Pathway Logic: A Modeling Tool for Experimental Biologists

Carolyn Talcott and Merrill Knapp
SRI International
firstname.lastname@sri.com

July 24, 2009

Abstract

In this paper we present the Pathway Logic System from the point of view of how a bench biologist might use it to understand experimental results, and develop theories and testable hypotheses about the system they are studying.

1 Introduction

Systems biology (SB) can be described as a global approach to biological modeling that is based on "omics" data sets obtained using automated high-throughput molecular biology and protein biochemistry techniques (e.g., [5]). SB thus represents a fundamental change from classic experimental biology—which is essentially reductionistic and hypothesis-driven—to a more integrated and quantitative analysis of large-scale biological processes (systems). Pathway Logic takes a *Symbolic Systems Biology* a approach, using logical and qualitative representation and reasoning. The aim is to provide tools for experimental biologists that complement and augment their ability to reason and hypothesize about laboratory findings. In particular, Pathway Logic tools are designed to enable users to explore how experimental conclusions concerning pathways of interest compare or integrate with those of other researchers studying the same or related pathways—in effect, helping to reconcile "bottom-up" conclusions with existing or "top-down" models. In this paper we describe these tools conceptually—in terms of pathway maps and builders; and obtaining and reasoning about experimental evidence.

How might an experimental biologist use Pathway Logic? What is the interest in intracellular signaling? Consider a cancer biologist. In cancer cells, proliferation is out of control. The signals that start cell division are growth factors such as the Epidermal Growth Factor (Egf). What breaks is intracellular. If you could fix the break you could cure cancer by non-brute force methods (killing lots of cells and possibly the host). To discover what is broken, biologists use experimental data to build models of small modules of the cellular signaling system. The problem then becomes how to put these models together to gain a bigger picture,

to combine relations obtained from experimental observations in order to infer mechanisms, and to test logical aspects of these hypotheses before going back to the laboratory.

Plan. In section 2 we discuss related work in building pathway maps and collecting evidence. In section 3 we give a brief introduction to Pathway Logic framework and tools and in 4 we discuss curation of the knowledge bases. In section 5 we give examples showing how an experimental biologist might use Pathway Logic, using our model of Egf stimulation as a case study. In section 6 we conclude and discuss future directions.

2 Pathway Maps: Related Work

2.1 Using Pathway Maps

Pathway maps are diagrammatic representations of biological processes such as signaling and metabolic reactions, which are commonly depicted as discrete transitions or events connected in time or space (e.g., see [12]). Two popular examples of pathway maps are offered by Biocarta (www.biocarta.com) and Cell Signaling Technology (www.cellsignal.com/pathways). These maps–which show canonical signaling or metabolic pathways–are visual aids that display only as much information as can be grasped by human inspection, and have a semantic consistency obtained by linking individual components to a protein database or a commercial reagent such as an antibody. A notable exception can be found in CellDesigner (www.celldesigner.org), which can produce maps that are both large and semantically unambiguous (e.g., see [9]). However, if literature references are given for a component of these maps, they are associated with the entire diagram rather than with specific events involving the component. GenMAPP (www.genmapp.org) illustrates an additional feature of publicly available pathway maps: relative expression levels of specific transcripts can be superimposed as colors on the maps to integrate data obtained from transcriptome experiments (e.g., differential gene expression studies) with canonical pathways (usually protein-level components). In this case, names and species of pathway components are standardized because they are linked to gene microarray annotation files from various data sources. The NCI Nature Protein Interaction Database (`pid.nci.nih.gov`) is a curated collection of information about known biomolecular interactions and key cellular processes assembled into signaling pathways. Each reaction has associated evidence in the form of one or more PubMed Ids and Reaction components are linked to UniProt pages. The database is searchable, however there is no way to combine or compare pathways or to select subpathways other than those that are predefined. Finally, there is an important and growing application of pathway maps for SB research: pathways are used as the basis for in silico simulations of biological processes using differential rate equation or statistical approaches (e.g., see [3, 7] and references therein). A collection of such pathways can be found at the Biomodels Database (www.ebi.ac.uk/biomodels-main). To enable an in silico simulation of a process, discrete transitions between component states in a pathway are assigned reaction equations containing measured, estimated, or probabilistic rate constants, and various parameters are inserted to determine the effect of computing from a

given starting state. This approach is hindered by a lack of experimental data for rate constants and other key parameters.

In contrast to the examples discussed above, Pathway Logic generates pathway maps that represent the combination of a review article(s) and an information storage device. To prepare a review, an author generally collects information about a system from the experimental literature, attempts to assemble a coherent view or perspective concerning the system, and presents a theory or model that is supported by evidence from the literature. The author of a Pathway Logic model uses essentially the same approach, except that the model takes the form of a set of transitions (reactions) in a formal representation system. Thus the model is available for visualization and analysis in various ways, and for export for use by different computational platforms–thus, the model can be shared by a broad community of users.

We propose that pathway maps meeting practical needs of experimental biologists

- describe entire systems rather than isolated pathways
- have properties of a process diagram [6] or a hierarchical graph [12]
- provide experimental evidence for each reaction in a system
- are easily revised
- can be viewed in different ways, combined or broken into smaller parts
- can be queried to find regions of interest
- can be used by different computational platforms.

Pathway maps are essentially collections of interpretations of experimental results (inferences) that are presented as a hypothetical model of how a molecular system is assembled. If we can use a computer to draw a model, we have a useful tool. If, however, we can use a computer to also access the experimental results used to make the inferences, we have a much more useful tool. This is one of the goals of the Pathway Logic project.

2.2 Tools for Building Pathway Maps

Many tools are available for building (assembling and visualizing) biological pathways and networks; each tool provides one or more solutions for specific needs–reference [12] provides extensive summary tables of the features of diverse tools publicly available as of 2007. In general, these tools have many useful features, but they also suffer from a significant disadvantage–the *only* way for a user to input new reactions is to draw them manually. Specifically, a user must employ a drawing tool to create a component or pathway of interest; which is then translated into a markup language such as systems biology markup language (SBML; e.g., [4]) or BioPax (www.biopax.org) to convert it for use by different platforms.

Overall, current building tools offer an intuitive way to generate pathway maps, and they are particularly useful for focusing on small pathways or the details of large networks. However these building tools typically do not provide functionality to compare pathways or combine

two or more pathways into model of a larger system. For such operations, automated layout is important. When using drawing tools that limit a user to graphical notation symbols provided by the tool, there can be serious limitations; e.g., new types of transitions, cellular locations, or other modifications to a pathway may require representing new concepts. This can force a user to build maps that reflect the notation of a tool rather than those that accurately represent new experimental findings. On the other hand, if there are no constraints on the use of notation, the interpretation of a hand-drawn graph (pathway map) can be ambiguous or its meaning can be jeopardized.

The Pathway Logic Assistant is capable of automatically assembling large numbers of reactions into one or more networks, while having the ability to extract subnets and paths that satisfy user-specified conditions. As in a review article, in which statements are supported by specific references to the scientific literature—the experimental evidence—the elements of a Pathway Logic network are also linked to the sources from which they were derived. The underlying logical formalism of Pathway Logic provides flexibility in choice of representation, while at the same time ensuring semantic consistency.

2.3 Obtaining Evidence for Building Pathway Maps

The standard method in which the biological and biomedical community presents experimental evidence online is to provide a PubMed Id linking to the Medline abstract for a relevant publication. This method, however, has the practical problem that it requires retrieving and reading the entire publication to attempt to find experiments that support a conclusion or inference.

Extracting experimental evidence from primary sources by curation is a major problem that has been addressed in different ways. The use of automated text mining through Natural Language Processing (e.g., see [18]) is promising–however, the practical implementation of automated curation currently faces serious obstacles. For example, because information in the scientific literature is highly specialized, semantically unpredictable, and often not textual (graphical images), automated curation is likely to miss or distort key forms of experimental evidence when reading text. Manual curation, in contrast, is labor intensive and prone to curator error or bias; however, it is commonly used by experimentalists when they start to investigate a new problem—it is practical and practiced.

The most common tool used for storing and retrieving biological information is conventional database software. Database approaches have been used extensively: e.g., by the National Center for Biotechnology Information (NCBI, `www.ncbi.nlm.nih.gov`) and the Universal Protein Resource (UniProt, `www.uniprot.org`) for storing sequence information; and by the Biomolecular Interaction Network Database (BIND, `bond.unleashedinformatics.com`), IntAct (`www.ebi.ac.uk/intact/site/index.jsf`), and the Human Protein Reference Database (HPRD, www.hprd.org) for storing protein-protein interaction information. To generate pathway maps, however, retrieving information about the sequences, interactions, and biochemical modifications of signaling molecules is necessary but not sufficient to explain how a signal is propagated. That is, crucial information about signal transduction is derived from experiments that measure the change in the interactions and modifications of signaling molecules in response to a perturbation of a system—information that a human curator

can obtain from experimental reports, but is often missing in available databases.

Ingenuity Systems (www.ingenuity.com) is a leader in the field of manual curation; Ingenuity has been extracting biological findings or conclusions from literature in the public domain since 1998. Its approach is to use a "highly descriptive, controlled vocabulary" and a unique ontology to "enforce semantic and linguistic consistency across all the reported literature findings". Although Ingenuity's original intent was to represent published scientific findings without re-interpretation or bias by curators (e.g., see [1]), in fact, the controlled vocabulary constrained curators to collect the conclusions of the authors rather than their experimental results.

To support reliable manual curation of the experimental literature, we are developing a system, called *datums*, to collect, store, and retrieve curated information so that it can be understood and shared by a community of experimental biologists, and used to is developing models of cellular processes.

3 About Pathway Logic

We introduce Pathway Logic, the underlying formal representation system (Maude), and the associated tool suite, the Pathway Logic Assistant and its Petri net representation.

3.1 Pathway Logic Concepts

As already mentioned, Pathway Logic (PL) [13, 14, 15, 16] is a symbolic systems biology approach to the modeling and analysis of molecular and cellular processes based on rewriting logic. In PL, biological molecules, their states, locations, and their roles in molecular or cellular processes can be modeled at very different levels of abstraction. For example, a complex signaling protein can be modeled either according to an overall activity state, its post-translational modifications, or as a collection of protein functional domains and their internal or external interactions. Similarly, biological processes can be represented at different levels of granularity using rewrite rules. Each rule represents a hypothesis about a step (at the chosen level of granularity) in a biological process such as metabolism or intra-/inter- cellular signaling. A rule may represent a family of reactions using variables to stand for families of molecular components. Rules express both change (reactants and products) and dependencies on biological context (also called modifiers), for example, an enzyme needed to enable the reaction or a scaffold needed to hold proteins in position to interact productively.

PL has two notions of "model". The first notion of model, called a rule knowledge base, is a general model of the basic mechanisms underlying biological processes. A rules knowledge base (RKB) is represented as a rewrite theory that consists of a collection of rules together with supporting data type descriptions. The data type descriptions (formally, algebraic signatures) specify the concepts and vocabulary to be used, and how complex entities are formed from basic entities. Each biological molecule that is declared in a PL RKB has associated metadata (formal annotations) linking it to standard database entries, e.g., HUGO and UniProt/Swiss-Prot for proteins and KEGG or PubChem for chemicals, along with other information such as category and synonyms. This information is important to place the knowledge in a broader context and to be able to integrate it with other knowledge sources. In addition to reactants,

products and required context, each rule has associated experimental evidence used to justify the rule. Such a collection of rules forms a network with rules connected by shared reactant, product or context elements.

The second notion of model is that of a biological system, how it changes over time, how it affects and is affected by its environment. A PL system model consists of a system state together with a RKB. A cellular system state is given by describing the components of a cell (such as proteins, their states and locations), and any stimuli in its environment. We call such a system state a 'dish', thinking of experiments carried out in a Petri dish. Such models are called executable models, and can be understood as specifying possible ways a system can evolve by application of rules in the RKB, as described below. Associated to a dish and a RKB is the subset of rules that are "reachable" starting in the state described by the dish. This is known as a "dishnet", and can be used in place of the full RKB to analyze the model. A pathway is a set of rules whose application corresponds to the steps in one possible execution of the system, e.g., one way in which a signal can propagate. In Pathway Logic, pathways are not predefined. Instead they are assembled by applying the rules starting from a given state and searching for a state meeting user-specified conditions.

Pathway Logic provides a pathway query language to describe simple conditions. A query consists of sets of goals and avoids. A goal describes a desired situation a state and location of a cellular component while an avoid describes a situation to be avoided. A query can also specify rules that should not be used. One query used in section 5 specifies that Erks be activated. A second query adds to this goal the requirement to avoid use of Sos1. A pathway satisfies a query if in the end state all the goal situations are achieved and none of the avoid situations have occurred as the rules are applied. Given a query, the relevant subnet is the set of rules that contains all (minimal) pathways that satisfy the query. Pathways satisfying queries are found using a formal analysis technique called "model checking". A model checker is an algorithm that explores the possible executions (sequences of rule applications) of a system and checks whether a given property is satisfied by the execution. If some execution violates the property, the set of rules applied is returned as a counter example. To find a pathway satisfying a query, we ask the model checker to show that the query cannot be satisfied. If a counterexample is found, it is a pathway satisfying the query. Relevant subnets are computed using computational analysis algorithms called forward and backward collection [14]. A subnet consisting of connections to a given set of molecular components can be generated by graph exploration techniques.

3.2 Rewriting Logic and Maude

Pathway Logic models of biological processes are developed using the Maude system [2], a formal language and tool set based on rewriting logic. The Rewriting logic formalism [8] is based on two simple ideas: states of a system are represented as elements of an algebraic data type; and the behavior of a system is given by local transitions between states described by rewrite rules. A rewrite rule has the form $t \Rightarrow t'$ **if** c where t and t' are patterns (terms possibly containing place holder variables) and c is a condition (a Boolean term). Such a rule applies to a system in state S if t can be matched to a part of S by supplying the right values for the

place holders, and if the condition c holds when supplied with those values. In this case the rule can be applied by replacing the part of S matching t by t' using the matching values for the place holders in t'. The process of application of rewrite rules generates computations (also thought of as deductions). In the case of biological processes these computations correspond to pathways.

Maude provides a high-performance rewriting engine featuring matching modulo associativity, commutativity, and identity axioms. Matching is used to determine if a rule applies to a system state and the result of application. The associativity, commutativity, and identity axioms are used to describe states that are mixtures. In this case, the order in which the elements are presented does not matter. This allows rules for reactions in such mixtures to be described very compactly and naturally. Maude also provides search and model-checking capabilities. Thus, given a specification S of a system, one can execute S by rewriting to find one possible behavior, use search to see if a state meeting a given condition can be reached; or model-check S to see if a temporal property is satisfied, and if not to see a computation that is a counter example.

In PL logical inference and analysis techniques provided by Maude and other tools are used for simulation to study possible ways a system could evolve, to assemble pathways as answers to queries, and to reason about dynamic assembly of complexes, cascading transmission of signals, feedback-loops, cross-talk between subsystems, and larger pathways. Logical and computational reflection are used to transform and further analyze models.

3.3 The Pathway Logic Assistant

The Pathway Logic Assistant (PLA) [17] provides an interactive visual representation of PL models that allow a user to query a model and to perform in silico experiments to study the effects of perturbations on these networks via a graphical interfaces. In PLA, models are represented as graphs with nodes for rules and cellular components, and edges connecting reactant components to rules and rules to product components. Using PLA a biologist can

- ask for a list of dishes available for study, and modify or create dishes
- display and navigate the network of signaling reactions for a specified model
- display information about components, with links to public databases
- display active links to evidence from which rules have been derived
- formulate and submit queries to find and display pathways and subnets
- compute and display possible alternative pathways resulting from the knockout of components or rules
- compare two pathways
- find knockouts—proteins whose omission prevents reaching a specified state
- incrementally explore network connections to given rules or components

Many of these features are illustrated in section 5.

Formally, the PLA graphs are Petri Nets [11, 10], a model of concurrent processes that correspond to special forms of rewrite theory, and for which a number of efficient analysis tools are available. Petri Nets were invented to model execution of concurrent processes and thus are nicely suited to modeling signals propagating through a cell. Figure 1 shows a Pathway Logic rule represented as a Petri net transition along with the Maude representation from which it was derived.

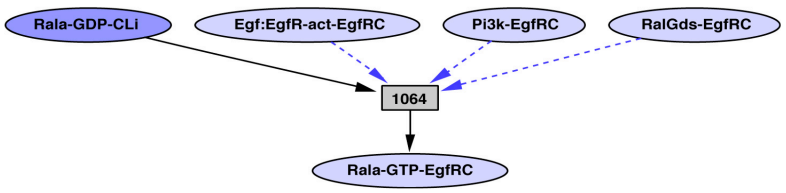


Figure 2a. A Pathway Logic rule represented graphically as a Petri net transition.

```
rl[1064.Rala.irt.Egf]:
  {EgfRC | egfrc ([EgfR - act] : Egf) ralagef:RalaGEF              }
  {CLi   | cli   [Rala - GDP]                                      }
  =>
  {EgfRC | egfrc ([EgfR - act] : Egf) ralagef:RalaGEF [Rala - GTP] }
  {CLi   | cli                                                     } .
```

Figure 2b. The same rule in Maude representation.

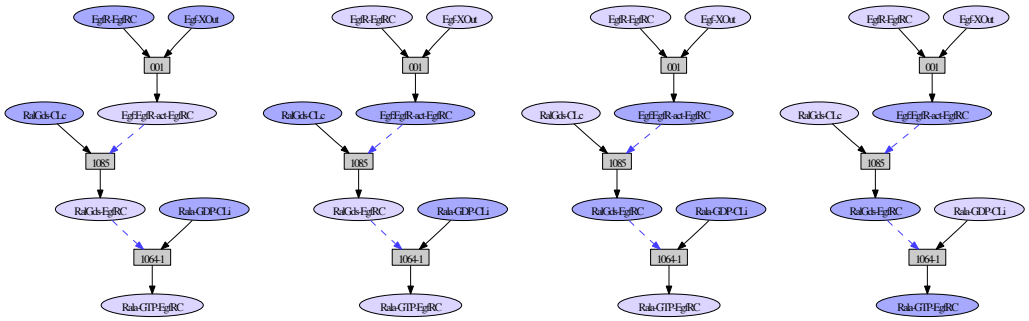
Figure 1: Rule 1064: activation of Rala

The rule says that if Rala is bound to GDP and is located in the inside of the cell membrane (-CLi), and Egf is bound to EgfR in the EgfR complex (-EgfRC), and if the Rala GEF (guanine nucleotide exchange factor) RalGds is also located in the EgfR Complex, then Rala will translocate to the EgfR complex and GDP will be exchanged for GTP. In a Pathway Logic network, places are called occurrences and transitions correspond to rules. Each occurrence represents a molecular component (typically a protein, small molecule or complex), its state and/or modifications, and its location. Occurrences are displayed as ovals and rules as rectangles. Ovals are labeled with a printed representation of the occurrence name, and rectangles are labeled with the rule identifier. The reactants and products of a rule are connected by solid black arrows lead to/from the rule. A dashed blue arrow is used when an occurrence is required for a transition to take place but the state of the occurrence itself remains unchanged (e.g., an unmodified enzyme). Dark-colored ovals represent occurrences present in the initial/current state of the model, and lighter-colored ovals represent occurrences that might become present as the system evolves. The Maude representation of this rule collects all components in a given location together in a list surrounded by brackets ($\{\ \}$) and tagged by the location name. The reactants and products are separated by an arrow ($\Rightarrow$) and the modifiers must be listed before and after the arrow to make explicit that they are required and unchanged. A modified component is represented by the component name and its modifications listed between []s and separated by a dash. Thus, Rala bound to GDP is represented by [Rala - GDP].

A set of Petri net rules corresponding to the rules of a PL knowledge base is called a transistion knowledge base (TKB). The analog of a PL dish is a PL Petri net state, which specifies which occurrences are present, that is, it specifies the state and location of each molecular

component. Given a state, a Petri net rule is enabled if all of its occurrences connected by incoming arrows (reactants and modifiers) are present in the state. When an enabled rule fires, the reactant occurrences are removed from the state and the product occurrences are added. The modifier occurrences are left unchanged.

Corresponding to a PL model, a Petri net model consists of a set of rules (a TKB) and an initial state. To execute a Petri net model one puts tokens on the ovals corresponding to occurrences present in the initial state, and moves tokens as rules become enabled and fired. Figure 2 shows the execution of a Petri net model of the pathway that leads to the activation of Rala, i.e. the firing of (an instance of) rule 1064.



(a) initial state (b) step 1 (c) step 2 (d) step 3

Figure 2: Execution of the Rala activation pathway

There are three rules in the pathway. Darker ovals represent occurrences that are present (marked with a token). Figure 2(a) shows the initial state with Egf on the outside (Egf-XOut), EgfR the single element of EgfRC, the EgfR complex (EgfR-EgfRC), RalGds in the cytoplasm (RalGds-CLc), and Rala-GDP at the inside of the cell membrane (Rala-GDP-CLi) marked as initially present. The only rule enabled is rule 001. Figure 2(b) shows the result of firing rule 001, removing tokens from Egf-XOut and EgfR-EgfRc and adding a token to Egf:EgfR-act-EgfRC (EgfR complexed with Egf and activated). Now rule 1085 is enabled and Figure 2(c) shows the result of firing rule 1085. This enables rule 1064-1 and Figure 2(d) shows the final state.

PLA is available for download for Mac OS X and Linux. There is a PLA Online demo version that runs a Java client on the users machine, accessed through WebStart. Downloads, the demo, sample models, tutorial material, papers and presentations are available from the Pathway Logic web site, `http://pl.csl.sri.com`.

4 Pathway Logic Knowledge Base Curation

Pathway Logic includes two knowledge bases for rules and evidence. The Pathway Logic rules knowledge base (RKB) contains over a thousand rules. The RKB contains a comprehensive model of signaling in response to Epidermal Growth Factore (Egf) stimulation developed to study cancer-related signaling networks (the topic of section 5). It also includes preliminary models of response to other ligands including Tnf (Tumor necrosis factor precursor), IL1

(Interleukin-1 alpha and beta), Igf1 (Insulin-like growth factor 1A), Lps (Lipopolysaccharide), and Ins (Insulin). It also contains a collection of rules called Common Rules summarizing experiments concerning interactions independent of specific stimuli.

Each rule in the RKB is linked to supporting evidence. Evidence items are represented in a formal system and stored in the evidence knowledge base (EKB). The EKB is designed to record units of evidence that we call "datums" to emphasize their individuality and distinguish them from "data" which we use to describe a collection of results. It is important that each datum capture objective information, rather than conclusions of the experimenter or curator. The knowledge base and its infrastructure is designed to be generally useful for experimental biologists, thus it is important for datums to be expressed using readily understood concepts with generally agreed-upon meaning, e.g., assays, detection methods and cells. Furthermore, each datum should contain a managable chunk of information, sufficient to unambiguously describe an experimental finding.

To guide our choice of attributes used in the formal representation of datums, we imagined that all the papers in Medline could be indexed for experimental content. We then asked: What questions would a cell biologist studying signal transduction want to ask? What information would need to be represented to answer those questions? In our preliminary formalization each datum includes a subject, the assay, a result, the experimental environment, and the source of the datum. Change datums also describe the treatment. Experiments often are repeated with variations, such as mutations, knock-outs, or knock-ins of components of interest. These are recorded in an extras element. A datum may also have a comments element to record additional unstructured information. These elements are described in some more detail in the following.

Evidence for Rule 1064

As an example we show the evidence for rule 1064 (see Figure 1). Recall that this rule says the Egf:EgfR complex must be formed and contain RalGds in order to exchange the GDP bound to Rala for GTP and recruit Rala into the complex. How was this rule inferred? First, EKB was searched for all datums in which Rala is the subject and Egf is the treatment. This search pulled up seven datums. Five of the datums are shown here in one of the possible printed formats for datums (notation is explained below, the missing datums are similar to those shown and omitted).

```
DID#05386: Rala[Ab] GTP[BD-PD] is increased irt Egf (tnr)
   cells: Cos7 in BMLS
   inhibited by: xCdGap [addition]
   source: 15034142-Fig-5a

DID#05387: Rala[Ab] GTP[BD-PD] is increased irt Egf (tnr)
   cells: Cos7 in BMLS
   inhibited by: Wortmannnin [chem]
   inhibited by: LY294002 [chem]
   source: 15034142-Fig-5c

DID#05395: Rala[Ab] GTP[BD-PD] is increased irt Egf (tnr)
```

```
   cells: Cos7 in BMLS
   inhibited by: xRac1-T17N ``DN-mutant'' [addition]
   source: 15034142-Fig-5a

DID#12876: xRala[xAb]IP GTP/GDP[32Pi-TLC] is increased irt Egf
   cells: Cos1-xRalGds in BMS
   times: 0 1+ 2++ 3++ 4+ 5 min
   reqs: xRalGds [omission]
   inhibited by: xRalGds-C203S ``membrane-binding-mutant'' [substitution]
   comment: cells were pretreated with Vanadate 30 min before Egf treatment
   source: 9416833-Fig-2

DID#15191: Rala[Ab] GTP[BD-PD] is increased irt Egf (10 min)
   cells: Hek293 in BMLS
   inhibited by: xRalGds-(1-297) ``C-term-mutant'' [addition]
   source: 11889038-Fig-7d
```

How do we read these datums? Consider the datum (DID#12876). The first line contains the subject, expressed Rala immunoprecipitated by an antibody to the expressed tag (xRala[xAb]IP); the assay, amount of GTP bound, determined by metabolic labeling of cells with radioactive inorganic phosphate followed by separation of GTP from GDP by thin-layer chromatography (GTP[32Pi-TLC]); the direction of change, (is increased); and the treatment (irt Egf). The "cells:" line tells us that the assay was performed on Cos1 cells transfected with RalGds (Cos1-xRalGds) growing in basal medium containing serum (in BMS). The "times:" line says that measurements were taken at 1, 2, 3, 4, and 5 minutes, with the most increase at 2 and 3 minutes. The "reqs:" line says that the observed increase did not occur if the cells did not contain expressed RalGds (omission). The "inhibited by:" line says that the observed increase was significantly inhibited if the expressed RalGds was substituted with RalGds with a C203S mutation. The final line tells us that this datum was taken from Figure 2 of the paper with pubmed id 9416833. The "comment:" line points out a variation from standard protocols that may explain discrepancies between the results of this experiment compared to similar experiments.

What do we learn from this list of datums? Three papers contained experiments that showed that the amount of Rala bound to GTP was increased in response to an Egf treatment. In 2/3 of the papers or 4/7 of the experiments, the Egf treatment time was less than 10 minutes, demonstrating that the effect was not due to new proteins being synthesized in response to Egf.

It is a general principal that the transformation of Rala-GDP to Rala-GTP requires a GEF for Rala. The Egf response is inhibited by the expression of a C-terminal mutant (DID#15191) or a C203S mutant (DID#12876) of RalGds indicating that RalGds might be required. The response is inhibited by a membrane-binding mutant of RalGds (DID#12876) indicating that RalGds might require translocation to the cell membrane to help activating Rala. The inhibition by over-expression of mutants of RalGds (DID#15191) (DID#12876) demonstrate that it is RalGds that is required and not just any Rala GEF.

The response is inhibited by the chemicals Wortmannin and LY294002, indicating a possible requirement for Pi3k (PI 3-kinase) (DID#05387). The response is also inhibited by the ex-

pression of CdGap (a Gap for Cdc42) indicating that Cdc42 might be required (DID#05386). The response is inhibited by the expression of a dominant-negative Rac1 indicating that Rac1 might be required (DID#05395). The evidence for these additional requirements is deemed weak, for example the two chemicals may inhibit more than Pi3k, and thus they are omitted from rule 1064. A separate rule set including hypothetical rules are with these requirements are provided for the user to include or not as they choose.

5 Using Pathway Logic

In this section we explain some of the ways an experimental biologist might use the PL knowledge bases and PLA in their research. As mentioned in section 4 our knowledges bases currently contain curated datums and rules relevant to mammalian intracellular signal transduction. We will focus on the PL model of response to Epidermal growth factor (Egf) stimulation. This is an important model for the study of cancer and many other phenomena as Epidermal growth factor receptor (EgfR) signaling regulates growth, survival, proliferation, and differentiation in mammalian cells.

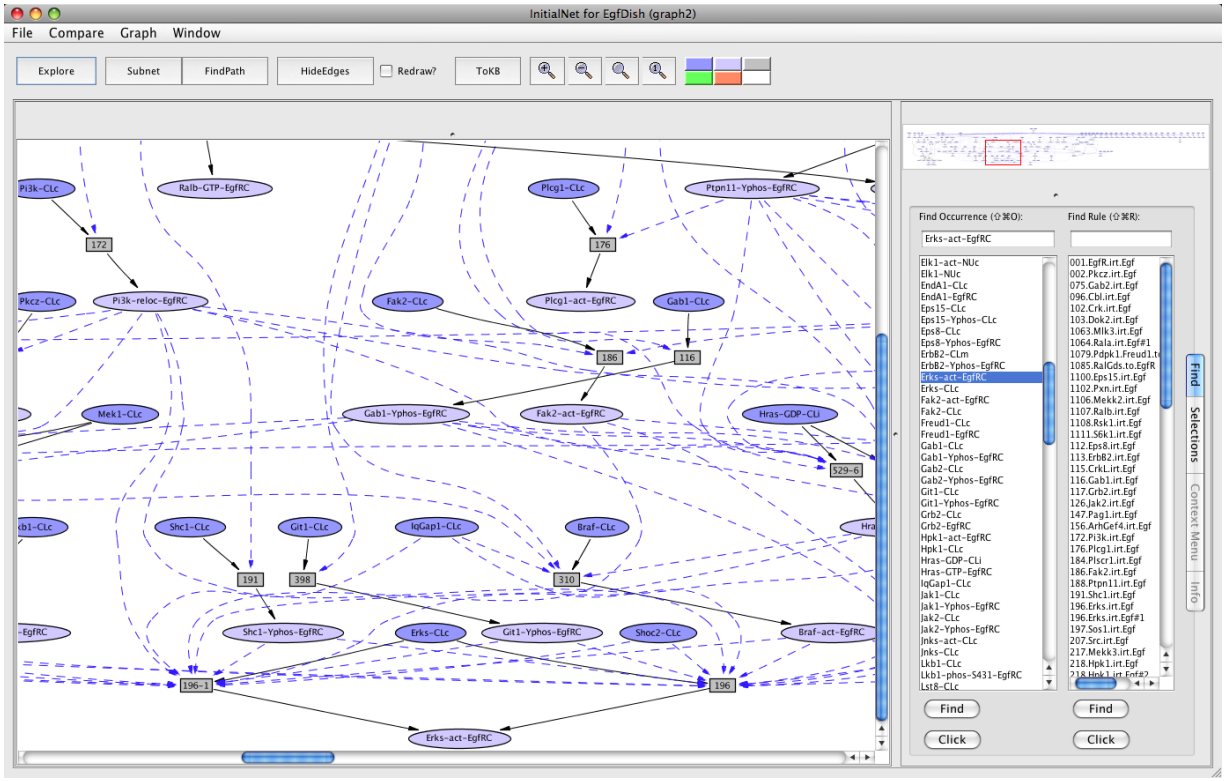


Figure 3: PL model of Egf stimulation viewed in PLA

In the PL RKB we have a model of these events obtained by curating datums from the literature reporting experiments studying cells treated with Egf and using these datums together with

general biological knowledge about e.g., activity of certain proteins, to formulate a set of rules characterizing individual events. These rules are extracted from the full RKB by specifying a basic cell together with the Egf ligand in the cells environment (supernatant) as the initial state. Figure 3 shows a screen shot of this model, represented as a Petri net and viewed in PLA. The upper right shows a thumbnail of the full model. An enlarged version of the portion in the viewport (redbox) is displayed on the left. The lower right is an information panel that allows the user to search for specific components or rules or specify query elements.

This version of the Egf network looks different from others because it includes all rules inferred from changes observed in response to Egf binding to the EgfR, not just a single pathway. The formulation of a rule includes all the biological context that has been shown to be required for the main event to happen. A link to the evidence supporting each rule is provided in the viewer.

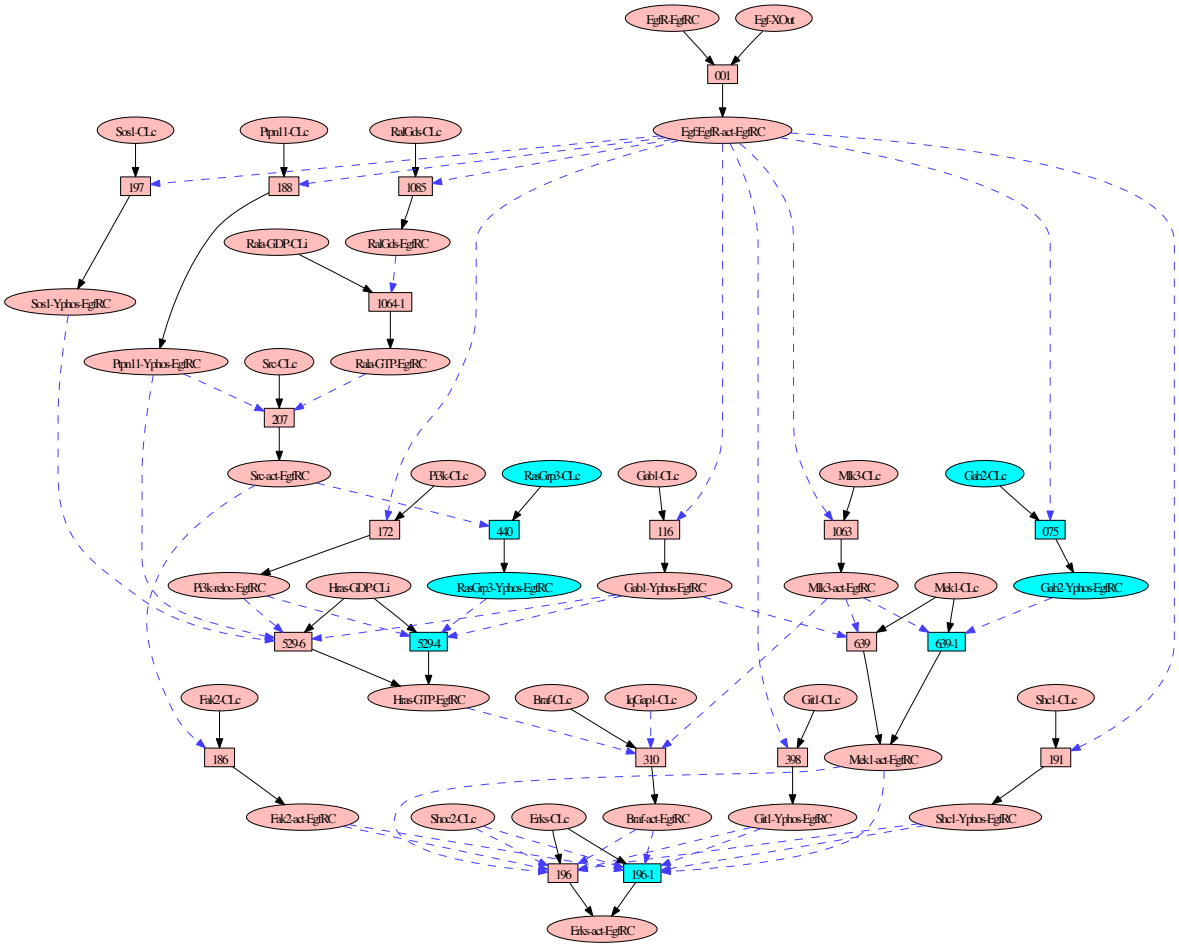


Figure 4: A pathway activatint Erks embedded in the relevant subnet

Activation of the MAPK (Mitogen Activated Protein Kinase) Erk is a key step along the way to activating transcription. This pathway is often represented as a linear sequence of events:

$$\text{Egf} \rightarrow \text{EgfR} \rightarrow \text{Grb2} \rightarrow \text{Sos1} \rightarrow \text{Ras} \rightarrow \text{Raf1} \rightarrow \text{Mek} \rightarrow \text{Erk}$$

Here is a biologist style explanation of what this picture or sequence represents.

> In this canonical pathway, Egf binds to the Egf receptor (EgfR) and stimulates its protein tyrosine kinase activity to cause auto-phosphorylation, thus activating EgfR. Next, the adaptor protein Grb2 and the guanine nucleotide exchange factor (GEF) Sos1 are recruited to the membrane and bind to the activated EgfR. The Sos1-containing EgfR complex activates a Ras family GTPase, and the activated Ras protein activates Raf1, a member of the RAF serine/threonine protein kinase family. Raf1 then activates the protein kinase Mek, which then activates Erk.

Is the pathway really a linear chain? What do these arrows mean? What specific proteins do the names refer to? For example there several memebers of the Erk family. Egf stimulation evidence concerns Erk1 and Erk2, but the experiments do not distiguish the behavior of these two members of the Erk family. Thus we introduce a constant Erks to stand for either Erk1 or Erk2 (or both). We can find out how Erks can be activated in our Egf model by making the active form (Erks-act-EgfRc) a goal and asking for the relevant subnet. This gives us the subnetwork that contains all (minimal) pathways. Figure 4 shows this subnet. We can also ask the subnet for one pathway using the findPath tool. This is the part of the subnet colored pink.

The teal colored components highlight possible alternatives. For example the picture suggests that rule 529-4 might be an alternative to 529-6. The difference in these two rules is the choice of GEF used to converts Hras-GDP to Hras-GTP. As in the classic Erk activation pathway, rule 529-6 uses active Sos1 in the EgfRC complex (Sos1-Yphos-EgfRC), while rule 529-4 uses RasGrp3-Yphos-EgfRC. Biologists know that Sos1 is a GEF for Hras but there is no evidence that Sos1 is required for activation of Erk. Suppose we do an in silico KO experiment, knocking out Sos1 by specifying that it should be avoided. Now findPath produces a pathway that uses RasGrp3. Thus our model accounts for at least two ways to activate Hras within an Egf to Erk pathway.

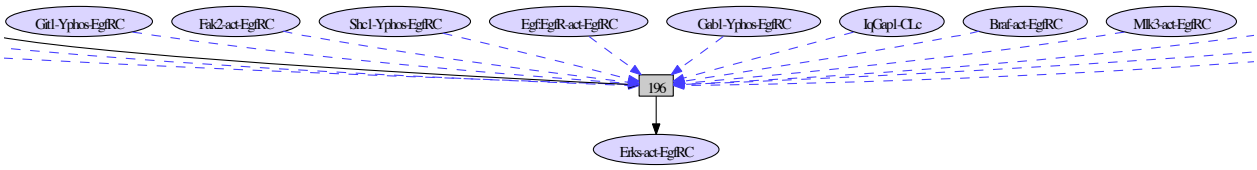


Figure 5: A portion of Rule 196–activation of Erk

The final rule, 196, shows many things are required, not just active Mek: Mek1-act-EgfRC, Fak2-act-EgfRC, Braf-act-EgfRC, Mlk3-act-EgfRC, Ptpn11-Yphos-EgfRC, Shc1-Yphos-EgfRC, Gab1-Yphos-EgfRC or Gab2-Yphos-EgfRC, Hras-GTP-EgfRC, IqGap1-CLc, Git1-CLc, Shoc2-CLc. Where did all those bizarre requirements come from? We can ask PLA for the evidence for rule 196. This evidence page reports on 194 experiments in which Erk1, Erk2, or both were either phosphorylated on their TEY site (T202/Y204 for Erk1, T185/Y187 for Erk2) or

activated in an in vitro kinase assay. The peak response occurred around 5-10 min. Thirty-nine different cell lines were represented.

Some of the required proteins such as Mek1, Hras, and Shc1 are expected. Surprisingly, although Raf1 is activated in response to Egf, there is plentiful evidence that Raf1 is NOT required for Erk phosphorylation or activation in response to Egf. There is also evidence that Mek1/2 activation in response to Egf does not require Raf1.

Another non-standard requirement is that Mlk3 is required for activation of Braf , Mek1 and Erks in response to Egf. Rule 1063 shows Mlk3 being activated in response to formation of the active Egf-EgfR complex. But is this a direct activation? From the datums we learn that although Mlk3 is a kinase, its requirement by Braf and Erks does not require its kinase activity . Furthermore, overexpression of Mlk3 actually inhibits the activation of Erks in response to Egf . This is indicative of a scaffold or adaptor protein rather than a active member of a phosphorylation cascade. We query the datums knowledge base for evidence about activity of Mlk3 and find (among others)

```
DID#29544 xMlk3[xAb]IP IVKA(auto)[32P-ATP] is detectable
 cells: HELA in BMS
 inhibited by: xMlk3(K144R)"kinase-inactive-mutant" [substitution]
 inhibited by: xMlk3(T277A)"phos-site-mutant" [substitution]
 partially inhibited by: xMlk3(S281A) [substitution]
 source: 11053428(D)
```

which suggests that Mlk3 activity requires phosphorylation on threonine-277 and serine-281. Can EgfR initiate this activation? We may know that EgfR is a tyrosine kinase, or we can follow the EgfR links to the PRO or Uniprot pages to learn the specificity of the kinase activity of EgfR.

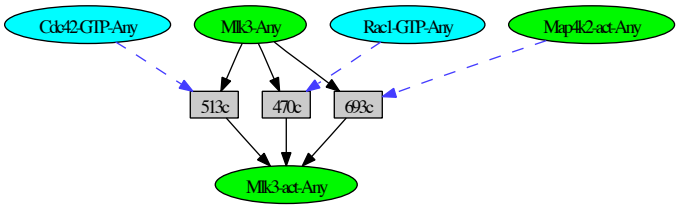


Figure 6: Common Rule activation of Mlk3

This makes us suspicious that EgfR does not activate Mlk3 directly. To investigate that hypothesis, we use the network explorer capability of PLA to find known upstream connections of Mlk3-act (with Any location). Figure 6 shows the result. There are three rules upstream of Mlk3-act in the knowledge base connecting to Map4k2-act, Cdc42-GTP, or Rac1-GTP. An oval colored green has only downstream connections, while and oval colored teal (cyan) has both upstream and downstream connections. Map4k2-act being green means that we don't have information about how it is activated. The teal color of Cdc42-GTP and Rac1-GTP suggest that that there may be a known path to activation, and looking in the Egf model, we find that they are indeed activated in response to Egf (see Figure 7).

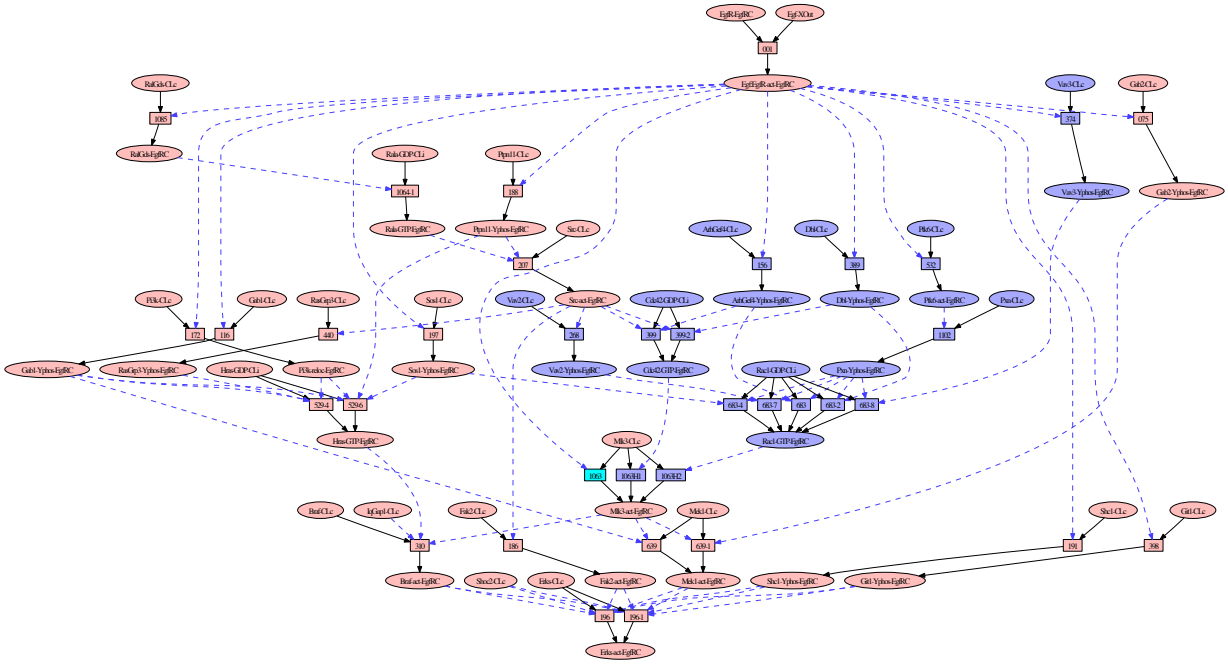


Figure 7: Erk subnets with/without hypothetical rules for Mlk3

We hypothesize that a possible missing link in the model is an additional requirement of Cdc42-GTP or Rac1-GTP for activation of Mlk3. We add corresponding hypothetical rules (1063H1 and 1063H2) to the knowledge base. The resulting subnet for activation of Erks and its comparison to the original is shown in Figure 7. The bluish nodes show the new elements needed to activate Mlk3, and hence to activate Erks. The remaining part of the network (pinkish nodes) is basically unchanged. The teal colored rule shows the original one step activation of Mlk3. Now the next step is to either find additional evidence in the literature or to check our hypotheses in the lab.

6 Conclusion and Future Directions

We have presented the Pathway Logic modeling and analysis system from the point of view of its use by experimental biologists. The system has much broader applicability including use by computational biologists and in collaborative efforts.

The network presented here is not intended to be cast in stone. It is an illustration of

- (1) how PLA can be used to help interpret experimental results. The ease with which pathways can be redrawn, rules modified.
- (2) how storing experimental results in the form of Datums can aid reasoning and discussion.

- (3) to provide a head start to anyone interested in persuing work in this direction (as in not reinventing the wheel versus take a guru's word for it)
- (4) A new kind of review article. This has been tried by using update articles but that assumes that all early conclusions still hold.

There are may future directions and opportunities for improvement and expansion of scope. Better was to explore and view networks are needed. For example, restricting the type of links to follow or the endpoints of interest when doing automated exploration would reduce the clutter in the generated subnetworks. A useful additional exploration mechanism would be finding pathways between network components. Again there needs to be ways to specify the pathways of interest, as there are usually many if there are any. Dual to local exploration is the ability collapse subnetwork modules in to single nodes to give a better understanding of the big picutre.

For expanding the scope and content we are looking at importing pathways from other curated resources (for example the NCI-Nature PID). Another promising direction is developing immune system models, that include inter-cellular signaling and integrate this with intracellular signaling. Finally, a rather different direction is to link Pathway Logic to drug target data bases, for example to automate analysis of potential effects of drugs beyond the intended effect.

References

[1] S. E. Calvano, W. Xiao, D. R. Richards, R. M. Felciano, H. V. Baker, R. J. Cho, R. O. Chen, B. H. Brownstein, J. P. Cobb, S. K. Tschoeke, C. Miller-Graziano, L. L. Moldawer, M. N. Mindrinos, R. W. Davis, R. G. Tompkins, and S. F. Lowry. A network-based analysis of systemic inflammation in humans. *Nature*, 437:1032–1037, 2005.

[2] Manuel Clavel, Francisco Durán, Steven Eker, Patrick Lincoln, Narciso Mart´-Oliet, José Meseguer, and Carolyn Talcott. *All About Maude: A High-Performance Logical Framework*. Springer, 2007.

[3] K. A. Janes and D. A. Lauffenburger. A biological approach to computational models of proteomic networks. *Curr Opin Chem Biol*, 10:73–80, 2006.

[4] D. B. Kell and P. Mendes. The markup is the model: reasoning about systems biology models in the semantic web era. *J Theor Biol*, 252:538–543, 2008.

[5] P. Kersey and R. Apweiler. Linking publication, gene and protein data. *Nat Cell Biol*, 8:1183–1189, 2006.

[6] H. Kitano, A. Funahashi, Y. Matsuoka, and K. Oda. Using process diagrams for the graphical representation of biological networks. *Nat Biotechnol*, 23:961–966, 2005.

[7] M. J. Lazzara and D. A. Lauffenburger. Quantitative modeling perspectives on the erbb system of cell regulatory processes. *Exp Cell Res*, 2008.

[8] J. Meseguer. Conditional Rewriting Logic as a unified model of concurrency. *Theoretical Computer Science*, 96(1):73–155, 1992.

[9] K. Oda and H. Kitano. A comprehensive map of the toll-like receptor signaling network. *Mol Syst Biol*, 2, 2006.

[10] J. L. Peterson. *Petri Nets: Properties, analysis, and applications*. Prentice-Hall, 1981.

[11] C. A. Petri. Introduction to general net theory. In Brauer, W., editor, *Net Theory and Applications, Proceedings of the Advanced Course on General Net Theory of Processes and Systems, Hamburg, 1979*, volume 84 of *LNCS*, pages 1–19, Berlin, Heidelberg, New York, 1980. Springer-Verlag.

[12] M. Suderman and M. Hallett. Tools for visually exploring biological networks. *Bioinformatics*, 23:2651–2659, 2007.

[13] C. Talcott, S. Eker, M. Knapp, P. Lincoln, and K. Laderoute. Pathway logic modeling of protein functional domains in signal transduction. In *Proceedings of the Pacific Symposium on Biocomputing*, January 2004.

[14] Carolyn Talcott. Formal executable models of cell signaling primitives. In Tiziana Margaria, Anna Philippou, and Bernhard Steffen, editors, *2nd International Symposium On Leveraging Applications of Formal Methods, Verification and Validation ISOLA06*, pages 303–307, 2006.

[15] Carolyn Talcott. Symbolic modeling of signal transduction in pathway logic. In L. F. Perrone, F. P. Wieland, J. Liu, B. G. Lawson, D. M. Nicol, and R. M. Fujimoto, editors, *2006 Winter Simulation Conference*, pages 1656–1665, 2006.

[16] Carolyn Talcott. Pathway logic. In *Formal Methods for Computational Systems Biology*, volume 5016 of *LNCS*, pages 21–53. Springer, 2008. 8th International School on Formal Methods for the Design of Computer, Communication, and Software Systems.

[17] Carolyn Talcott and David L. Dill. Multiple representations of biological processes. *Transactions on Computational Systems Biology*, 2006.

[18] G. A. Viswanathan, J. Seto, S. Patil, G. Nudelman, and S. C. Sealfon. Getting started in biological pathway construction and analysis. *PLoS Comput Biol*, 4(e16), 2008.