USING PATHWAY LOGIC TO INTEGRATE SIGNAL TRANSDUCTION AND GENE EXPRESSION DATA

Linda Briesemeister, Merrill Knapp, Keith Laderoute,
Andy Poggio, and Carolyn Talcott*

SRI International
Menlo Park, CA, 94025, USA
City, State ZIP/Zone, Country
Email: {firstname.lastname}@sri.com

Joe Gray, Laura Heiser, and Paul Spellman

Lawrence Berkeley National Laboratory
Berkeley, CA, USA
Email: {JWGray,LMHeiser,PTSpellman}@lbl.gov

This poster presents first steps towards using symbolic models of signaling pathways in combination with statistical methods to analyze gene expression data. As a case study, data from 51 breast cancer cell lines is used.

1. INTRODUCTION

Pathway Logic [3, 4, 6, 7] is an approach to the modeling and analysis of molecular and cellular processes based on rewriting logic [5]. A Pathway Logic knowledge base includes data types representing cellular components such as proteins, small molecules, complexes, compartments/locations protein state, and post-translational modifications. Rewrite rules describe the behavior of proteins and other components depending on modification state and biological context. Each rule represents a step in a biological process such as metabolism or intra/intercellular signaling. A collection of such facts forms a formal knowledge base. Logical inference and analysis techniques 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.

The poster illustrates the use of the Pathway Logic knowledge base in combination with statistical methods to analyze gene expression data obtained from a collection of breast cancer cell lines. From the gene expression data for a cell line, a network of potentially reachable signaling reactions (rules) is extracted from the knowledge base. These networks of rules are clustered to find small signaling modules whose elements appear together or are absent together in the networks for the different cell lines. Genes whose presence/absence differs across the cell lines have also been mapped onto the PL Egf signaling network.

2. PATHWAY LOGIC MODELS

Pathway Logic models are written using the rewriting logic language Maude (`http://maude.cs.uiuc.edu`). They are structured in four layers: (1) sorts and operations, (2) components, (3) rules, and (4) queries. The *sorts and operations* layer declares the main sorts and subsort relations, the logical analog to ontology. The sorts of entities include Chemical, Protein, Complex, and Location (cellular compartments), and Cell. These are all subsorts of the sort, Soup, that represents 'liquid' mixtures, as multisets (unordered collections) of entities. The sort Modification is used to represent post-translational protein modifications. Modifications are applied using the operator `[ - ]`. For example the term `[PrlR - act]` represents the Prolactin receptor in an activated state. A cell state is represented by a term of the form

```
[cellType | locs]
```

*Corresponding author.

where `cellType` specifies the type of cell, for example Macrophage, and `locs` represents the contents of a cell organized by cellular location. Each location is represented by a term of the form `{ locName | components }` where `locName` identifies the location, for example `CLm` for cell membrane, and `components` stands for the mixture of proteins and other compounds in that location.

The *components* layer specifies particular entities (proteins, genes, chemicals) and introduces additional sorts for grouping proteins in families. The *rules* layer contains rewrite rules specifying individual steps of a process. As an example we show one of the rules relevant to the gene expression analysis. The rule has a label, `766.PrlR.by.Prl`, a before pattern and an after pattern, separated by a =>. In English, the rule says that for any cell containing PrlR in its membrane, if Prl is present in the supernatant containing the cell (the before pattern), then PrlR will become activated and Prl will be bound to its receptor on the outside surface of the cell (the after pattern).

```
rl[766.PrlR.by.Prl]:
   Prl
  [any:CellType | ct
    {CLo | clo}{CLm | clm PrlR}]
  =>
  [any:CellType | ct
    {CLo | clo [Prl - bound]}
    {CLm | clm [PrlR - act]}] .
  ---------------------------------
  *** 11566606(R) PrlR is a homodimer
```

'Any cell' is represented by the variable (placeholder) `any:CellType` in the cell term `[any:CellType | ... ]`. Formally, using this variable means the rule applies to any cell type. Two of the cell's locations are represented explicitly: terms of the form `{CLo | ... }` represent the outside of the cell membrane; and terms of the form `{CLm | ... }` represent the the cell membrane. The symbol `ct` is a variable that stands for the remaining locations whose contents are not important for this reaction. Similarly the symbols `clo` and `clm` are variables standing for other components of their respective locations. The premis 'containing PrlR in its membrane' is represented by `PrlR` in the location named `CLm`, and 'Prl is present in the supernatant containing the cell' is represented by `Prl` adjacent to the cell term. Activated PrlR is represented by `[PrlR - act]` and 'Prl bound to its receptor on the outside surface of the cell' is represented by `[Prl - bound]` in the location named `CLo`. The line beginning `***` is the annotation for this rule.

The *queries* layer specifies initial states. Initial states are in silico Petri dishes containing a cell, with its components, and ligands of interest in the supernatant.

The Pathway Logic Assistant (PLA) is a tool that provides an interactive visual representation of a PL model, and allows user to browse and query the network of reactions associated to an initial state. PLA, sample models, tutorial material, papers and presentations are available from the Pathway Logic web site, `http://pl.csl.sri.com/`

3. MAPPING GENE EXPRESSION DATA INTO PL

To illustrate the use of the PL knowledge base and the PLA tool, in connection with other tools, to analyze gene expression data we consider data represents expression levels for a set of 51 breast cancer tumor cell lines in their exponentially growing state. For each cell line, an initial state was determined based on the gene expression data, and using PLA a network of reachable signaling reactions was generated for each initial state.

We were interested in whether the networks for the 51 cell lines could be grouped by their network properties. To address this issue, we performed an unsupervised hierarchical clustering on the network components that varied across the 51 networks. This resulted in twenty rule clusters. Figure 1 show three of these rule clusters. Some of these clusters were clearly a consequence of the model bias, some were of dubious interest due to inclusion of components in the initial state that may not be relevant. However, certain clusters are of direct relevance to the breast cancer project in that they suggest differences in malignant progression among the cell lines. For example, five of the cell lines were predicted to preferentially use signaling from the prolactin receptor (PrlR, green cluster rules), which binds the pleiotropic cytokine and mammary epithelial growth factor prolactin (Prl) [1] . Enhanced activity of the PrlR may be a significant risk factor for human breast cancer [8], highlighting the oncogenic potential of this system in certain breast cancer cells. Another signaling system predicted to be used preferentially by a subgroup of the breast cancer cell lines involves the c-Met receptor (Met, orange cluster rules), which binds hepatocyte growth factor (Hgf; also called Scatter Factor) [2]. Met is a potent source of signals both for the pro-

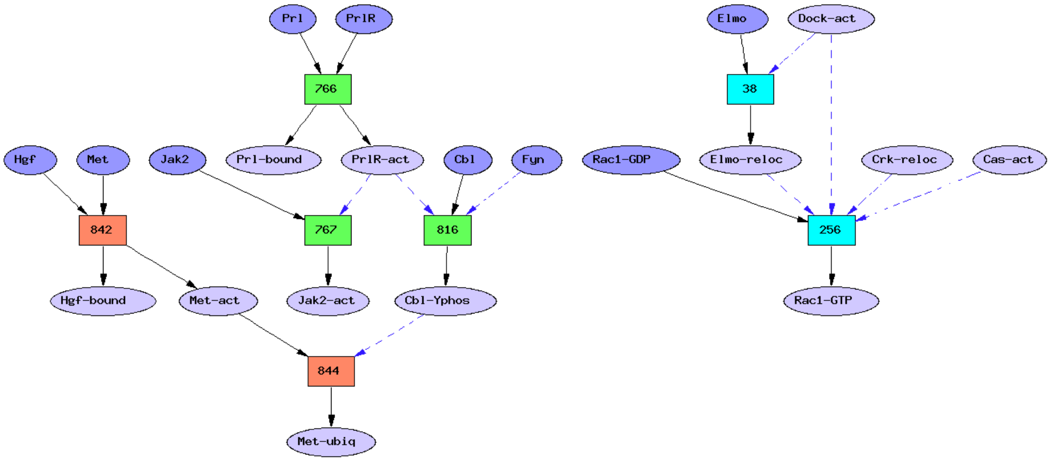

Fig. 1. Three rule clusters. Rectangles represent reaction rules (clusters identified by the color of the reactangle). Ovals represent components (proteins in different states). Solid arrows lead from reactants to rules and rules to products. When a rule fires, reactants are replaced by products. Dashed arrows lead from context components (that must be present but are not changed, such as enzymes) to rules.

liferation and chemotaxis of various human cancer cells, including breast cancer cells [2].

4. CONCLUSIONS

This poster illustrates initial efforts to use the PL curated knowledge base and the PLA tool to analyze gene expression data derived from breast cancer tumor cells. It will be important to experimentally confirm these and other predictions of this rule clustering analysis of Pathway Logic signaling network models of this set of breast cancer cell lines.

Next steps include developing additional algorithms to infer new pathways connected to genes discovered to be distinctive by statistical and information theoretic analyses. Future plans also include using the PL knowledge base to analyze protein data for these same cell lines.

Acknowledgments

This work is part of the Lawrence Berkeley National Laboratory Integrative Cancer Biology Program project (`http://icbp.lbl.gov`). It was supported in part by a Public Health Service Grant CA112970 from the National Cancer Institute.

References

1. N. Ben-Jonathan, K. Liby, M. McFarland, and M. Zinger. Prolactin as an autocrine/paracrine growth factor in human cancer. *Trends Endocrinol Metab*, 13:245–250, 2002.
2. J. G. Christensen, J. Burrows, and R. Salgia. c-met as a target for human cancer and characterization of inhibitors for therapeutic intervention. *Cancer Letters*, 225:1–26, 2005.
3. Steven Eker, Merrill Knapp, Keith Laderoute, Patrick Lincoln, José Meseguer, and Kemal Sonmez. Pathway Logic: Symbolic analysis of biological signaling. In *Proceedings of the Pacific Symposium on Biocomputing*, pages 400–412, January 2002.
4. Steven Eker, Merrill Knapp, Keith Laderoute, Patrick Lincoln, and Carolyn Talcott. Pathway Logic: Executable models of biological networks. In *Fourth International Workshop on Rewriting Logic and Its Applications (WRLA'2002), Pisa, Italy, September 19 — 21, 2002*, volume 71 of *Electronic Notes in Theoretical Computer Science*. Elsevier, 2002. `http://www.elsevier.nl/locate/entcs/volume71.html`.
5. J. Meseguer. Conditional Rewriting Logic as a unified model of concurrency. *Theoretical Computer Science*, 96(1):73–155, 1992.
6. 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.
7. Carolyn Talcott and David L. Dill. The pathway logic assistant. In Gordin Plotkin, editor, *Third International Workshop on Computational Methods in Systems Biology*, pages 228–239, 2005.
8. S. S. Tworoger and S. E. Hankinson. Prolactin and breast cancer risk. *Cancer Letters*, 2006.