Creating Metabolic Network Models using Text Mining and Expert Knowledge

J.A. Dickerson[1], D. Berleant[1], Z. Cox[1], W. Qi [1], and E. Wurtele[2]
Iowa State University, Ames, IA, 50011

Abstract:

This paper describes the initial development of a tool that helps find and visualize metabolic networks. The tool is written in Java™ and consists of two parts. The first part is a text-mining tool that pulls out potential metabolic relationships from the existing database PubMed. These relationships are then reviewed by a domain expert and added to an existing network model. The result is visualized using a graph display module. The basic metabolic or regulatory flow is illustrated using a binary network. An example from the regulatory network for the plant hormone gibberellin shows how this tool operates.

I. Introduction

Despite the large amounts of research in genomics, more effort is needed for development of methodologies to identify and analyze complex biological networks. RNA profiling analysis is yielding vast amounts of data on gene expression. New techniques such as proteomics will further add to this glut of information.

The goal of this project is to develop a publicly available software suite called the Gene Expression Toolkit. This toolkit will aid in the analysis and comparison of large microarray, proteomics, and metabolomics data sets. The user can select parameters for comparison such as species, experimental conditions, and developmental stage. Two of the key tools in the Gene Expression Toolkit are a text-mining tool, called PathBinder, which permits the mining of online literature and a network modeling tool called FCModeler. The PathBinder citations will be available to the researcher and smoothly transferable for use in annotating displays in other parts of the package and as links in building models. The FCModeler tool for gene regulatory and metabolic networks is intended to easily capture the intuitions of biologists and help test hypotheses along with providing a modeling framework for assessing the large amounts of data captured by microarrays and other high-throughput experiments. This tool uses fuzzy methods for modeling networks and interprets the results using fuzzy cognitive maps (Dickerson and Kosko 1994; Kosko 1986; Kosko 1986).

II. Structure of Concepts and Links

Metabolic networks form the basis for the net accumulation of biomolecules in living organisms. Regulatory networks modulate the action of these metabolic networks, leading to physiological and morphological changes. The modeling tool will integrate our understanding of the interactions within and between these regulatory and metabolic networks. The nodes represent specific biochemicals such as proteins, RNA, and small molecules, or stimuli, such as light, heat, or nutrients. Three types of links are specified as shown in Figure 1. In a conversion link (black arrow), a node (typically a chemical(s)) is converted into another node (chemical(s)), and used up in the process. In a regulatory link (green and red arrows), the node activates or deactivates another node, and is not used up in the process. A catalytic link (blue arrows) represents an enzyme that enables a chemical conversion and does not get used up in the process.

Other key features include concentrations of the molecules (nodes), strengths of the links, and subcellular compartmentation. These data can be added as they are identified experimentally. Currently the biologist user can include or ignore a variety of parameters, such as subcellular compartmentation and link strength. Furthermore, because the node and link data is entered on simple Microsoft Excel™ spreadsheets, individual biologists can easily sort, share, and post data on the web.

Future versions will distinguish between regulation that results in changes in concentrations of the regulated molecule, and regulation that involves a reversible activation or deactivation.

III. PathBinder: Document Processing Tool for Finding Metabolic Pathways

PathBinder identifies information about the pathways that mediate biological processes from the scientific

---

[1] Electrical and Computer Engineering Department

[2] Botany Department

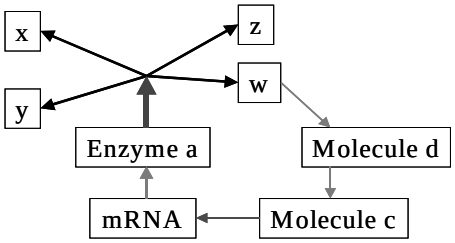


Figure 1: Links in the metabolic and regulatory network model. Black arrows indicate conversion links. Blue arrows represent catalytic links, green and red arrows are positive and negative. regulatory links, respectively.

literature. Sentences are useful units of information for this purpose, and are relatively straightforward units into which to segment a document. This tool searches through documents in Medline and Agricola for sentences containing terms that indicate relevance to signal transduction or metabolic pathways. Microarray data can be used to hypothesize causal relationships between genes, and PathBinder will then mine Medline and Agricola for information about these putative pathways, extracting passages most likely to be relevant to a particular pathway and storing this desired information in highly concentrated form. The information is presented in a user-friendly format that supports efficiently investigating the pathways.

A number of researchers have addressed extraction of protein interactions from MEDLINE. Blaschke et al. (1999) extracts by first identifying phrases conforming to the template: …protein …verbclass…protein…, where verbclass is one of 14 sets of pathway relevant verbs (such as "bind") and inflections. Ng and Wong (1999) describe a system, of which BioNLP is one component which extracts sentences containing pathway relevant verbs and applies templates to identify path relevant relationships among proteins. Rindflesch et al. (1999) apply non-trivial NLP to extract assertions about binding relationships in particular. Thomas et al. (2000) distinguish between verbs that are relatively more and less reliable in indicating protein interactions in their extraction work. The PathBinder work differs from these due to a combination of system design decisions. PathBinder avoids syntactic analysis of text in favor of word experts for pathway relevant verbs. Word experts are sets of rules for interpreting words (Berleant 1995). PathBinder also is oriented toward assisting humans in constructing pathways rather than fully automatic construction, thus avoiding information retrieval precision limitations. We are also investigating the relative performances of several algorithms for identifying relevant sentences, including verb-free algorithms that rely instead on protein term co-occurrences.

How PathBinder Works

Step 1: user input. Keyboard input of biomolecule names in pathways of interest by the user.

Step 2: synonym extraction. A user-editable synonym file combined with a more advanced module that will automatically access the HUGO (`www.gene.ucl.ac.uk.publicfiles/nomen/nomenclature.txt`) and OMIM (`www.ncbi.nlm.nih.gov/htbinpost/Omim/`) nomenclature databases, and extract synonyms.

Step 3: document retrieval. The PubMed and Agricola document repositories are accessed and queried using terms input in Step 1. The output of this step is a list of URLs with high relevance probabilities.

Step 4: sentence extraction. Each URL is downloaded and scanned for pathway-relevant sentences that satisfy the query. These sentences constitute pathway-relevant information "nuggets."

Repetition of steps 2 through 4, using different biomolecule names extracted from qualifying sentences. These new biomolecule names are candidates for inclusion in the pathways of interest. *Step 5: sentence index.* Process the messy collection of qualifying sentences into a far more user-friendly form, a multi-level index, with the number of levels dependent on the sentence extraction criteria. This index conforms to a pattern (Figure **2**), displayed by a Web browser, and the sentences in it are clickable. When a sentence is clicked, the document from which it came appears in the Web browser.

Step 6: integration with the rest of the software and the microarray data sets. The index can be used to create a graphical representation in which verbs are represented by lines, interconnecting the biomolecule names and forming a web-like relationship diagram of the extracted information.

PathBinder is useful as both a standalone tool and an integrated subsystem of the complete system. The multilevel indexes transform naturally into inputs for the network modeling tools. The networks that PathBinder helps identify will form valuable input to the clustering, display, and analysis software modules.

Example of a sample PathBinder Query:

The query is to find sentences containing (either gibberellin, gibberellins, or GA) AND (either SPY, SPY-4, SPY-5, or SPY-7). Three relevant results were found and incorporated into the metabolic and regulatory visualization. A single sentence example is show below.

Protein A
    Protein B
        Associates/Associated/etc.
            Sentence1
            Sentence2
            . . .
        Binds/Binding/Bind/etc.
            SentenceM
            SentenceN
            . . .
        Regulates/Regulating/etc.
        . . .
    Protein C
        Associates/Associated/etc.
            SentenceY
            . . .
        Binds/Binding/Bind/etc.
        . . .
Protein B
    Protein D
        Associates/Associated/etc.
            Sentence1000
. . . .

Figure 2. Format of index produced by the PathBinder software. A, B, C, etc., are placeholders for the "official" name of a protein followed by a list of its synonyms, and the SentenceN's stand for actual extracted sentences.

Sentence: "The results of these experiments show that spy-7 and gar2-1 affect the GA dose-response relationship for a wide range of GA responses and suggest that all GA-regulated processes are controlled through a negatively acting GA-signaling pathway."

Source Information: UI - 99214450, Peng J, Richards DE, Moritz T, Cano-Delgado A, Harberd NP, Plant Physiol 1999 Apr;119(4):1199-208.

IV. Visualizing the Network

The next step is to visualize the known and unknown biological information using a graph visualization program called *Graphviz* developed at AT&T research labs (http://www.research.att.com/sw/tools/graphviz/) to do the initial graph layout. The front end of the FCModeler tool is a Java ™ interface that reads and displays data from an Excel™ spreadsheet of links and nodes. This system handles such non-standard graphing techniques as links that modify other links as in a catalytic reaction. Figure 3a shows a small part of a graph for the Arabadopsis metabolic and regulatory network. Figure 3b shows the original graph plus the three new links discovered in the PathBinder search.

Eventually, the expression of the strength of a connection relative to another connection will be added to the graph. Connection strength can also reflect the user's confidence in the link between concepts. The system will check for conflicts between different network models by looking for paths that cancel each other out will be added to the software. When hypothesized edges are added, the software will check for direct and indirect causal conflicts as well as redundant information between pairs of concepts or nodes. When conflicts are discovered, the source of the conflict will be reported to the scientist doing the modeling.

V. FCModeler: Fuzzy Cognitive Map Modeling Tool for Regulatory Networks

The FCModeler tool models regulatory networks so that important relationships and hypotheses can be mined from the data. Some types of models that have been studied for representing gene regulatory networks are Boolean networks (Liang, 1998; Akutsu, 1999), linear weighting networks (Weaver, 1999), differential equations (Akutsu 2000), and Petri nets (Matsuno et al 2000). Circuit simulations and differential equations require detailed information that is not yet known about the regulatory mechanisms between genes. Boolean networks analyze binary state transition matrices to look for patterns in gene expression. Each part of the network is either on or off depending on whether a signal is above or below a pre-determined threshold. Linear weighting networks have the advantage of simplicity since they use simple weight matrices to additively combine the contributions of different regulatory elements. Petri nets can handle a wide variety of information however their complexity does not scale up well to systems that have both continuous and discrete inputs (Alla, 1998; Reisig, 1998).

Fuzzy cognitive maps (FCMs) have the potential to answer many of the concerns that arise from the existing models. Fuzzy logic allows a concept or gene expression to occur to a degree – it does not have to be either on or off (Kosko 1986). FCMs have been successfully applied to systems that have uncertain and incomplete models that cannot be expressed compactly or conveniently in equations. Some examples are modeling human psychology (Hagiwara 1992), modeling slurry rheology (Banini and Bearman 1998), and on-line fault diagnosis at power plants (Lee et al., 1996). All of these problems have some common features. The first is the lack of quantitative information on how different variables interact. The second is that the direction of causality is at least partly known and can be articulated by a domain expert. The third is that they link concepts from different domains together using arrows of causality. These features are shared by the problem of modeling the signal transduction and gene regulatory networks.

We will use a series of +/- links that model known signal transduction pathways and hypothesized pathways. A third link type will suggest a relationship between concepts with no implied causality. These links will be constructed by mining the literature using PathBinder and from the expert knowledge of biologists. Given the partial signal transduction network so constructed, we will augment the system with advanced tools that:

- Locate and visualize closely coupled subgraphs or signal transduction networks.
- Develop simulation tools for modeling intervention in the network (e.g. what happens when a node is shut off) and search for critical paths and control points in the network.
- Capture information about how edges between graph nodes change when different regulatory factors are present

Fuzzy cognitive maps are fuzzy digraphs that model causal flow between concepts or in this case genes, proteins, and transcription factors (Kosko 1986; Kosko 1986). The concepts are linked by edges that show the degree to which the concepts depend on each other. FCMs can be binary state systems with causality directions that are +1, a positive causal connection, -1, a negative connection, or zero, no causal connection. Simple binary limit cycles show "hidden patterns" of actions. The fuzzy structure allows the gene expression to be expressed in the continuous range [0, 1]. The input is the sum of the product of the fuzzy edge values. The system nonlinearly transforms the weighted input to each node using a threshold function or other nonlinear activation.

The edges between nodes can also be time dependent functions that create a complex dynamical system. Neural learning laws and expert heuristics encode limit cycles and causal patterns. One learning method is differential Hebbian learning in which the edge matrix updates when a causal change occurs at the input (Dickerson and Kosko 1994).

VI. Conclusions

The integration of FCModeler with PathBinder will allow biologists to gather and combine information from the literature, their expert knowledge, and the public databases of mRNA results.

References:

Akutsu, T., S. Miyano, et al. (1999). "Identification of Genetic Networks from a Small Number of Gene Expression Patterns Under the Boolean Network Model," *Pacific Symposium on Biocomputing* 4, Hawaii.

Akutsu, T., Miyano, S., Kuhara, S. (2000). "Algorithms for Inferring Qualitative Models of Biological Networks," *Pacific Symposium on Biocomputing* 5, Hawaii.

Alla, H. and R. David (1998). "Continuous and Hybrid Petri Nets." *Journal of Circuits, Systems, and Computers* **8**(1): 159-188.

Banini, G. A. and R. A. Bearman (1998). "Application of fuzzy cognitive maps to factors affecting slurry rheology," *International Journal of Mineral Processing* **52**(4): 233-244.

Berleant, D., "Engineering Word Experts for Word Disambiguation," *Natural Language Engineering* 1 (4) (Dec. 1995) 339-362.

Blaschke, C.; Andrade, M. A.; Ouzounis, C.; and Valencia, A. 1999. "Automatic extraction of biological information from scientific text: Protein-protein interactions," In *Proceedings of the AAAI Conference on Intelligent Systems in Molecular Biology*, 60--67.

Dickerson, J. A. and B. Kosko (1994). "Virtual Worlds as Fuzzy Cognitive Maps." *Presence* **3**(2, Spring): 173-189.

Hagiwara, M. (1992). "Extended Fuzzy Cognitive Maps," *92 IEEE Int Conf Fuzzy Syst FUZZ-IEEE*, San Diego, IEEE.

Ideker, T. E., Thorsson, V., Karp, R.M. (2000). "Discovery of Regulatory Interactions Through Perturbation: Inference and Experimental Design," *Pacific Symposium on Biocomputing* 5, Hawaii.

Kosko, B. (1986). "Fuzzy Cognitive Maps." *International Journal Man-Machine Studies* **24**: 65-75.

Kosko, B. (1986). "Fuzzy Knowledge Combination." *International Journal of Intelligent Systems* **1**: 293-320.

Lee, K., S. Kim, et al. (1996). "On-line fault diagnosis by using fuzzy cognitive maps." *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences* **E79-A,**(6): 921-922.

Liang, S., S. Fuhrman, et al. (1998). "REVEAL, A general reverse engineering algorithm for inference of genetic network architectures," *Pacific Symposium on Biocomputing* 3, Hawaii.

Matsuno, H., Doi, A., Nagasaki, M. and Miyano, S. (2000). "Hybrid Petri Net Representation of Gene Regulatory Network," *Pacific Symposium on Biocomputing* 5, Hawaii.

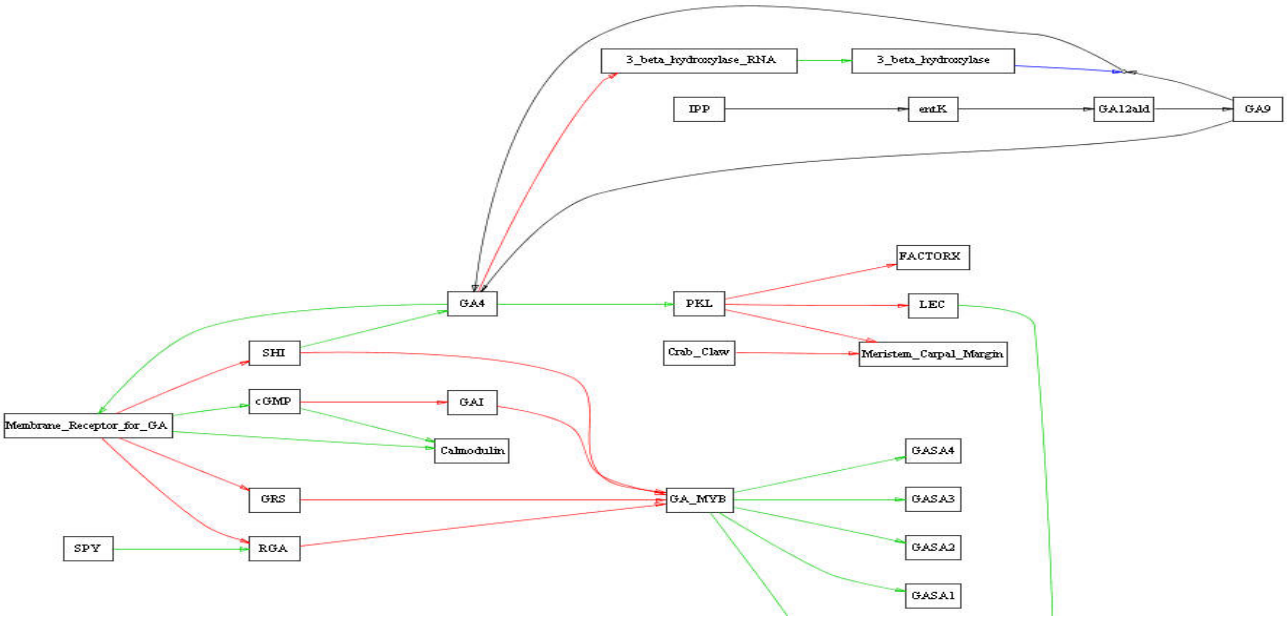
Ng, S.-K. and M. Wong. "Toward routine automatic pathway discovery from on-line scientific text abstracts," *Genome Informatics*, 10:104--112, 1999.

Reisig, W. and G. Rozenberg (1998). Lectures on Petri Nets I: Basic Models. Berlin, Springer.

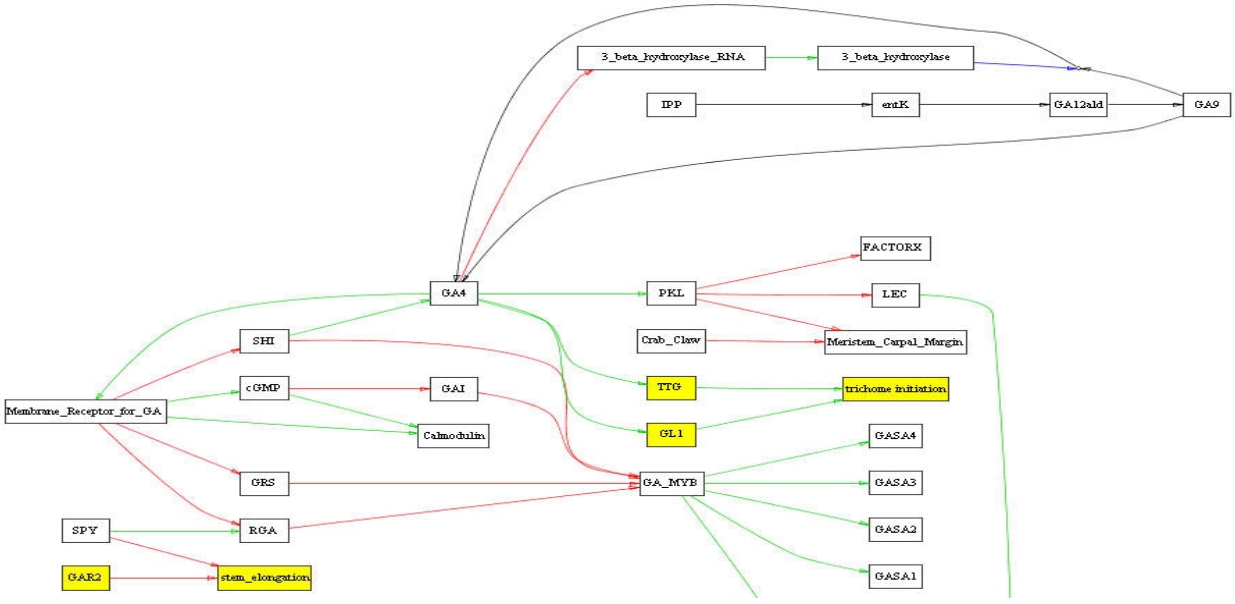
Rindflesch, T.C., L. Hunter, A. R. Aronson, "Mining Molecular Binding Terminology from Biological Text," *Proceedings of the AMIA '99 Annual Symposium.*

Thomas, J., D. Milward, C. Ouzounis, S. Pulman, and M. Carrol, "Automatic Extraction of Protein Interactions from Scientific Abstracts," *Pacific Symposium on Biocomputing* 5:538-549, 2000.

Weaver, D. C., C. T. Workman, et al. (1999). Modeling Regulatory Networks with Weight Matrices. Pacific Symposium on Biocomputing 4, Hawaii.



3a



3b

Figure 3. a) shows a small part of a graph for the Arabadopsis metabolic and regulatory network. b) shows the original graph plus the three new links discovered in the PathBinder search which are highlighted in yellow.