclic B-hyperpaths in *k-hypergraphs*, where each hyperedge has at most k hypernodes in its tail or in its head. We can prove that even this more biologically-valid version of the problem is also NP-complete.

THEOREM 1. *Finding the hyperpath $\Pi(s,t)$ of a signaling hypergraph $\mathcal{H} = (V, \mathcal{V}, \mathcal{E})$ where each hyperedge $e \in \mathcal{E}$ has at most k hypernodes in $T(e)$ or in $H(e)$ is NP-hard for $k \geq 3$.*

We can prove Theorem 1 by reduction from Minimum k-Set Cover; due to lack of space, we omit the proof.

3.4 Properties of Acyclic B-Hyperpaths

In this section, we state several properties of B-hyperpaths and B-connectedness, which we will use in Section 4 to prove

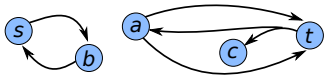

Figure 2: A hypergraph that satisfies constraints (2)-(5) where t is not B-connected to s.

the correctness of our algorithm. We omit most proofs in this section and the next due to lack of space.

First, we relate a valid ordering and the acyclicity of a hypergraph $\mathcal{H}$. Similar to a topological ordering in DAGs, a hypergraph $\mathcal{H} = (V, \mathcal{V}, \mathcal{E})$ has a valid ordering iff it is acyclic. Second, if a hypergraph is acyclic, then there must be a hypernode with no incoming hyperedges. In other words, given an acyclic hypergraph $\mathcal{H} = (V, \mathcal{V}, \mathcal{E})$, there exists some hypernode $U \in \mathcal{V}$ such that $BS(U) = \emptyset$.

We now turn our attention to B-hyperpaths in $\mathcal{H}$. Since B-hyperpaths are defined using the notion of B-connection, all hypernodes in a B-hyperpath $\Pi(s,t)$ are B-connected to s in the sub-hypergraph $\Pi(s,t)$. More formally, if $\Pi(s,t)$ is a B-hyperpath, then $\mathcal{B}_{\Pi(s,t)}(s) = \mathcal{V}_{\Pi(s,t)}$. A corollary of this statement is that if $\Pi(s,t)$ is a B-hyperpath in $\mathcal{H}$, then $\mathcal{V}_{\Pi(s,t)} \subseteq \mathcal{B}_{\mathcal{H}}(s)$.

In this work, our goal is to compute acyclic B-hyperpaths. Observe that B-hyperpaths need not be acyclic, even though they are minimal with respect to deletion of hypernodes and hyperedges. For example, the B-hyperpath defined by the red hyperedges in Figure 1C contains the simple cycle $(\{a,b\}, b, \{a,b\})$. Since there are no simple cycles in acyclic B-hyperpaths, we can characterize the "beginning" of the B-hyperpath as the set of hypernodes that have an empty backward star. The final claim states that for an acyclic B-hyperpath $\Pi(s,t)$, s is the only hypernode in $\Pi(s,t)$ that does not occur in the head of any hyperedge in $\Pi(s,t)$.

4. ALGORITHMS

For ease of exposition, we describe our approach to computing minimum s-t B-hyperpaths in two parts. First, we develop an MILP to compute an acyclic B-connected sub-hypergraph that contains s and t. Although we can compute such a sub-hypergraph in polynomial time [11], we present an MILP so that we can augment it with an objective function in the second part to solve the NP-complete problem of computing optimal B-hyperpaths.

4.1 Acyclic B-Connected Sub-Hypergraphs

Given a hypergraph $\mathcal{H}$ and two hypernodes s and t in $\mathcal{V}_{\mathcal{H}}$, we wish to compute an acyclic sub-hypergraph $\mathcal{H}'$ that contains s and t such that $t \in \mathcal{B}_{\mathcal{H}'}(s)$, i.e. t is B-connected to s in $\mathcal{H}'$. Note that $\mathcal{H}'$ may not be an s-t B-hyperpath because it is not necessarily minimal.

First, for every $U \in \mathcal{V}$, we introduce a binary (0-1) variable α_U. We also introduce a binary variable α_e for every hyperedge $e \in E$. The output sub-hypergraph $\mathcal{H}'$ is defined by the values of these variables: the hypernode U (respectively, hyperedge e) is in $\mathcal{H}'$ iff $\alpha_U = 1$ (resp., $\alpha_e = 1$). Given a setting of the variables in α, we will henceforth refer to the corresponding sub-hypergraph as $\mathcal{H}(\alpha)$ and to the hypernodes and hyperedges in this sub-hypergraph as $\mathcal{V}(\alpha)$ and $\mathcal{E}(\alpha)$, respectively. The α variables must satisfy the following linear constraints:

$$\forall\, U \in \mathcal{V} \setminus \{s\}: \quad \begin{cases} \alpha_U \leq \sum_{e \in BS(U)} \alpha_e & \text{if } BS(U) \neq \emptyset \\ \alpha_U = 0 & \text{otherwise} \end{cases} \tag{2}$$

$$\forall\, e \in E: \quad \sum_{U \in T(e)} \alpha_U \geq |T(e)|\alpha_e \tag{3}$$

$$\forall\, e \in E: \quad \sum_{U \in H(e)} \alpha_U \geq |H(e)|\alpha_e \tag{4}$$

$$\alpha_t = 1. \tag{5}$$

These constraints have the following meaning. With the exception of hypernode s, every hypernode U such that $\alpha_U = 1$ has at least one incoming hyperedge e such that $\alpha_e = 1$ (constraint (2)). For every hyperedge e such that $\alpha_e = 1$, all hypernodes in the tail (constraint (3)) and the head (constraint (4)) must have α values of one. Constraints (2) and (3) encode the definition of B-connectedness, while constraint (3) says that if we can ensure that all hypernodes in the tail of a hyperedge are B-connected to s, then we automatically ensure the B-connectedness of all hypernodes in the head. Finally, we require that t is in $\mathcal{H}(\alpha)$ (constraint (5)).

Together, constraints (2)–(5) seek to ensure that t is in $\mathcal{H}(\alpha)$ and that all the hypernodes in $\mathcal{H}(\alpha)$ are B-connected to s. However, disconnected sub-hypergraphs with cycles may satisfy the constraints (Figure 2). To address this issue, we introduce a real-valued order variable o_U for each $U \in \mathcal{V}$. We ensure that o defines a valid ordering with respect to $\mathcal{H}(\alpha)$ through the following constraint, which ensures that for each edge in $\mathcal{H}(\alpha)$, every hypernode in the tail of the edge must have an order value smaller than the order value of every hypernode in the head of the edge:

$$\forall\, e \text{ such that } \alpha_e = 1; \forall (U,W) \in T(e) \times H(e): \quad o_U < o_W.$$

These constraints apply only to those edges e where $\alpha_e = 1$. Furthermore, linear programs require weak inequalities to define boundary regions. To address both points, we introduce two constants: ϵ, which takes a very small value and C, which takes a very large value. The following linear constraint applies to all edges in $\mathcal{H}$:

$$\forall e \in E; \forall (U,W) \in T(e) \times H(e): o_U \leq o_W - \epsilon + C(1-\alpha_e). \tag{6}$$

Eq. (6) is only enforced when $\alpha_e = 1$ for hyperedge e; when $\alpha_e = 0$, the large constant C dominates the right hand side, trivially satisfying the inequality. We relax the strict inequality by requiring that o_W is at least ϵ larger than o_U.[2]

Given a hypergraph $\mathcal{H}$ and two hypernodes s and t in $\mathcal{H}$, we say that an assignment of α and o variables is *feasible* if it simultaneously satisfy the constraints in Equations (2)-(6). Given a feasible assignment, we make a number of claims about the resulting sub-hypergraph $\mathcal{H}(\alpha)$ using the lemmas presented in Section 3.4.

LEMMA 1. *$\mathcal{H}(\alpha)$ is acyclic.*

LEMMA 2. *The only hypernode with an empty backward star in $\mathcal{H}(\alpha)$ is s.*

LEMMA 3. *Hypernode s has the smallest value for the order variable in $\mathcal{H}(\alpha)$; i.e., $s = \arg\min_{U \in \mathcal{H}(\alpha)} o_U$.*

From these claims, we prove the following lemma:

LEMMA 4. *All hypernodes in $\mathcal{V}(\alpha)$ are B-connected to hypernode s in $\mathcal{H}(\alpha)$, i.e., $\mathcal{B}_{\mathcal{H}(\alpha)}(s) = \mathcal{V}(\alpha)$.*

[2]To reduce the search space for the MILP, we bound the order variables so that $o_U \in [0,1]$ for all hypernodes $U \in \mathcal{V}$.

PROOF. We will use strong induction on the order variables. Without loss of generality, rename the hypernodes in $\mathcal{V}(\alpha)$ by increasing value of the order variables, so that $o_{U_i} < o_{U_{i+1}}$, for all $1 \leq i < n$, where $n = |\mathcal{V}(\alpha)|$. Note that $s = U_1$ by Lemma 3. By the definition of B-connected, s is B-connected to itself, establishing the base case. Now consider hypernode U_2. Eq. (2) requires that U_2 must have at least one hyperedge e in its backward star in $\mathcal{H}(\alpha)$. Constraint (6) requires that if U_i is a hypernode in the tail of e, for some value of i, then $o_{U_i} < o_{U_2}$. The only possible value of i is 1. Thus, there must exist a hyperedge e such that $T(e) = \{s\}$ and $U_2 \in H(e)$, proving that U_2 is B-connected to s in $\mathcal{H}(\alpha)$.

For the inductive hypothesis, we assume that hypernodes $U_1, U_2, \ldots, U_{k-2}, U_{k-1}$ are B-connected to s in $\mathcal{H}(\alpha)$. To prove that U_k is also B-connected to s in $\mathcal{H}(\alpha)$, we must show that there exists a hyperedge $e \in \mathcal{E}(\alpha)$ such that $U_k \in H(e)$ and every hypernode $W \in T(e)$ is B-connected to s. Constraint (2) requires that there exists some hyperedge $e \in \mathcal{E}(\alpha)$ that is in the backward star of U_k. Now constraint (3) applies to e. Therefore, all hypernodes in $T(e)$ are in $\mathcal{V}(\alpha)$. Finally, Eq. (6) requires that for any hypernode U_i in $T(e)$, $o_{U_i} < o_{U_k}$, i.e., $i < k$. Therefore, by the inductive hypothesis, each hypernode in $T(e)$ is B-connected to s. Together, these statements establish that hypernode U_k is B-connected to s. □

Observe that t is in $\mathcal{V}(\alpha)$ because we fix α_t to 1 in constraint (5); thus, o_t will have a value and we will consider t in the inductive proof, leading to the following corollary.

COROLLARY 5. *Hypernode t is B-connected to hypernode s in $\mathcal{H}(\alpha)$, i.e., $t \in \mathcal{B}_{\mathcal{H}(\alpha)}(s)$.*

From the proof of Lemma 4, we also see that there must be a hyperedge in $\mathcal{E}(\alpha)$ connecting s to U_2 (and possibly other nodes). Thus, there is at least one hyperedge e in $\mathcal{E}(\alpha)$ such that hypernode $s \in T(e)$, which allows us to prove the following lemma.

LEMMA 6. *$\mathcal{H}(\alpha)$ contains an acyclic s-t B-hyperpath as a sub-hypergraph.*

The previous lemmas establish that if the MILP has a feasible solution, then $\mathcal{H}(\alpha)$ is acyclic, contains both s and t, and that all hypernodes in $\mathcal{V}(\alpha)$ are B-connected to s in $\mathcal{H}(\alpha)$. Moreover, $\mathcal{H}(\alpha)$ contains an acyclic s-t B-hyperpath. The next lemma establishes the inverse property: if the hypergraph $\mathcal{H}$ contains an acyclic s-t B-hyperpath, then the MILP has a feasible assignment.

LEMMA 7. *If $\mathcal{H}$ contains an acyclic B-hyperpath $\Pi(s,t) = (\mathcal{V}_{\Pi(s,t)}, \mathcal{E}_{\Pi(s,t)})$, then there is a feasible assignment where $\mathcal{H}(\alpha) = \Pi(s,t)$.*

PROOF. (Sketch) We can easily construct the feasible assignment that corresponds to $\mathcal{H}' = \Pi(s,t)$. Define an assignment A of the α variables as follows:

$$\alpha_U = \begin{cases} 1 & \text{if } U \in \mathcal{V}_{\Pi(s,t)} \\ 0 & \text{otherwise.} \end{cases} \qquad \alpha_e = \begin{cases} 1 & \text{if } e \in \mathcal{E}_{\Pi(s,t)} \\ 0 & \text{otherwise.} \end{cases}$$

It is not difficult to show that the assignment A satisfies constraints (2)–(5). To complete the proof, we can use the fact that a valid ordering exists for $\Pi(s,t)$ (it is acyclic) to determine an assignment for the order variables in the MILP that satisfy constraint (6). □

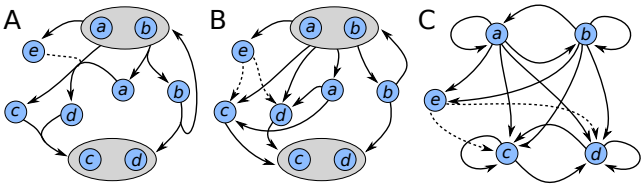

Figure 3: Converting a hypergraph $\mathcal{H}$ (A) into a graph with complexes (B) and a graph (C).

4.2 Minimum Acyclic B-Hyperpaths

We now augment the MILP developed so far with an objective function in order to compute a minimum acyclic s-t B hyperpath, i.e., one that minimizes Eq. (1). We compute an assignment of variables that solves the following optimization problem:

$$\arg\min_{\alpha,o} \sum_{e \in E} \alpha_e \text{ subject to constraints (2)-(6).} \tag{7}$$

The following theorem captures our main result.

THEOREM 2. *Given a hypergraph $\mathcal{H} = (V, E, c)$ and two hypernodes $s, t \in V$, a feasible solution of (7) is an acyclic s-t B-hyperpath $\Pi(s,t)$ with minimal cost $C(\Pi(s,t))$ over all acyclic s-t B-hyperpaths in $\mathcal{H}$.*

PROOF. (Sketch) Let $\mathcal{H}(\alpha)$ be the sub-hypergraph of $\mathcal{H}$ that minimizes the objective function in (7). We can prove that $\mathcal{H}(\alpha)$ is acyclic (Lemma 1), that $t \in \mathcal{B}_{\mathcal{H}(\alpha)}(s)$ (Corollary 5), and that $\mathcal{H}(\alpha)$ is minimal with respect to the deletion of hyperedges and hypernodes (since $\mathcal{H}(\alpha)$ is the optimal solution to the MILP).

Thus, the sub-hypergraph $\mathcal{H}(\alpha)$ that minimizes the objective function (7) is indeed an acyclic s-t B-hyperpath $\Pi(s,t)$. The value of the objective function for $\Pi(s,t)$ is $|\mathcal{E}_{\Pi(s,t)}|$ because $\alpha_e = 1$ for all $e \in E_{\Pi(s,t)}$. Therefore, $\Pi(s,t)$ has the smallest number of hyperedges over all acyclic s-t B-hyperpaths in $\mathcal{H}$. □

4.3 Conversion to Graph Representations

For the purpose of comparing signaling hypergraphs to graphs, we convert each signaling hypergraph $\mathcal{H} = (V, \mathcal{V}, \mathcal{E})$ to two different graph representations (Figure 3). First, we build a directed graph with complexes $G_C = (\mathcal{V}, \mathcal{E}_{G_C})$ whose nodes are the hypernodes in $\mathcal{H}$ and where $\mathcal{E}_{G_C}$ consists of all pairs of hypernodes in the tail and head of each hyperedge in $\mathcal{H}$, i.e., $\mathcal{E}_{G_C} = \bigcup_{e \in \mathcal{E}} \{T(e) \times H(e)\}$. Note that each edge in $\mathcal{E}_{G_C}$ connects exactly two hypernodes (Figure 3B). The graph with complexes is akin to a compound graph, except that it does not explicitly represent the nested structure of complexes. However, it is not difficult to compute this structure from the hypernodes.

Second, we convert the graph with complexes G_C into a graph $G = (V, E_G)$ (Figure 3C). The nodes of G_C are identical to the set of hypernodes of $\mathcal{H}$. The edges in E_G are the union of two sets of edges: (a) For each hypernode U in $\mathcal{V}$, we connect all pairs of nodes in U by an undirected edge, corresponding to the common practice of representing a complex by a clique in a graph. (b) For each edge (U, V) in G_C, we connect every node in the hypernode U to every node in the hypernode in V by a directed edge. Finally, we replace every undirected edge by two directed edges.

	Small Wnt	Large Wnt	NCI-PID
# Pathways	2	6	213
# Neg. Regs	6	25	856
Signaling Hypergraph $\mathcal{H}$			
# Nodes	47	304	6793
# Hypernodes	57	354	8779
# Hyperedges	34	223	7735
Graph With Complexes G_C			
# Nodes	47	304	6793
# Hypernodes	57	354	8779
# Edges	80	541	15622
Graph G			
# Nodes	47	304	6793
# Edges	294	1435	40346

Table 1: Selected Signaling Pathways for Analysis. Negative regulators (second row) are ignored in each pathway.

4.4 NCI-PID Pathways

NCI-PID [25] contains curated human pathways that include biochemical reactions, complex assembly, cellular compartment transport, transcriptional regulation, and regulation of biological processes. We focused on the Wnt signaling pathway, in part due to its central role in development and a number of cancers.

We automatically constructed three sets of signaling hypergraphs by combining different sets of signaling pathways annotated in NCI-PID: a small Wnt signaling pathway, a large Wnt signaling pathway, and the entire set of all NCI-PID pathways. The small Wnt signaling pathway consisted of the union of two NCI-PID pathways: "degradation of β-catenin" and "canonical Wnt signaling". The large Wnt signaling pathway included four additional NCI-PID pathways: "noncanonical Wnt signaling," "Wnt signaling network," "regulation of nuclear β-catenin" and "presenilin action in Notch and Wnt signaling", which corresponded to non-canonical branches of Wnt signaling. The NCI-PID pathways are freely available in BioPAX format [7], and we processed them using an in-house parser built upon PaxTools [6]. Signaling hypergraphs do not currently support negative regulation; thus negative regulators were ignored (Table 1). NCI-PID represents complexes as sets of unique NCI-PID protein IDs; thus we were able to extract complexes and parse them as hypernodes. Multiple forms of the same protein may be present with attributes such as compartmentalization, activation, and post-translational modifications (PTMs) such as phosphorylation and ubiquitination. We treated each variant as a distinct entity. We used this information to analyze and visualize our results, as the reader can see in the figures in Section 5.

Table 1 displays statistics on these three sets of pathways when represented as signaling hypergraphs $\mathcal{H}$ and upon conversion to graphs with complexes G_C and graphs G. For the full NCI-PID database, the number of edges in G_c is twice the number of signaling hyperedges in $\mathcal{H}$; the number of edges in G is about five times as many. These statistics suggest that the information loss incurred upon making these conversions of signaling hypergraphs is accompanied by a significant inflation in the number of edges.

5. RESULTS

Given a signaling hypergraph $\mathcal{H} = (V, \mathcal{V}, \mathcal{E})$ and two hypernodes $s, t \in \mathcal{V}$, we wished to compute s-t B-hyperpaths with the smallest number of hyperedges in $\mathcal{H}$. We outline the general procedure for computing B-hyperpaths in $\mathcal{H}$ in Algorithm 1. First, we computed all hypernodes that were B-connected to s in $\mathcal{H}$ [1] and the sub-hypergraph $\mathcal{H}'$ of $\mathcal{H}$ induced by these hypernodes, returning an infeasible solution if t was not B-connected to s (lines 1-5). We then solved for α and o variables using the MILP that optimizes Eq. (7) (Section 4), and stored the optimal objective (lines 6-7). Since there may be many B-hyperpaths with the same number of hyperedges, we iteratively solved the MILP after adding a constraint that forced a new B-hyperpath (line 11). We returned all the B-hyperpaths with the smallest number of hyperedges.

Algorithm 1 RunMILP($\mathcal{H}$,s,t)

Require: $\mathcal{H} = (V, \mathcal{V}, \mathcal{E}); s \in \mathcal{V}, t \in \mathcal{V}$
1: $\mathcal{B}_{\mathcal{H}}(s)$:= Set of hypernodes in S that are B-connected to s
2: $\mathcal{H}'$:= sub-hypergraph of $\mathcal{H}$ induced on $\mathcal{B}_{\mathcal{H}}(s)$
3: **if** t is not in $\mathcal{B}_{\mathcal{H}}(s)$ **then**
4: **return** Infeasible Assignment
5: **end if**
6: α, o := Solve Eq. (7) on $\mathcal{H}'$, s, and t
7: opt := $|\mathcal{E}(\alpha)|$
8: $R := \emptyset$
9: **while** $|\mathcal{E}(\alpha)|$ = opt **do**
10: $R := R \cup \mathcal{H}(\alpha)$
11: Add constraint such that $\sum_{e \in \mathcal{E}_{\mathcal{H}'}} \alpha_e < |\mathcal{E}(\alpha)|$.
12: α, o := Re-solve the MILP on $\mathcal{H}'$, s, and t
13: **end while**
14: **return** R

We applied this procedure to signaling hypergraphs as well as their graph-with-complexes and graph counterparts. We acknowledge that the minimal acyclic B-hyperpath path problem can be solved in polynomial time using Dijkstra's algorithm. However, we continued to use the MILP approach to ensure uniformity of analysis across all inputs.

5.1 Small Wnt Signaling Pathway

The NCI-PID pathway describing the degradation of β-catenin terminates at ubiquitinated β-catenin. The NCI-PID canonical Wnt signaling pathway terminates at nuclear β-catenin, which is a transcriptional co-regulator. To answer Question 1 from the introduction, we asked what reactions terminate at the (a) ubiquitinated form of β-catenin and (b) the nuclear form of β-catenin.

We made the following modifications to the small Wnt signaling pathway before applying the MILP. We introduced a source hypernode s and connected s to 21 hypernodes with an empty backward star. We also connected s to a hypernode representing a complex of APC, Axin1, and β-catenin; this complex is part of a cycle involving cytoplasmic β-catenin, and including the hypernode in the set of sources "breaks" this loop. Finally, we removed one self-loop that contained the same form of β-catenin in the head and the tail. The modified signaling hypergraph consisted of 58 hypernodes, 48 nodes, and 56 hyperedges. All hypernodes and

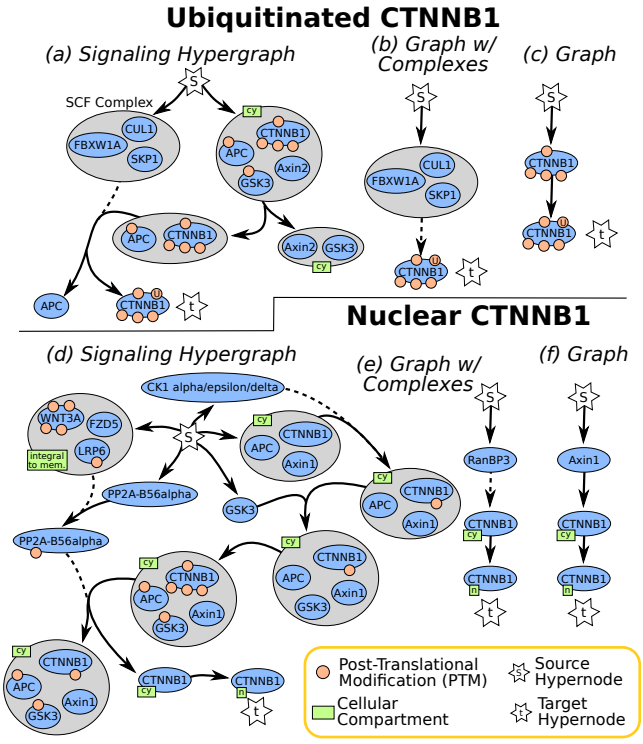

Figure 4: Optimal B-hyperpaths in the small Wnt signaling pathway to (a)-(c) ubiquitinated β-catenin, denoted by 'U,' and (d)-(f) nuclear β-catenin.

hyperedges were B-connected to s.[3] We computed the optimal B-hyperpaths from the source hypernode s (i) to the ubiquitinated form of β-catenin and (ii) to the nuclear form of β-catenin.

Reactions involved in ubiquitinated β-catenin. The MILP returned one optimal B-hyperpath with four hyperedges (Figure 4(a)). Since β-catenin is marked for degradation in the absence of Wnt signaling, the absence of Wnt proteins from this B-hyperpath was not surprising. The APC/GSK3/Axin2/β-catenin complex splits to produce two smaller complexes: APC/β-catenin and Axin2/GSK3. The SCF ubiquitin ligase complex composed of CUL1, SKP1, and an F-box protein then splits the APC/β-catenin complex and ubiquitinates β-catenin, marking it for degradation.

In the graph-with-complexes representation, there was a single path of length two from s to ubiquitinated β-catenin through the SCF complex (Figure 4(b)). This path corresponded to a simple path in the signaling hypergraph. In the graph representation, there were five paths of length two. Each of the first three paths was a simple path in the signaling hypergraph and contained one of the members of the SCF complex (e.g., the path (s,CUL1,ubiquitinated β-catenin)). However, the other two paths, through APC and through phosphorylated β-catenin, were *not* simple paths in the signaling hypergraph. The graph representation collapsed the two complexes in the solution for signaling hypergraphs, yielding the path from s to ubiquitinated β-catenin through phosphorylated β-catenin (Figure 4(c)).

[3]We say that a hyperedge is B-connected to s if all hypernodes in its tail are B-connected to s.

Reactions involved in nuclear import of β-catenin. The MILP returns a single optimal B-hyperpath consisting of 11 hyperedges (Figure 4(d)). Here, Wnt signaling is necessary for the formation of the WNT3A/FZD5/LRP6 complex at the cell membrane, which activates a regulator (PP2A-B56α) that dissociates β-catenin from its complex with the destruction box APC/Axin1/GSK3[4] and dephosphorylates β-catenin, which in turn translocates to the nucleus. The optimal B-hyperpath also contains details about the formation of the APC/GSK3/Axin1/β-catenin complex: first, CK1 family proteins phosphorylate β-catenin in the APC/Axin1/β-catenin complex, and GSK3 then joins the complex and becomes activated.

In the graph-with-complexes representation, there is a single path of length three from s to nuclear β-catenin that contains RanBP3 (Figure 4(e)). RanBP3 is in the Wnt signaling pathway because it aids in the nuclear export of β-catenin *back* to the cytoplasm [14]; thus, this path is misleading in this context. There are seven paths of length three from s to nuclear β-catenin in the graph representation; the path through RanBP3 is the only one that corresponds to a simple path in the signaling hypergraph. The other paths (through WNT3A, Axin1, GSK3, APC, FZD5, and phosphorylated β-catenin) are all present in multiple complexes that are collapsed in the graph representation. The path through Axin1 is shown in Figure 4(f).

5.2 Large Wnt Signaling Pathway

TCF1 and LEF1, transcription factors involved in Wnt signaling, are also downstream targets of Wnt. To answer Question 1 for the large Wnt signaling pathway, we computed minimum acyclic B-hyperpaths to identify reactions that regulate the transcription of genes *tcf1* and *lef1*.

We introduced a source hypernode s and connected it to 149 hypernodes with an empty backward star. We also connected s to the same hypernode in the cycle involving cytoplasmic β-catenin as for the small Wnt signaling pathway, and removed eight self-loops. The modified signaling hypergraph consisted of 356 hypernodes, 306 nodes, and 374 hyperedges, of which 354 hypernodes and 372 hyperedges were B-connected to s. To identify a series of reactions that regulate both *tcf1* and *lef1* gene transcription, we added a target hypernode t and a single hyperedge ({TCF1, LEF1},t) to the signaling hypergraph. For t to be B-connected to s, both TCF1 and LEF1 must be B-connected to s.

There were four optimal B-hyperpaths in the signaling hypergraph (Figure 5(a) and Table 2). The B-hyperpaths shared a majority of the hyperedges, which established that nuclear β-catenin is B-connected to s. These hyperedges were identical to those used to connect nuclear β-catenin to s in the small Wnt pathway (Figure 4) except that they included the formation of the WNT3A/FZD5/LRP6 complex. The four B-hyperpaths differed in the complexes containing TLE and TCF family proteins that bind to the promoter regions of LEF1 and TCF genes (Table 2). For example, the transcription factor TCF1E can be replaced by TCF4E in Figure 5 to regulate LEF1 transcription.

The five optimal paths from s to t in the graph-with-complexes representation echoed these differences, though they did not contain the steps to transport β-catenin to the

[4]This reaction in NCI-PID does have one copy of β-catenin among the reactants and two copies of β-catenin among the products.

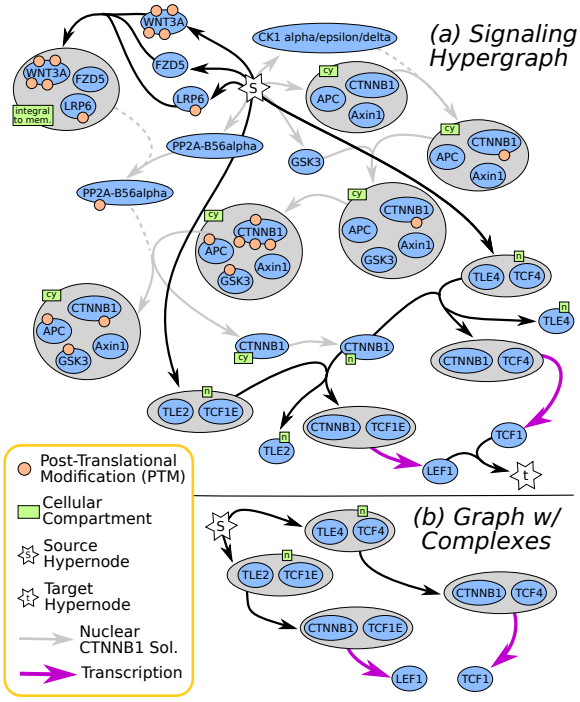

Figure 5: (a) Optimal B-hyperpath in the signaling hypergraph and (b) Steiner tree in the graph with complexes. The Steiner tree was comprised of two of the five optimal paths in the graph with complexes representation (Table 2).

nucleus (Table 2). The graph representations connected s to t directly through LEF1 and TCF1, since LEF1 and TCF1 also happen to be members of complexes with empty backward stars (Table 2). We also explored Steiner trees in the graph with complexes. Steiner trees find a sub-graph that span a set of terminal hypernodes in a graph, which are s, LEF1, and TCF1 in our case. We computed the Steiner tree with the smallest number of edges connecting s to TCF1 and to LEF1 in the graph-with-complexes representation.[5] The resulting Steiner trees contained six edges of edge-disjoint simple paths from s to LEF1 and to TCF1. Figure 5(b) illustrates one of these Steiner trees. The Steiner tree, like all of the shortest paths in the graph-with-complexes representation, did not include the transport of cytoplasmic β-catenin to the nucleus, a crucial component of TCF1 and LEF1 transcriptional activation.

5.3 Full NCI-PID Signaling Pathway

Finally, we analyzed the full NCI-PID signaling pathway to address Question 2 from the introduction. We asked the following complementary two questions:

- **New Sources to Known Targets:** are there reactions currently not annotated to the Wnt pathway that are connected to transcriptional regulators in the Wnt signaling pathway?
- **Known Sources to New Targets:** do reactions in the Wnt pathway connect to transcriptional regulators or factors that are not currently annotated to the Wnt signaling pathway?

[5]We used MSGSteiner [2] to find a prize-collecting Steiner tree that includes all terminals.

	Shortest B-Hyperpath/Path from s	#[1]
$\mathcal{H}$	(see Figure 5)	21
	{TLE4,TCF4} replaced by {TLE2,TCF4}[2]	21
	{TLE2,TCF1E} replaced by {TLE4,TCF4E}[2]	21
	{TLE2,TCF1E} replaced by {TLE2,TCF4E}, and {CTNNB1,TCF1E} replaced by {CTNNB1,TCF4E}[2]	21
G_C	(s,{TLE2,TCF1E},{CTNNB1,TCF1E},LEF1,t)	4
	(s,{TLE4,TCF4},{CTNNB1,TCF4},TCF1,t)	4
	(s,{TLE4,TCF4E},{CTNNB1,TCF4E},LEF1,t)	4
	(s,PITX2,{CTNNB1,LEF1,PITX2},LEF1,t)	4
	(s,{TLE2,TCF4},{CTNNB1,TCF4},TCF1,t)	4
G	(s,LEF1,t)	2
	(s,TCF1,t)	2

[1] Number of hyperedges in optimal solution.
[2] Major differences compared to solution in Figure 5.

Table 2: Optimal B-hyperpaths and paths for the large Wnt signaling pathway.

These types of questions will not only help improve the manual curation of signaling pathway databases, but will also provide insight into potential means of *pathway crosstalk* (where the stimulation of one pathway affects the downstream targets of another). To initiate this analysis, we removed self-loops from hypernodes that appeared in the head and the tail of 43 hyperedges.

For the "New Sources to Known Targets" problem, we connected a source hypernode s to 3,065 elements that did not appear in the large Wnt signaling pathway and had an empty backward star. We connected 84 hypernodes from the large Wnt signaling pathway that were located in the nucleus t. The modified signaling hypergraph contained 8,781 hypernodes and 10,876 hyperedges, which reduced to 7,341 hypernodes and 8,569 hyperedges after finding the hypernodes that were B-connected to s.

The optimal B-hyperpath consisted of six hyperedges (s_1 to t_1 in Figure 6). The nuclear complex containing the Androgen receptor (AR) and the hormone dihydrotestosterone (T-DHT) is present in the Wnt signaling pathway due to a reaction with a complex involving β-catenin [20] (this reaction does not appear in our solution). The optimal B-hyperpath included two upstream biological events: the formation of this complex in the cytosol followed by its translocation of the complex to the nucleus. These upstream events are included in the "Regulation of Androgen receptor activity" NCI-PID pathway.

For the "Known Sources to New Targets" problem, we connected s to 143 hypernodes in the large Wnt signaling pathway that had an empty backward star. We connected 939 hypernodes that did not appear in the large Wnt signaling pathway and were located in the nucleus to t. The modified signaling hypergraph contained 8,781 hypernodes and 8,809 hyperedges; this number reduced to 260 hypernodes and 268 hyperedges after finding the hypernodes that are B-connected to s. There were three optimal B-hyperpaths containing three hyperedges; all involve simple paths leading to post-translational modifications of JUN that are not in the Wnt signaling pathway. This result was not surprising, since Jun has many regulators and over 15 different post-translational forms. To find the "next" best B-hyperpath,

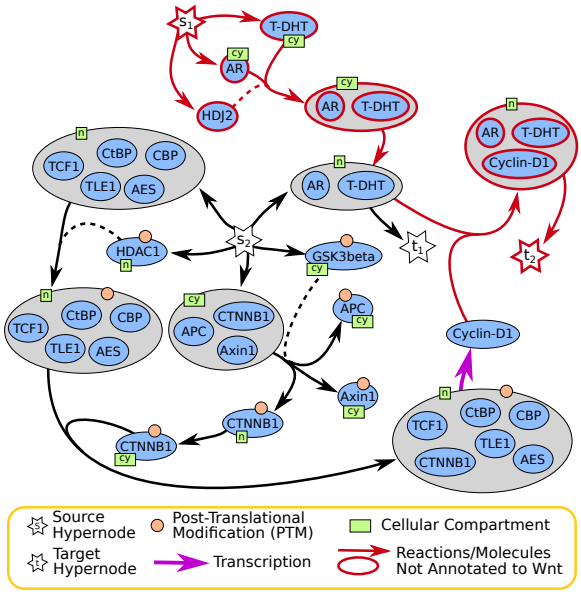

Figure 6: Two B-hyperpaths computed in the full NCI-PID signaling pathway. The B-hyperpath from s_1 to t_1 is the optimal result for "New Sources to Known Targets", and the B-hyperpath from s_2 to t_2 is the optimal result for "Known Sources to New Targets."

we removed the hyperedge connecting the Jun proteins to t. In this modified hypergraph, the optimal B-hyperpath contained 12 hyperedges (s_2 to t_2 in Figure 6). The transcription of Cyclin-D1 (and the events leading up to it) are members of the Wnt signaling pathway; the AR/T-HDT complex is in the pathway as well. Surprisingly, the formation of the AR/T-HDT/Cyclin-D1 complex is not in the Wnt signaling pathway. Cyclin-D1 is a co-repressor of AR [19], and the formation of the AR/T-DHT/Cyclin-D1 complex appears in NCI-PID's "Coregulation of Androgen receptor activity" pathway. Further, the complex formation is a spontaneous reaction.

On hindsight, the results appear to be unsurprising. One optimal B-hyperpath describes the formation of the AR/T-DHT complex and its transport to the nucleus. The other optimal B-hyperpath culminates in the spontaneous complexing of AR/T-DHT with Cyclin-D1. However, the NCI-PID curators selected to include these complexes and reactions in three different pathways. Manual discovery of these connections is likely to be very difficult. Signaling hypergraph theory offers a facile way to make such discoveries.

5.4 Performance Evaluation

The MILP was implemented in Python version 2.7.3, and in practice ran in a manner of seconds for all experimental scenarios. For the Wnt signaling pathways, the runtime of the signaling hypergraph representations ranged from 0.1s to 1.29s; comparable to that of graphs with complexes (0.11s to 0.49s) and graphs (0.79s to 1.49s) The runtime of the signaling hypergraph MILP for the full NCI-PID pathway took considerably longer in the "New Sources to Known Targets" scenario (36.57s) compared to the "Known Sources to New Targets" scenario (0.52s), reflecting the large difference in the relative size of the signaling hypergraphs.

6. CONCLUSIONS

The limitations of graph-based approaches for signaling pathways analysis have been recognized for years. A number of representations have been developed that involve directed hypergraphs and hypergraph-like notions. We have proposed a related representation called signaling hypergraphs that allow better characterization of reactions that involve multiple complexes and proteins. Signaling hypergraphs produce more informative hyperpaths than corresponding graph representations on NCI-PID curated pathways.

We have described an MILP to compute optimal acyclic B-hyperpaths in signaling hypergraphs. As we have noted earlier, characterizing signaling hypergraphs that handle all forms of regulation (including inhibition) and developing algorithms to compute cyclic B-hyperpaths are points of future work. Both of these aspects may require generalizing B-connectedness. Further, other notions of connectedness (including F-connection, which defines connectedness among hypernodes according to the forward star rather than the backward star) are worth considering for signaling pathway analysis [1]. We also note that finding B-hyperpaths that optimize other hyperpath measures, such as hyperpath traversal cost and hyperpath rank, admit polynomial-time solutions [1, 26] and may be useful in the context of signaling pathways. Logic models that contain information about the "state" of a protein or complex are a special case of directed hypergraphs [24]. Incorporating this type of information in signaling hypergraphs may provide a scalable alternative to dynamic models.

We initially chose NCI-PID to interrogate because it contains a balance of manually-curated reactions and annotated signaling pathways that are relatively well-connected. We note that NCI-PID is not longer actively maintained, and we have found minor inconsistencies and ambiguities upon closer inspection of the Wnt signaling pathway. We plan to convert other signaling pathway databases such as Reactome [5] and KEGG [16] to signaling hypergraphs and apply the MILP to these pathways.

We have reported optimal B-hyperpaths in Wnt signaling, both within the annotated pathway as well as in the context of the larger NCI-PID dataset. The corresponding shortest paths and Steiner trees found in graph representations miss crucial components of the underlying reactions. Further, some of the paths are misleading, as in the case with RanBP3 in the Figure 4. Through the development of new hypergraph-based algorithms, signaling hypergraphs have the potential to more accurately reflect the complexity of reactions in signaling pathway analysis.

7. ACKNOWLEDGMENTS

National Institute of General Medical Sciences of the National Institutes of Health grant R01-GM095955, National Science Foundation grant DBI-1062380, and Environmental Protection Agency grant EPA-RD-83499801 supported this work.

8. REFERENCES

[1] G. Ausiello, R. Giaccio, G. F. Italiano, and U. Nanni. Optimal traversal of directed hypergraphs. Technical report, 1992.

[2] M. Bailly-Bechet, C. Borgs, A. Braunstein, J. Chayes, A. Dagkessamanskaia, J. M. François, and R. Zecchina. Finding undetected protein associations in cell signaling by belief propagation. *Proceedings of the National Academy of Sciences*, 108(2):882–887, Jan. 2011.

[3] C. Berge. *Hypergraphs, Volume 45: Combinatorics of Finite Sets (North-Holland Mathematical Library)*. North Holland, 1 edition, Aug. 1989.

[4] R. Cambini, G. Gallo, and M. Scutellà. Flows on hypergraphs. *Mathematical Programming*, 78(2):195–217, 1997.

[5] D. Croft, G. O'Kelly, G. Wu, R. Haw, M. Gillespie, L. Matthews, M. Caudy, P. Garapati, G. Gopinath, B. Jassal, S. Jupe, I. Kalatskaya, S. Mahajan, B. May, N. Ndegwa, E. Schmidt, V. Shamovsky, C. Yung, E. Birney, H. Hermjakob, P. D'Eustachio, and L. Stein. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Research*, 39(Database issue):D691–D697, Jan. 2011.

[6] E. Demir, Ö. Babur, I. Rodchenkov, B. A. Aksoy, K. I. Fukuda, B. Gross, O. S. Sümer, G. D. Bader, and C. Sander. Using biological pathway data with Paxtools. *PLOS Computational Biology*, 9(9):e1003194, 2013.

[7] E. Demir *et al.* The BioPAX community standard for pathway data sharing. *Nature Biotechnology*, 28(9):935–942, 2010.

[8] U. Dogrusoz, A. Cetintas, E. Demir, and O. Babur. Algorithms for effective querying of compound graph-based pathway databases. *BMC Bioinformatics*, 10(376), 2009.

[9] K. Fukuda and T. Takagi. Knowledge representation of signal transduction pathways. *Bioinformatics*, 17(9):829–837, 2001.

[10] S. R. Gallagher and D. S. Goldberg. Clustering coefficients in protein interaction hypernetworks. In *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*, BCB'13, pages 552:552–552:560, New York, NY, USA, 2013. ACM.

[11] G. Gallo, G. Longo, S. Pallottino, and S. Nguyen. Directed hypergraphs and applications. *Discrete Applied Mathematics*, 42(2–3):177 – 201, 1993.

[12] I. Gat-Viks and R. Shamir. Refinement and expansion of signaling pathways: the osmotic response network in yeast. *Genome Research*, 17(3):358–67, 2007.

[13] L. S. Heath and A. A. Sioson. Semantics of multimodal network models. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6(2):271–280, 2009.

[14] J. Hendriksen, F. Fagotto, H. van der Velde, M. van Schie, J. Noordermeer, and M. Fornerod. RanBP3 enhances nuclear export of active (beta)-catenin independently of CRM1. *J. Cell Biol.*, 171(5):785–797, Dec 2005.

[15] Z. Hu, J. Mellor, J. Wu, M. Kanehisa, J. M. Stuart, and C. DeLisi. Towards zoomable multidimensional maps of the cell. *Nature Biotechnology*, 25(5):547–554, May 2007.

[16] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research*, 40(Database issue):D109–114, Jan 2012.

[17] G. Karlebach and R. Shamir. Modelling and analysis of gene regulatory networks. *Nature Reviews Molecular Cell Biology*, 9(10):770–780, September 2008.

[18] S. Klamt, U.-U. Haus, and F. Theis. Hypergraphs and cellular networks. *PLoS Computational Biology*, 5(5):e1000385, 2009.

[19] K. E. Knudsen, W. K. Cavenee, and K. C. Arden. D-type cyclins complex with the androgen receptor and inhibit its transcriptional transactivation ability. *Cancer Res.*, 59(10):2297–2301, May 1999.

[20] H. Li, J. H. Kim, S. S. Koh, and M. R. Stallcup. Synergistic effects of coactivators GRIP1 and beta-catenin on gene activation: cross-talk between androgen receptor and Wnt signaling pathways. *J. Biol. Chem.*, 279(6):4212–4220, Feb 2004.

[21] D. Machado, R. S. Costa, M. Rocha, E. C. Ferreira, B. Tidor, and I. Rocha. Modeling formalisms in Systems Biology. *AMB Express*, 1(1):45–14, Dec. 2011.

[22] H. Matsuno, Y. Tanaka, A. H., A. Doi, M. Matsui, and S. Miyano. Biopathways representation and simulation on hybrid functional petri net. *In Silico Biol.*, 3(3):389–404, 2003.

[23] A. Ritz, A. N. Tegge, H. Kim, C. L. Poirel, and T. Murali. Signaling hypergraphs. *Trends in Biotechnology*, 32(7):356–362, 2014.

[24] R. Samaga and S. Klamt. Modeling approaches for qualitative and semi-quantitative analysis of cellular signaling networks. *Cell Commun. Signal*, 11(1):43, 2013.

[25] C. F. Schaefer, K. Anthony, S. Krupa, J. Buchoff, M. Day, T. Hannay, and K. H. Buetow. PID: the Pathway Interaction Database. *Nucleic Acids Research*, 37(Database issue):D674–D679, 2009.

[26] M. Thakur and R. Tripathi. Linear connectivity problems in directed hypergraphs. *Theoretical Computer Science*, 410(27):2592–2618, 2009.

[27] C. J. Vaske, S. C. Benz, J. Z. Sanborn, D. Earl, C. Szeto, J. Zhu, D. Haussler, and J. M. Stuart. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*, 26(12):i237–i245, 2010.