Chapter 17

CREATING, MODELING, AND VISUALIZING METABOLIC NETWORKS

FCModeler and PathBinder for Network Modeling and Creation

Julie A. Dickerson[1,2], Daniel Berleant[1,2], Pan Du[1,2], Jing Ding[1,2], Carol M. Foster[3], Ling Li[3], and Eve Syrkin Wurtele[2,3]

[1]Electrical and Computer Engineering Dept; [2]Virtual Reality Applications Center; [3]Genetics Development and Cell Biology, Iowa State University, Ames, IA

Chapter Overview

Metabolic networks combine metabolism and regulation. These complex networks are difficult to understand and create due to the diverse types of information that need to be represented. This chapter describes a suite of interlinked tools for developing, displaying, and modeling metabolic networks. The metabolic network interactions database, MetNetDB, contains information on regulatory and metabolic interactions derived from a combination of web databases and input from biologists in their area of expertise. PathBinderA mines the biological "literaturome" by searching for new interactions or supporting evidence for existing interactions in metabolic networks. Sentences from abstracts are ranked in terms of the likelihood that an interaction is described and combined with evidence provided by other sentences. FCModeler, a publicly available software package, enables the biologist to visualize and model metabolic and regulatory network maps. FCModeler aids in the development and evaluation of hypotheses, and provides a modeling framework for assessing the large amounts of data captured by high-throughput gene expression experiments.

Keywords

fuzzy logic; microarray analysis; gene expression networks; fuzzy cognitive maps; text mining; naïve Bayes

1. INTRODUCTION

The field of systems biology in living organisms is emerging as a consequence of publicly-available genomic, transcriptomics, proteomics, and metabolomics datasets. These data give us the hope of understanding the molecular function of the organism, and being able to predict the consequences to the entire system of a perturbation in the environment, or a change in expression of a single gene. In order to understand the significance of this data, the functional relationships between the genes, proteins, and metabolites must be put into context. This chapter describes an iterative approach to exploring the interconnections between biomolecules that shape form and function in living organisms. This work focuses on the model plant system, Arabidopsis. The systems biology approach itself can be used as a prototype for exploration of networks in any species.

The two major motivating currents in this work are the need to build systems for biologists and the need to better understand the science of knowledge extraction from biological texts. The biologist's information about the function of each RNA and protein is limited. Currently, about 50% of Arabidopsis genes are annotated in databases (e.g., TAIR (Rhee, Beavis et al., 2003) or TIGR (www.tigr.org)). In part because the process of evolution results in families of genes with similar sequences and related functions, much of the available annotation is not precise, and some annotation is inaccurate. Even more limited is our understanding of the interactions between these biomolecules. To help bridge this gap, metabolic networks are being assembled for Arabidopsis (e.g., AraCyc, KEGG). To date, these contain many derived pathways based on other organisms; consequently, they have errors and do not capture the subtleties of the Arabidopsis (or even plant) biochemistry and molecular biology that are necessary for research.

Considerable high-quality information is buried in the literature. A given pathway is known predominantly to those researchers working in the area. Such a pathway is not easily generated by curators whom are not experts in the particular field. This information is not rapidly accessible to a biologist examining large and diverse datasets and investigating changing patterns of gene expression over multiple pathways in which she/he may have little expertise. Furthermore, the interconnections between the multiple complex pathways of a eukaryotic organism cannot be envisioned without computational aid. To assist biologists in drawing connections between genes, proteins and metabolites, cumulative knowledge of the known and hypothesized metabolic and regulatory interactions of Arabidopsis must be supported by advanced computing tools integrated with the body of existing knowledge.

2. OVERVIEW

2.1 Metabolic Pathway Databases

The database used in this work is MetNetDB (Wurtele, Li et al., 2003). MetNetDB, combines knowledge from experts, Aracyc (Mueller, Zhang et al., 2003) and more specialized pathway sequence data, with experimental data from microarrays, proteomics and metabolomics and dynamically displays the results in FCModeler (Dickerson, D. Berleant et al., 2003; Wurtele, Li et al., 2003). The database is designed to include information about subcellular location, and to handle both enzymatic and regulatory interactions.

There are a few major database projects designed to capture pathways: What Is That? (WIT) Project (Overbeek, Larsen et al., 2000) (http://wit.mcs.anl.gov/WIT2/WIT), Kyoto Encyclopedia of Genes and Genomes (KEGG (Kanehisa and Goto, 2000), and EcoCyc/MetaCyc (Karp, Riley et al., 2000; Karp, Riley et al., 2002)).

WIT and KEGG contain databases of metabolic networks, which focus on prokaryotic organisms. The WIT2 Project produced static "metabolic reconstructions" for sequenced (or partially sequenced) genomes from the Metabolic Pathway Database. KEGG computerizes current knowledge of molecular and cellular biology in terms of the interacting genes or molecules and links the pathways with the gene catalogs being produced by the genome sequencing projects. EcoCyc is a pathway/genome database for *E. coli* that describes its enzymes and transport proteins. It has made significant advances in visualizing metabolic pathways using stored layouts, and linking data from microarray tests to the pathway layout (Karp, Krummenacker et al., 1999; Karp, 2001). The metabolic-pathway database, MetaCyc, describes pathways and enzymes for many different organisms (e.g. *Arabidopsis thaliana*, AraCyc), and combines information from sequences.

Other database designs emphasize data visualization. Cytoscape visualizes existing molecular interaction networks and gene expression profiles and other state data using Java (Shannon, Markiel et al., 2003). Cytoscape has facilities for constructing networks and displaying annotations from fixed files. MetNetDB is web-accessible and users can create their own custom-pathways, which can then be used to analyze expression data.

2.2 Network Modeling and Reconstruction

There have been many attempts to reconstruct gene regulatory and metabolic networks from microarray data using machine learning methods

(Weaver, Workman et al., 1999; Akutsu, Miyano et al., 2000; D'haeseleer, Liang et al., 2000; Matsuno, 2000; Genoud, Trevino Santa Cruz et al., 2001; Hartemink, Gifford et al., 2001; Wessels, Someren et al., 2001; Hanisch, Zien et al., 2002). However, these methods are based on the assumption that genes in the same pathways are co-regulated and show the same expression patterns. This assumption does not always hold for genes in pathways. Additionally, several pathways can show similar responses to a stimulus which leads to many false positive links. Data must be combined from multiple sources such as gene function data and expert information to give a complete picture of the interactions.

2.3 Extracting Biological Interactions from Text

Mining of the biological "literaturome" is an important module in a comprehensive creation, representation, and simulation system for metabolic and regulatory networks. Without it, many biomolecular interactions archived in the literature remain accessible in principle but underutilized in practice. Competitions to test the performance of automatic annotation such as the BioCreative Workshop (EMBO BioCreative Workshop, 2004), the TReC (Text Retrieval Conference) genomic track, and the KDD Cup 2002 show encouraging results, but high rates of error show that the systems are not yet accurate enough. None of these competitions directly focused on the problem of finding, and combining, evidence from sentences describing biomolecular interactions. This is a key need for a biological database system like MetNetDB, in which evidence provided by sentences must be rated to support ranking in terms of the likelihood that an interaction is described, must be combined with evidence provided by other sentences, and must support efficient human curation. Furthermore, sentence-based retrieval can be useful in and of itself to biologists, who are typically limited to retrieval based on larger text units as supported e.g. by PUBMED and Agricola in the biological domain and common Web search engines in general.

2.3.1 Empirical Facts about Biological Texts.

Although many researchers have investigated mining of biomolecular interactions from text, the reporting of empirical facts about interaction descriptions remains quite limited. Craven and Kumlien (Craven and Kumlien, 1999) investigated word stems and the ability of each to predict that a sentence describes the subcellular location of the protein if it contains a stem, a protein name, and a subcellular location. Marcotte et al., (Marcotte, Xenarios et al., 2001) gave a ranked list of 20 words found useful

in identifying abstracts describing protein interactions in yeast-related abstracts. Results were derived from and therefore may be yeast-specific. Ono et al., (Ono, Hishigaki et al., 2001) quantitatively assessed the abilities of four common interaction-indicating terms, each associated with a custom set of templates, to indicate protein-protein interactions. The quantitative performances of the four are hard to interpret because each used a different template set, but it is interesting that their ranks in terms of precision were the same for both the yeast and the *E. coli* domains, suggesting domain independence for precision.

Thomas et al., (Thomas, Milward et al., 2000) proposed four categories of biological text passages using a rule-based scoring strategy, and gave the information retrieval (IR) performance of each category. Sekimizu et al., (Sekimizu, Park et al., 1998) measured the IR performances of 8 interaction-indicating verbs in the context of a shallow parser. The IR capabilities of the verbs could be meaningfully compared, but whether these results would hold across different parsers or other passage analyzers is an open question. In our lab, we have obtained similar results using passages containing two protein names. Counting passages describing interactions as hits and others as misses, sentences had slightly higher IR effectiveness than phrases despite lower precision, and considerably higher IR effectiveness than whole abstracts (Ding, Berleant et al., 2002). Ding et al., (Ding, 2003; Ding, Berleant et al., 2003) applied an untuned link grammar parser to sentences containing protein co-occurrences, finding that using the presence of a link path as an additional retrieval criterion raised the IR effectiveness by 5 percentage points (i.e. 7%). These works highlight the gap in knowledge of empirical facts about biological texts. In the future, researchers will focus increasing attention on this important gap.

2.3.2 Combining Evidence

Combining different items of evidence can result in a single composite likelihood that a sentence describes a biomolecular or other interaction. This can enable putative interactions in automatically generated biomolecular interaction network simulators to be rated, or sentences to be ranked for human curation. The following paragraphs compare two methods for evidence combination: the Naïve Bayes model and semi-naïve evidence combination.

Naïve Bayes and semi-naïve evidence combination both have a similar scalability advantage over full Bayesian analysis using Bayes Theorem to account for whatever dependencies may exist. That scalability is why they are useful. However, when used to estimate probabilities that an item (e.g. a sentence) is in some category (e.g. describes a biomolecular interaction),

semi-naïve evidence combination makes fewer assumptions (Berleant, 2004).

Evidence combination with the Naïve Bayes model. This standard method produces probability estimates that can used for categorization (Lewis, 1998). The formula is:

$$p(h \mid f_1,\ldots f_n) = \frac{p(h)\,p(f_1,\ldots f_n \mid h)}{p(f_1)\,p(f_2)\,p(f_3)\ldots p(f_n)} \approx \frac{p(h)\,p(f_1 \mid h)\,p(f_2 \mid h)\,p(f_3 \mid h)\ldots p(f_n \mid h)]}{p(f_1)\,p(f_2)\,p(f_3)\ldots p(f_n)} \quad (1)$$

where h is the probability that a sentence is a "hit" (has a description of the expected interaction), and f_i is feature i. The approximation provides a computationally tractable way to estimate the desired probability given the assumption that the features occur independently of one another. A readable derivation is provided by Wikipedia (Wikipedia: The Free Encyclopedia, 2004).

Semi-naïve evidence combination. This method is scalable in the number of features, like Naïve Bayes, but has the advantage of making fewer independence assumptions. Unlike the Naïve Bayes model, it does not assume that the features are independent regardless of whether sentences are hits or not.

The parsimonious formula for semi-naïve evidence combination, in terms of odds (the ratios of hits to misses) is (Berleant, 2004):

$$O(h|f_1,\ldots,f_n)=O_1\ldots O_n/(O_0)^{n-1} \quad (2)$$

where the odds that a sentence describes an interaction if it has features $f_1,\ldots,f_n$ are $O(h|f_1,\ldots,f_n)$. The odds that a sentence with feature k is a hit are O_k, and the prior odds (i.e. over all sentences in the test set irrespective of their features) that a sentence is a hit are O_0. The odds of flipping a head are 1/1=1 (1 expected success per failure), while the odds of rolling a six are 1/5 (one success expected per five failures). Odds are easily converted to the more familiar probabilities by applying $p=O/(O+1)$. Similarly, $O=p/(1-p)$.

Comparison of the Naïve Bayes and semi-naïve evidence combination models. Naïve Bayes is often used for category assignment. The item to be classified is put into the category for which Naïve Bayes gives the highest likelihood. In the present context there are two categories, one of hits and one of non-hits, but in general there can be N categories. In either case, the denominator of the Naïve Bayes formula is the same for each category, so it can be ignored. However, when the Naïve Bayes formula is used for estimating the *probability* that a sentence is in a particular category, the denominator must be evaluated. This is problematic because the assumption

of unconditional independence is not only unsupported, but most likely *wrong*. The reason is that the features that provide evidence that the sentence belongs in a particular category are probably correlated.

For the problem of estimating the *probability* that a particular sentence is a hit (or, more generally, belongs to a particular category), semi-naïve evidence combination appears more suitable because it estimates odds (which are easily converted to probabilities) without requiring the problematic assumption that features occur unconditionally independently (i.e. independently regardless of whether the sentence is a hit or not).

3. METNET

MetNet is designed to provide a framework for the formulation of testable hypotheses regarding the function of specific genes, proteins, and metabolites, and in the long term provide the basis for identification of genetic regulatory networks that control plant composition and development. Our approach to reveal complex biological networks is to extract information from gene expression data sets and combine it with what is already know about metabolic and regulatory pathways to achieve a better understanding of how metabolism is regulated in a eukaryotic cell.

3.1 Metabolic Networking Data Base (MetNetDB)

A critical factor both in establishing an efficient system for mining the literaturome and in modeling network interactions is the network database itself. MetNetDB is a searchable database with a user-friendly interface for creating and searching the *Arabidopsis* network map (Wurtele, Li et al., 2003). ***MetNetDB contains a growing metabolic and regulatory map of Arabidopsis.*** Entities (represented visually as nodes) in the database include metabolites, genes, RNAs, polypeptides, protein complexes, and 37 hierachically-organized interaction types, including catalysis, conversion, transport, and various types of regulation. MetNetDB currently contains more than 50,000 entities (from KEGG, TAIR and BRENDA), 1000 expert-user-added entity definitions, and 2785 expert-user-added interactions, including transport, together with associated information fields. In addition, it contains interactions from *Arabidopsis* Lipid Gene Database, and partially curated interactions from AraCyc. Synonyms for each term in MetNetDB are obtained from sources including expert users, TAIR, and BRENDA; an adequate library of synonyms is particularly important in text mining. Database nomenclature is modeled after the *Arabidopsis* Gene Ontology

(http://arabidopsis.org/info/ontologies/), for ease of information transfer between MetNetDB and other biological databases.

3.2 FCModeler: Network Visualization and Modeling

FCModeler is a Java program that dynamically displays complex biological networks and analyzes their structure using graph theoretic methods. Data from experiments (i.e., microarray, proteomics, or metabolomics) can be overlayed on the network map.

3.2.1 Network Visualization and Graph Theoretic Analysis

Visual methods allow the curator to investigate the pathway one step at a time and to compare different proposed pathways. Graph union and intersection functions assist curators in highlighting these differences. FCModeler uses graph theoretic methods to find cycles and alternative paths in the network. Alternative path visualizations help curators search for redundant information in pathways. For example, a sketchy pathway may need to be replaced with more details as they become available. Cycles in the metabolic network show repeated patterns. These cycles range from simple loops, for example, a gene causing a protein to be expressed, and accumulation of the protein inhibiting the gene's transcription. More complex cycles encompass entire metabolic pathways. The interactions or overlaps between the cycles show how these control paths interact. FCModeler searches for elementary cycles in the network. Many of the cycles in a pathway map are similar, and several similarity measures and pattern recognition models are available for grouping or clustering the cycles (Cox, Fulmer et al., 2002; Dickerson and Cox, 2003).

3.2.2 Multi-Scale Fuzzy K-Means Clustering

The analysis and creation of gene regulatory networks involves first clustering the data at different levels, then searching for weighted time correlations between the cluster center time profiles. Link validity and strength is then evaluated using a fuzzy metric based on evidence strength and co-occurrence of similar gene functions within a cluster. The Fuzzy K-means algorithm minimizes the objective function (Bezdek, 1981):

$$J(F,V) = \sum_{i=1}^{N} \sum_{j=1}^{K} m_{ij}^2 d_{ij}^2 \qquad (3)$$

$F = \{X_i,\ i = 1,...,N\}$ are the N data samples; $V = \{V_j,\ j = 1,...,K\}$ represents the K cluster centers. m_{ij} is the membership of X_i in cluster j, and d_{ij} is the Euclidean distance between and V_j. This work uses a windowed membership function:

$$m_{ij} = \frac{1/d_{ij}^2}{\sum_{k=1}^{K} 1/d_{ik}^2} W\left(d_{ij}\right) \tag{4}$$

Adding a window function $W(d)$ to the membership function limits the size of clusters. This work uses truncated Gaussian windows with values outside the range of 3σ set to zero:

$$W\left(d_{ij}\right) = \begin{cases} e^{-\left(d_{ij}\right)^2 / \left(2\sigma^2\right)} & d_{ij} < 3\sigma \\ 0 & elsewhere \end{cases} \tag{5}$$

The window function, $W(d)$, insures that genes with distances larger than 3σ from the cluster center will have no effect on the new cluster center estimates.

3.2.3 Multi-Scale Algorithm

The multi-scale algorithm is similar to the ISODATA algorithm with cluster splitting and merging (Ball, 1965; Ball and Hall, 1965). There are four parameters: K (initial cluster number), σ (scale of the window $W(d)$), T_{split} (split threshold), $T_{combine}$ (combine threshold). Whenever the genes are further away from the cluster center than T_{split}, the cluster is split and faraway genes form new clusters. Also, if two cluster centers are separated by less than $T_{combine}$, then the clusters are combined. Usually $T_{combine} \le \sigma$ and $2\sigma \le T_{split} \le 3\sigma$. The algorithm is given in Table 17-1. ε_1 and ε_2 are small numbers to determine whether the clustering converged, and $\varepsilon_1 > \varepsilon_2$. If one cluster has elements far away from the cluster centers then the cluster is split. The advantage of this algorithm is that it dynamically adjusts the number of clusters based on the splitting and merging heuristics.

Table 17-1. Multi-Scale Fuzzy K-Means Algorithm

Step	Description
1	Initialize parameters: K, σ, T_{split} and $T_{combine}$
2	Iterate using Fuzzy K-means until convergence to a given threshold ε_1
3	Split process: do split if there are elements farther away from cluster center than T_{split}.

continued

Step	Description
4	Iterate using Fuzzy K-means until convergence to a given threshold ε_1
5	Combine Process: combine the clusters whose distance between cluster centers is less than $T_{combine}$. If the cluster after combining has elements far away from cluster center (distance larger than 3σ), stop combining.
6	Iterate steps 1-5 until no splits or combination occur. Converging to a given threshold ε_2.

3.2.4 Effects of Window Size

Changing the window size can affect the level of detail captured in the clusters. If σ<<1, then clusters are individual elements. As σ increases, the window gets larger. The result is a hierarchical tree that shows how the clusters interact at different levels of detail. This work uses three level of multi-scale fuzzy K-mean clustering (σ=0.1, 0.2 and 0.3). The initial number of clusters is *K* = *N*, the total number of data points, $T_{combine} = \sigma$, and $T_{split} = 3\sigma$. Clustering results with different window sizes provide different levels of information. At σ=0.1, the cluster sizes are very small. These clusters represent very highly correlated profiles (correlation coefficients between gene profiles within 1- σ window size are larger than 0.9) or just the individual gene profiles because many clusters only contain a single element. At σ=0.2, smaller clusters are combined with nearby clusters. Highly correlated profiles are detected. The σ=0.3 level is the coarsest level.

3.2.5 Construction of Rene Regulatory Networks

Clustering provides sets of genes with similar RNA profiles. The next step is finding the relationships among these coregulated genes. If gene *A* and gene *B* have similar expression profiles, there are several possible relationships: 1. *A* and *B* are coregulated by other genes; 2. *A* regulates *B* or vice versa; 3. There is no causal relationship, just coincidence. Here the regulation may be indirect, i.e., interact through intermediates. These cases cannot be differentiated solely by clustering. We use cubic spline interpolation for simplicity and get equally sampled profiles.

The gene regulatory model can be simplified as a linear model (D'Haeseleer, Liang et al., 1999):

$$x_A(t+\tau_A) = \sum_B w_{BA} x_B + b_A \tag{6}$$

x_A is the expression level of gene *A* at time *t*, τ_A is the gene regulation time delay of gene *A*, w_{BA} is the weight indicating the inference of gene *B* to *A*, b_A is a bias indicating the default expression level of gene *A* without regulation.

Standardizing gene expression profiles to 0 mean and 1 standard deviation removes the bias term, b_A. The goal is to find out if genes *A* and *B* have a regulatory relationship so the weight is $w_{AB} = [0,1]$ (0 means no regulatory relation, 1 means strongly regulated). The time correlation between genes *A* and *B* can be expressed in discrete form as

$$R_{AB}(\tau) = \sum_n x_A(n) x_B(n-\tau) \tag{7}$$

Where x_A and x_B are the standardized (zero mean, standard deviation of unity) expression profiles of genes *A* and *B*. τ is the time shift. For the periodic time profile, we can use circular time correlation, i.e., the time points at the end of the time series will be rewound to the beginning of series after time shifting. For multiple data sets, the time correlation results of each data set are combined as:

$$R_{AB}^C(\tau) = \sum_k w_k R_{AB}^k(\tau) \tag{8}$$

Where $R_{AB}^C(\tau)$ is the combined time correlation result, $R_{AB}^k(\tau)$ is the time correlation result of the k^{th} data set, w_k is the weight of k^{th} data set that depends on the experiment reliability and the length of the expression profile.

The value $\max|R_{AB}^C(\tau)|$ can be used to estimate the time delay τ' between expression profiles of genes *A* and *B*. Given a correlation threshold T_R, if $\max|R_{AB}^C(\tau)| > T_R$ there is significant regulation between genes or clusters. By defining the clusters as nodes and significant links as edges, we can get the gene regulation network of these clusters. Assuming that the time delays are caused by regulation, we can define four types of regulation:

$R_{AB}^C(\tau') > 0$, $\tau' \neq 0$, positive regulation between genes *A* and *B*;
$R_{AB}^C(\tau') < 0$, $\tau' \neq 0$, negative regulation between genes *A* and *B*;
$R_{AB}^C(\tau') > 0$, $\tau' = 0$, genes *A* and *B* are positively coregulated;
$R_{AB}^C(\tau') < 0$, $\tau' = 0$, genes *A* and *B* are negatively coregulated.

The sign of τ' determines the direction of regulation. $\tau' > 0$ means gene *B* regulates gene *A* with time delay τ'; $\tau' < 0$ means gene *A* regulates gene *B* with time delay τ'.

3.3 Network Validation Using Fuzzy Metrics

The available gene ontology (GO) annotation information can estimate a fuzzy measure for the types or functions of genes in a cluster. The GO terms in each cluster are weighted according to the strength of the supporting evidence information for the annotation of each gene and the distance to cluster center. An additive fuzzy system is used to combine this

information(Kosko, 1992). Every GO annotation indicates the type of evidence that support it. Among these types of evidence, several are more reliable and several are weaker. This evidence is used to set up a bank of fuzzy rules for each annotated data point. Different fuzzy membership values are given to each evidence code. For example, evidence inferred by direct assays (IDA) or from a traceable author statement (TAS) in a refereed journal has a value of one. The least reliable evidence is electronic annotation since it is known to have high rates of false positives.

Table 17-2. Evidence Codes and Their Weights

Evidence Code	Meaning of the Evidence Code	Membership Value, w_{evi}
IDA	Inferred from direct assay	1.0
TAS	Traceable author statement	1.0
IMP	Inferred from mutant phenotype	0.9
IGI	Inferred from genetic interaction	0.9
IPI	Inferred from physical interaction	0.9
IEP	Inferred from expression pattern	0.8
ISS	Inferred from sequence, structural similarity	0.8
NAS	Non-traceable author statement	0.7
IEA	Inferred from electronic annotation	0.6
	Other	0.5

Each gene in a cluster is weighted by the Gaussian window function in equation (5). This term weights the certainty of the gene's GO annotation using product weighting. Each gene and its associated GO term are combined to find the possibility distribution for each single GO term that occurs in the GO annotations in one cluster. One gene may be annotated by several GO terms, and each GO term has one evidence code. Each GO term may occur *K* times in one cluster, but with a different evidence code and in different genes. For the *n*th unique GO term in the *j*th cluster, the fuzzy weight is the sum of the weights for each occurrence of the term:

$$W_{GO}(j,n) = \sum_{i=1}^{K} w_{GO,j}(i,n) \tag{9}$$

Where $w_{GO,j}(i,n) = w_{evi}(i,n) \cdot W\left(d_{ij}\right)$, w_{evi} is shown in Table 17-2, and $W(d_{ij})$ is the same as equation (5).

This provides a method of pooling uncertain information about gene function for a cluster of genes. This gives an additive fuzzy system that assesses the credibility of any GO terms associated to a cluster (Kosko 1992). The results can be left as a weighted fuzzy set or be defuzzified by selecting the most likely annotation. For each cluster, the weight is

normalized by the maximum weight and the amount of unknown genes. This is the weighted percentage of each GO term p_{weight} :

$$p_{weight}(j,n) = \frac{W_{GO}(j,n)}{W_{root}(j) - W_{unknown}(j)} * 100\% \quad (10)$$

Where $W_{GO}(j,n)$ represents the weight of the *n*th GO term in the *j*th cluster. $W_{unknown}(j)$ is the weight of GO term in cluster *j*: xxx unknown, e.g., GO: 0005554 (molecular_function unknown). $W_{root}(j)$ is the weight of root in cluster *j*. GO terms are related using directed acyclic graphs. The root of the graph is the most general term. Terms further from the root provide more specific detail about the gene function and are more useful for a researcher. The weight of each node is computed by summing up the weights of its children (summing the weights of each of the *N* GO terms in a cluster):

$$W_{root}(j) = \sum_{n=1}^{N} W_{GO}(j,n) \quad (11)$$

The higher weighted nodes further from the root are the most interesting since those nodes refer to specific biological processes.

3.4 PathBinderA: Finding Sentences with Biomolecular Interactions

The objective of the PathBinderA component of the system is to mine sentences describing biomolecular interactions from the literature. This functionality forms a potentially valuable component of a range of systems, by supporting systems for automatic network construction, systems for annotation of high-throughput experimental results, and systems that minimize the high costs of human curation. Such a component should typically mine all of MEDLINE, the *de facto* standard corpus for bioscience text mining. For the plant domain, full texts in the plant science domain should also be addressed, requiring cooperative agreements with publishers in general. The feasibility of different PathBinder components is illustrated by the system at http://metnetdb.gdcb.iastate.edu/pathbinder/. The PathBinderA system, used in this work, has been prototyped and is undergoing further development.

Attaining the desired results requires a well-motivated and tested method for processing biological texts. The design includes a two-stage algorithm. Each stage is based on probability theory. In stage 1, evidence for interaction residing in sentence features is combined to compute the sentence's credibility as an interaction description. In stage 2, the credibilities of the "bag" of sentences that mention two given biomolecules are combined to

rate the likelihood that the literature describes those biomolecules as interacting. The practical rationale for this process is that an important resource is being created for use by the scientific community, as well as an important module of the overall MetNetDB system. This resource is aimed at effective curation support, which in turn is aimed at feeding the construction of interaction networks.

3.4.1 PathBinder Component Design Issues

There are three major phases of a PathBinder component such as the PathBinderA component of METNET. The mining process comprises stages 1 and 2, and using the results of the mining constitutes the third phase.

Text mining, stage 1. This involves assessing the credibility of a given sentence as a description of an interaction between two biomolecule names in it. To do this the evidence provided by different features of a sentence must be combined. Semi-naïve evidence combination, described earlier in this chapter, is one such method. The Naïve Bayes model provides another possibility. Syntactic parsing to analyze sentences in depth is an alternative approach.

The abilities of various features of sentences to predict whether they describe an interaction can be determined empirically in order to enable those features to be used as input to a method for assessing sentences. One such feature is whether a sentence with two biomolecule names has those names in the same phrase or, instead, the names occur in different phrases within the sentence. We have investigated this feature using the IEPA corpus (Ding, Berleant et al., 2002). Table 17-3 (rightmost column) shows the results. Another feature is whether or not an interaction term intervenes between the co-occurring names. An interaction term is a word that can indicate that an interaction between biomolecules takes place, like "activates," "block," "controlled," etc. Such a term can appear between two co-occurring names, can appear in the same sentence or phrase but not between them, or can be absent entirely. The table below (middle two columns) shows the data we have collected, also using the IEPA corpus.

Table 17-3. Analysis of the recall and precision of co-occurrence categories with respect to mining interaction descriptions.

	Interactor intervening		Interactor elsewhere		Interactor anywhere	
Phrase co-occurrences	r=0.55	p=0.63	r=0.18	p=0.24	r=0.72	p=0.45
Sentence co-occurrences	r=0.22	p=0.30	r=0.058	p=.09	r=0.28	p=0.21
All co-occurrences	r=0.77	p=0.48	r=0.23	p=0.17	r=1	p=0.34

Text mining, stage 2. In this stage, the evidence for an interaction provided by multiple relevant sentences is combined to get a composite probability estimate for the interaction. This becomes possible after stage 1 has given a probability for each sentence. The basic concept underlying stage 2 is that if even one sentence in a "bag" containing two given names describes an interaction between them, then the interaction is present in the literature (Skounakis and Craven, 2003). The need for as little as a single example to establish an interaction leads directly to a probability calculation for combining the evidence provided by the sentences in a bag. The reasoning goes as follows.

Let notation $p(x)$ describe the probability of x. Assume the evidence provided by each sentence s_i in a bag is independent of the evidence provided by the other sentences, allowing us to multiply the probabilities of independent events to get the probability of their simultaneous occurrence.

p(one or more sentence in bag b describes an interaction between n_1 and n_2)
=1–p(zero sentences in bag b describe an interaction between n_1 and n_2),
=1–p(s_1 does not describe an interaction AND s_2 does not describe an interaction AND s_3...)
=1–p(s_1 does not describe an interaction)·p(s_2 does not describe an interaction)·p(s_3 does not describe…
=1–[1-p(s_1 describes an interaction)]·[1-p(s_2 describes an interaction)]· [1-p(s_3 does not describe…
$=1-\prod_i 1-p(s_i \text{ describes an interaction}).$

This equation is not only mathematically reasonable but considerably simpler than the more complex formulas given by Skounakis and Craven (Skounakis and Craven, 2003).

Using the mining results, stage 3. This phase integrates the extraction capability into the larger MetNetDB system. The integration supports the following functionalities.

1. *Support for curation.* Because networks of interactions are built from individual interactions, it is important not only to mine potential interactions from the literature but to present these to curators so that they can be efficiently verified. Curation is a serious bottleneck because it requires expert humans, a scarce resource. Therefore efficient support for curation is an important need. PathBinderA supports curation by presenting mined potential interactions to curators starting from the best, most likely interactions. The goal of this design is to minimize the labor required by the curation process.
2. *Generating interaction hypotheses.* When mining the literature produces a strong hypothesis of an interaction, that interaction may be tentatively

added to the interaction database without curation. Interactions whose probabilities are assessed at 90% or better are likely to fall into this category, although this threshold is adjustable. Such likely interactions are made available pending curation.

3. *More efficient literature access.* High-volume information resources can benefit from providing convenient access to the literature relevant to particular items in the resource. Such functionality is clearly useful to non-expert users, and even expert users can benefit since no individual can be intimately familiar with the full range of the literature on biomolecular interactions even in one species. The system design provides for integrating literature access with an easy-to-use community curation functionality. In this design, users anywhere can click a button associated with the display of any sentence they retrieve from the system. This brings up a form with two other buttons. One of these registers an opinion that the sentence describes an interaction, and one registers an opinion to the contrary. Comments may be typed into an optional comment area. Submitted forms will then be used by the official curators.

Users can choose *species and other taxa* to focus their search. Users may, for example, specify *viridiplantae* (green plants) to see sentences related to *Arabidopsis* as well as any other green plant species. The current prototype of PathBinderA allows users to specify two biomolecules, an interaction-relevant verb, and a subcellular location. Sentences with the two biomolecules and the verb which are associated with the specified subcellular location can then be retrieved (see Figure 17-1).

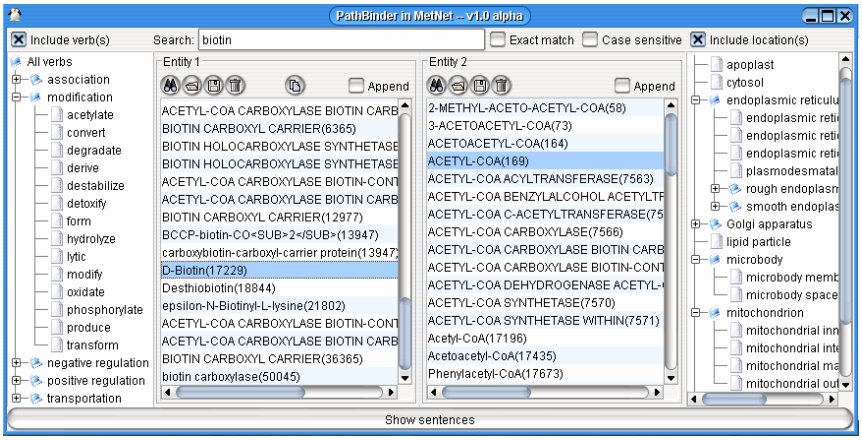

Figure 17-1. PathBinderA interface, showing the four choices a user can make. These choices include two biomolecules, a subcellular location, and an interaction-relevant verb.

4. BUILDING ON METABOLIC NETWORKS: USING METNET

Regulatory networks from Arabidopsis can be built using a combination of expert knowledge from MetNetDB, fuzzy clustering and correlation from FCModeler. The constructed networks can be validated using PathBinderA to access the literaturome and the weighted GO scores derived in FCModeler.

We illustrate with a test data set comparing wild-type (WT) *Arabidopsis thaliana* plants with those containing antisense *ACLA-1* behind the constitutive CaMV 35S promoter (referred to as *aACLA-1*) (Fatland, Nikolau et al., 2004). Plants were grown under a short day cycle of 8 hours light and 16 hours dark. Affymetrix GeneChip microarrays were used. The data consisted of two replicates; each with eleven time points (0, 0.5, 1, 4, 8, 8.5, 9, 12, 14, 16, 20 hours), harvested during the light (0 to 8 hours) and dark (8 to 20 hours) (Foster, Ling et al., 2003). Only *ACLA-1* seedlings exhibiting features characteristic of the antisense phenotype were used. Total RNA was extracted from leaves and used for microarray analyses.

The Affymetrix microarray data were normalized with the Robust Multichip Average (RMA) method (Gautier, Cope et al., 2004). Both replicates of each gene expression profile are standardized to zero mean, one standard deviation. The data was filtered by comparing the expression values of each gene between the WT and *ACLA-1* plants at 1, 8.5 and 12 hours; differentially expressed genes having larger than 2 fold changes at any time point were retained for further analysis. There are 484 such genes in total. The expression patterns of these 484 genes in the wild-type plants at all 11 time points were clustered.

4.1 Construct the Genetic Network Using Time Correlation

The genetic networks among the clusters of coregulated genes can be constructed based on their cluster center profiles. Since the data used were unequally sampled with 0.5h as minimum interval, we interpolated the gene expression profiles as equally sampled 41 time points with 0.5h intervals using cubic spline interpolation. The time correlation of each replicate was computed using equation (7), then combined using equation (8). The time period was limited to the range of [-4h, 4h] because the light period only lasted 8 hours in this data set. The genetic networks were constructed with a correlation threshold of $T_R = 0.65$. The strength of correlation was mapped into three categories: [0.65, 0.75), [0.75, 0.85), and [0.85, 1]. Three types of line thickness from thin to thick represent the strength of the correlation.

Black dashed lines represent positive coregulation; green dashed lines represent negative coregulation; solid red lines with bar head represent negative regulation; solid blue lines with arrowheads represent positive regulation.

Figure 17-2 shows the constructed gene regulatory networks based on the cluster center with window sigma equals 0.3. The networks indicate that cluster 1 and 5 are highly coregulated (0 time delay), cluster 1 and 5 positively regulate cluster 4 with a time delay of 2.5 hours and 3h, and both negatively regulated cluster 3 with a time delay of 1.5 hours; cluster 4 is negatively regulated by cluster 3 with a delay of 1 hour, the correlation between cluster 2 and cluster 4, and cluster 1 and 3 is not strong. All of these relations are coincident with the cluster center profiles.

Figure 17-3 shows the constructed regulatory networks of the 28 cluster centers at the σ= 0.2 level. The graph notations are the same as Figure 17-2. The graph shows that there is one highly connected group of clusters. The other clusters at the upper right corner are less connected. The relations between clusters may become complex with a large number of edges. Simplification of the networks is necessary when there are many highly connected clusters.

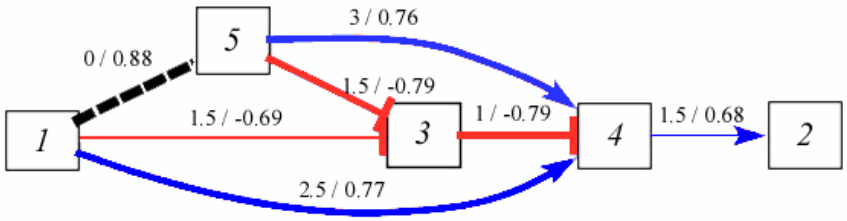

Figure 17-2. Gene regulatory networks inferred from the case with sigma equal to 0.3. The numbers on each link show the time delay for the interaction on top and the correlation coefficient of the interaction on the bottom.

Figure 17-3 shows possible duplicate relationships. This can be analyzed using the path search function in FCModeler. From cluster 15 to 19, there are two paths: one is directly from cluster 15 → 19 with time delay 1h and correlation coefficient, ρ = -0.85; another path is cluster 15 → 7 with time delay 0.5 h and correlation coefficient, ρ = -0.89, and then from 7→ 19 with time delay 0.5h and ρ = 0.81. The total time delays of both paths are the same. So it is very possible one of the paths is redundant. Figure 17-4 shows part of the simplified graph.

4.2 Cluster and Network Validation

Cluster validation can use the available literature and the GO annotation

for each gene to find out what kind of functions or processes occur within each cluster and to search for potential interactions between genes in different clusters. In Figure 17-3, the clusters in the portion of the graph at the upper right corner are less connected both to each other and to the main graph. Most of the genes in these "less-connected" clusters are not annotated in GO. This means these genes have no biological evidence of a direct relation with the highly connected group. It also shows how the fuzzy hierarchical algorithm successfully separates these genes.

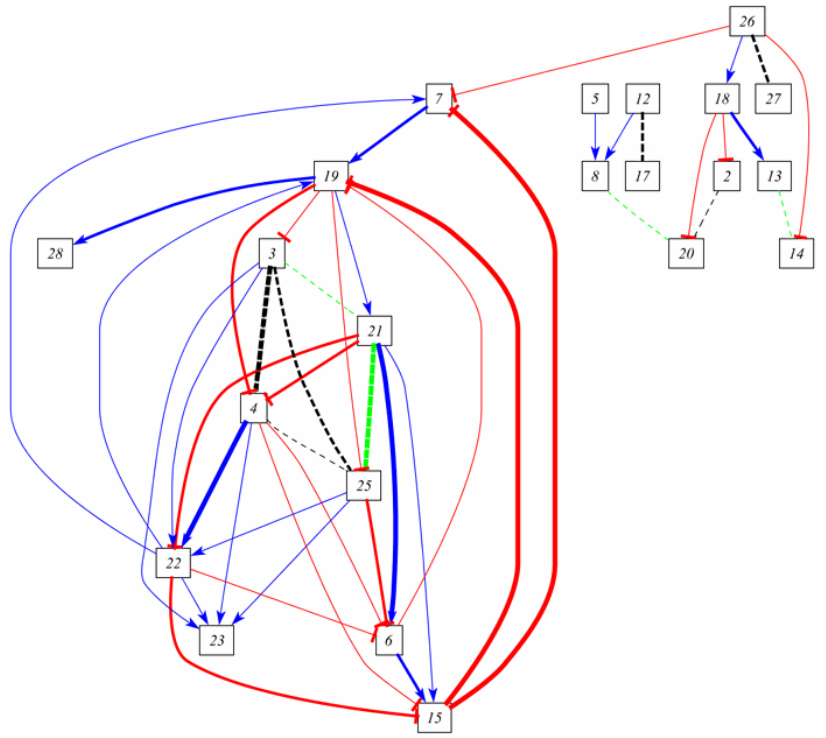

Figure 17-3.. Regulatory networks among cluster centers at the window size sigma = 0.2 level.

Figure 17-4 shows that clusters 3 and 4 are highly coregulated (correlation coefficient between cluster centers is 0.91). The cluster is split because the combined cluster 3 and 4 has a cluster diameter larger than 3σ. Table 17-4 shows the fuzzy weights for the GO terms for the genes in each cluster. The BP (Biological Process) GO annotations show that the genes in clusters 3 and 4 function in several similar biological processes. For example, both clusters contain genes of "Carboxylic acid metabolism", "Regulation of transcription, DNA-dependent", and "Protein amino acid phosphorylation". Cluster 3 has more "Regulation of transcription, DNA-dependent" genes, while cluster 4 emphasizes "Protein amino acid

phosphorylation" genes. Clusters 3 and 4 provide an example of the overlapping of fuzzy clusters in which the separation of two clusters may make sense, and suggests additional biological analyses.

Clusters 21 and 25 are two highly negatively coregulated clusters. Cluster 21 contains genes of "Photosynthesis, dark reaction" and hormone response genes, while cluster 25 mainly contains genes of catabolism and stress-associated genes. Cluster 21 contains genes for "Trehalose biosynthesis". Trehalose plays a role in the regulation of sugar metabolism, which has recently been identified for Arabidopsis (Eastmond and Graham 2003). Clusters 6 and 21 involve sugar metabolism (carbohydrate metabolism GO term). This provides an interesting biological implication for understanding regulation in this experiment.

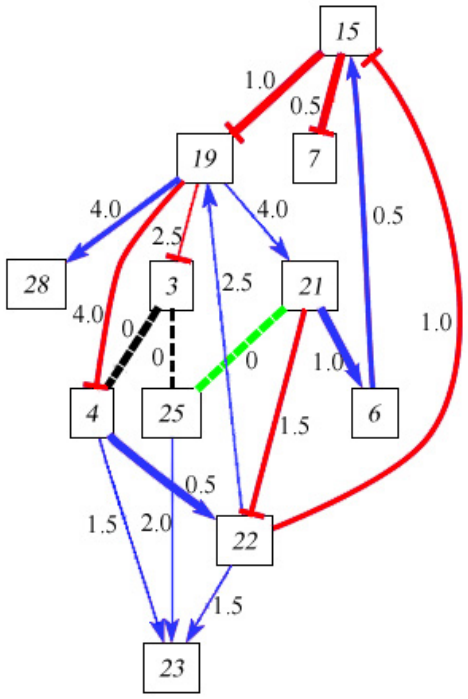

Figure 17-4. Simplified regulatory network with redundant edges removed for the window size σ = 0.2 level. The number on each link represents the estimated time delay.

Figures 17-3 and 17-4 show that cluster 19 regulates clusters 3, 4, 21, 22, 25 and 28. After checking the BP GO annotations, we found many of the annotated genes in cluster 19 fall in three categories: "Protein Metabolism" ("N-terminal protein myristoylation", and "Protein folding"), "Response to auxin stimulus", and "Cell-cell signaling". "N-terminal protein myristoylation," and "Protein folding" modulate protein activity, while

"Response to auxin stimulus" and "Cell-cell signaling" involve the processes of receiving stimulus or signals from others. Therefore these BP GO annotations are consistent with our network structures. Clusters 23 and 28 have no out-going edges. This implies that they may not be involved in regulatory activity. The genes in these clusters are metabolic or have no known function.

Table 17-4. Cluster annotation of Biological Process GO with the weights as defined in equations 9-11.

CLUSTER INDEX (W_{ROOT})	MAJOR GO TERM	$W_{GO}(J,N)$	$P_{WEIGHT}(J,N)$
Cluster 3	Response to water deprivation	4.11	16.6
(24.81)	Regulation of transcription, DNA-dependent	3.16	12.7
	Carboxylic acid metabolism	2.82	11.4
	Protein amino acid phosphorylation	2.63	10.6
Cluster 4	Protein amino acid phosphorylation	8.34	**23.1**
(36.03)	Carboxylic acid metabolism	3.58	9.9
	Response to abiotic stimulus	3.35	9.3
	Regulation of transcription, DNA-dependent	2.44	6.8
Cluster 6	Regulation of transcription, DNA-dependent	1.99	**23.5**
(8.48)	myo-inositol biosynthesis	0.95	11.2
	Abscisic acid mediated signaling	0.83	9.8
	Protein amino acid phosphorylation	0.57	6.7
Cluster 7	Carbohydrate metabolism	3.02	**22.2**
(13.58)	Cell surface receptor linked signal transduction	1.71	12.6
	Nucleobase, nucleotide, nucleic acid metabolism	1.62	11.9
	Protein amino acid phosphorylation	1.59	11.7
Cluster 15	Regulation of transcription, DNA-dependent	1.32	**52.4**
(2.52)	Electron transport	0.7	27.8
Cluster 19	Cell-cell signaling	0.78	23.5
(3.32)	Response to auxin stimulus	0.68	20.5
	Protein folding	0.65	19.6
	N-terminal protein myristoylation	0.61	18.4
Cluster 21	Carbohydrate metabolism	2.93	**29.1**
(9.71)	Response to gibberellic acid stimulus	1.86	19.2
	Photosynthesis, dark reaction	0.91	9.4
Cluster 22	Protein amino acid phosphorylation	6.74	**28.4**
(23.76)	Macromolecule biosynthesis	3.38	14.2
	Regulation of transcription DNA-dependent	2.50	10.5
	Signal transduction	2.30	9.7
Cluster 23	Response to endogenous stimulus	2.79	**60.5**
(4.61)	Response to biotic stimulus	1.83	39.7

continued

CLUSTER INDEX (W_{ROOT})	MAJOR GO TERM	$W_{GO}(J,N)$	$P_{WEIGHT}(J,N)$
Cluster 25	Carboxylic acid metabolism	8.19	**20.9**
(39.16)	Response to pest/pathogen/parasite	5.66	14.5
	Lipid biosynthesis	3.55	9.1
	Transport	3.52	9.0
Cluster 28	Carbohydrate metabolism	0.95	**100**
(0.95)			

Using PathbinderA to explore the relationship between genes in clusters is illustrated in a comparison between two genes in clusters 19 and 4. Cluster 19 contains the ethylene response gene "ethylene-induced esterase". Cluster 4 contains jasmonic acid response and several jasmonate biosynthesis genes. A search encompassed both these terms together with all of the synonyms for these terms in the MetNetDB database. We used "ethylene" and "jasmonate" to search in Pathbinder and retrieved 18 sentences (Figure 17-5 shows a subset of these sentences). Clicking on each sentence gives the entire abstract. Many of the sentences provided useful connections between these two nodes. For example, the abstract for the highlighted sentence delineates a relationship between the ethylene and jasmonate signaling pathways, as shown in Figure 17-6.

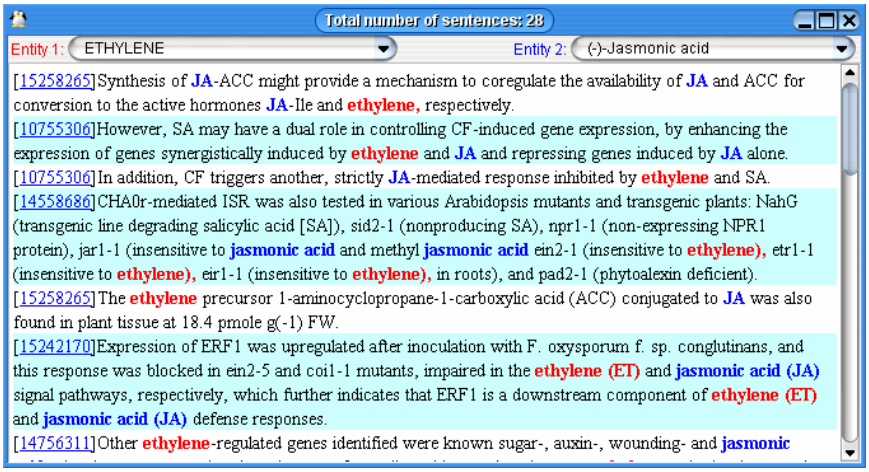

Figure 17-5. PathBinderA output for the terms ethylene and jasomonic acid. The relevant sentences and their Medline identification numbers are given.

5. DISCUSSION

The MetNet software focuses on understanding the complex molecular network in the model plant eukaryotic species, Arabidopsis. This work enables biologists to capture relationships at different levels of detail, to integrate gene expression data, and to model these relationships. Text mining in the PathBinderA system can help confirm relationships discovered by machine learning algorithms and will eventually be used to discover new relationships as methods of evidence combination are improved. Because of our absence of knowledge about many biological interactions, the software is designed to model at many levels of detail.

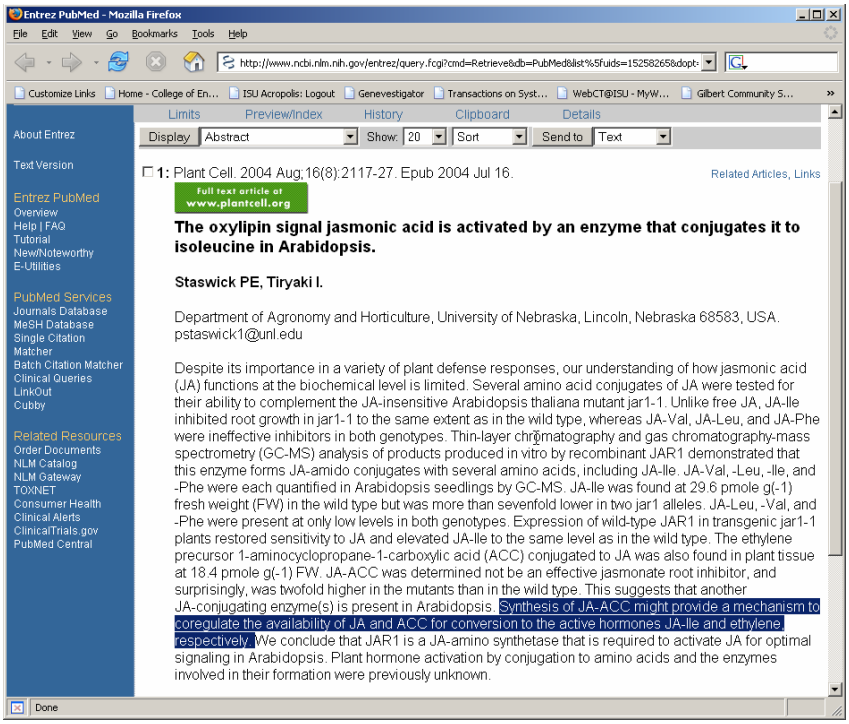

Figure 17-6.. The complete abstract for the selection shown above gives more details on the relationship between the ethylene and jasmonate signaling pathways.

6. ACKNOWLEDGEMENTS

This work is supported by grants from NSF (MCB-9998292, Arabidopsis 2010 DBI-0209809 and ITR-0219366), and the Plant Sciences Institute at Iowa State University. The network visualization was performed using the

facilities at the Virtual Reality Application Center at Iowa State University. We thank Lucas Mueller and TAIR for helpful advice and Aracyc data.

REFERENCES

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

Ball, G. H. (1965). "Data Analysis in the Social Sciences: What About the Details?" in *AFIPS Proc. Cong. Fall Joint Comp*, 27(1): 533-559.

Ball, G. H. and D. J. Hall (1965). *Isodata, a Novel Method of Data Analysis and Pattern Classification*, Stanford Research Institute.

Berleant, D. (2004). *Combining Evidence: The Naïve Bayes Model Vs. Semi-Naïve Evidence Combination*. Ames, IA, Software Artifact Research and Development Laboratory, Iowa State University.

Bezdek, J. C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*. New York, Plenum Press.

Cox, Z., A. Fulmer and J. A. Dickerson (2002). *Interactive Graphs for Exploring Metabolic Pathways*. ISMB, 2002, Edmonton, CA.

Craven, M. and J. Kumlien (1999). "Constructing Biological Knowledge Bases by Extracting Information from Text Sources," in *Proc Int Conf Intell Syst Mol Biol*: 77-86.

D'Haeseleer, P., S. Liang and R. Somogyi (1999). "Gene Expression Analysis and Modeling," in *Pacific Symposium of Biocomputing (*Tutorial).

D'haeseleer, P., S. Liang and R. Somogyi (2000). "Genetic Network Inference: From Co-Expression Clustering to Reverse Engineering*," Bioinformatics* 16(8): 707-26.

Dickerson, J. A. and Z. Cox (2003). "Using Fuzzy Measures to Group Cycles in Metabolic Networks," in *North American Fuzzy Information Processing Society (NAFIPS) Annual Meeting*, Chicago, IL.

Dickerson, J. A., D. Berleant, Z. Cox, W. Qi, D. Ashlock, E. S. Wurtele and A. W. Fulmer (2003)." Creating and Modeling Metabolic and Regulatory Networks Using Text Mining and Fuzzy Expert Systems" in *Computational Biology and Genome Informatics*. J. T. L. Wang, C. H. Wu and P. Wang. Singapore, World Scientific Publishing: 207-238.

Ding, J., D. Berleant, D. Nettleton and E. Wurtele (2002). "Mining Medline: Abstracts, Sentences, or Phrases?" in *Pacific Symposium on Biocomputing* (PSB 2002), Kaua'i, Hawaii.

Ding, J. (2003). *Pathbinder: A Sentence Repository of Biochemical Interactions Extracted from Medline.* Dept. of Electrical and Computer Engineering. Ames, IA, Iowa State University.

Ding, J., D. Berleant, J. Xu and A. W. Fulmer (2003). "Extracting Biochemical Interactions from Medline Using a Link Grammar Parser," in *Proceedings of the Fifteenth IEEE Conference on Tools with Artificial Intelligence (ICTAI 2003)*, Sacramento, CA, USA.

Eastmond, P. J. and I. A. Graham (2003). "Trehalose Metabolism: A Regulatory Role for Trehalose-6-Phosphate?" *Curr Opin Plant Biol* 6(3): 231-5.

EMBO BioCreative Workshop (2004). "A Critical Assessment for Information Extraction in Biology," (Biocreative), At. Granada, ES. 2004.

Fatland, B., B. J. Nikolau and E. S. Wurtele (2004). "Reverse Genetic Characterization of Cytosolic Acetyl-Coa Generation by Atp-Citrate Lyase in Arabidopsis*," Plant Cell*, in press.

Foster, C. M., L. Ling, A. M. Myers, M. G. James, B. J. Nikolau and E. S. Wurtele (2003). "Expression of Genes in the Starch Metabolic Network of Arabidopsis During Starch Sythesis and Degradation," In Preparation.

Gautier, L., L. Cope, B. Bolstad and R. Irizarry (2004). "Affy--Analysis of Affymetrix Genechip Data at the Probe Level," Bioinformatics 20(3): 307-315.

Genoud, T., M. B. Trevino Santa Cruz and J. P. Metraux (2001). "Numeric Simulation of Plant Signaling Networks," Plant Physiol. 126: 1430-1437.

Hanisch, D., A. Zien, R. Zimmer and T. Lengauer (2002). "Co-Clustering of Biological Networks and Gene Expression Data," in *Intelligent Systems for Molecular Biology (ISMB), 10th International Conference*, Edmonton, Canada, International Society for Computational Biology.

Hartemink, A. J., D. K. Gifford, T. S. Jaakkola and R. A. Young (2001). "Using Graphical Models and Genomic Expression Data to Statistically Validate Models of Genetic Regulatory Networks," in *Pacific Symposium on Biocomputing*, Hawaii.

Kanehisa, M. and S. Goto (2000). "Kegg: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Research* 28(1): 27-30.

Karp, P. D., M. Krummenacker, S. Paley and J. Wagg (1999). "Integrated Pathway/Genome Databases and Their Role in Drug Discovery," *Trends in Biotechnology* 17(7): 275-281.

Karp, P. D., M. Riley, M. Saier, I. T. Paulsen, S. M. Paley and A. Pellegrini-Toole (2000). "The Ecocyc and Metacyc Databases," *Nucleic Acids Research* 28(1): 56-59.

Karp, P. D. (2001). "Pathway Databases: A Case Study in Computational Symbolic Theories," *Science* 293(5537): 2040-4.

Karp, P. D., M. Riley, S. M. Paley and A. Pellegrini-Toole (2002). "The Metacyc Database," *Nucl. Acids. Res.* 30: 59-61.

Kosko, B. (1992). *Neural Networks and Fuzzy Systems.* Englewood Cliffs, Prentice Hall.

Lewis, D. (1998). "Naïve Bayes at Forty: The Independence Assumption in Information Retrieval," in *Conf. Proc. European Conference on Machine Learning*, Chemnitz, Germany.

Marcotte, E. M., I. Xenarios and D. Eisenberg (2001). "Mining Literature for Protein-Protein Interactions," *Bioinformatics* 17(4): 359-63.

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

Mueller, L. A., P. Zhang and S. Y. Rhee (2003). "Aracyc: A Biochemical Pathway Database for Arabidopsis," *Plant Physiol.* 132(2): 453–460.

Ono, T., H. Hishigaki, A. Tanigami and T. Takagi (2001). "Automated Extraction of Information on Protein-Protein Interactions from the Biological Literature," *Bioinformatics* 17(2): 155-61.

Overbeek, R., N. Larsen, G. D. Pusch, M. D'Souza, E. S. Jr, N. Kyrpides, M. Fonstein, N. Maltsev and E. Selkov (2000). "Wit: Integrated System for High-Throughput Genome Sequence Analysis and Metabolic Reconstruction," *Nucl. Acids. Res.* 28: 123-125.

Rhee, S. Y., W. Beavis, et al., (2003). "The Arabidopsis Information Resource (Tair): A Model Organism Database Providing a Centralized, Curated Gateway to Arabidopsis Biology, Research Materials and Community," *Nucl. Acids. Res.* 31(1): 224-228.

Sekimizu, T., H. S. Park and J. Tsujii (1998). "Identifying the Interaction between Genes and Gene Products Based on Frequently Seen Verbs in Medline Abstracts," *Genome Inform Ser Workshop Genome Inform* 9: 62-71.

Shannon, P., A. Markiel, O. Ozier, N. Baliga, J. Wang, D. Ramage, N. Amin, B. Schwikowski and T. Ideker (2003). "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks," *Genome Research* 13(11): 2498-504.

Skounakis, M. and M. Craven (2003). "Evidence Combination in Biomedical Natural-Language Processing," in *3rd ACM SIGKDD Workshop on Data Mining in Bioinformatics.*

Thomas, J., D. Milward, C. Ouzounis, S. Pulman and M. Carroll (2000). "Automatic Extraction of Protein Interactions from Scientific Abstracts," *in Pacific Symposium of Biocomputing*: 541-52.

Weaver, D. C., C. T. Workman and G. D. Stormo (1999). "Modeling Regulatory Networks with Weight Matrices," in *Pacific Symposium on Biocomputing* 4, Hawaii.

Wessels, L. F. A., E. P. V. Someren and M. J. T. Reinders (2001). "A Comparison of Genetic Network Models," in *Pacific Symposium on Biocomputing*, Hawaii.

Wikipedia: The Free Encyclopedia (2004). Naive Bayesian Classification. Wikipedia: the free encyclopedia (http://en.wikipedia.org/). 2004.

Wurtele, E. S., J. Li, et al., (2003). "Metnet: Software to Build and Model the Biogenetic Lattice of Arabidopsis," *Comparative and Functional Genomics* 4: 239-245.

SUGGESTED READINGS

On Information Retrieval: Modern Information Retrieval, by Ricardo Baeza-Yates, Berthier Ribiero-Neto, Addison-Wesley Pub Co; 1st edition, 1999, ISBN: 020139829X.
An excellent survey of the key issues surrounding IR, from algorithms to presentation of IR results. This book contains clear explanations of all major algorithms along with quantitative analyses of the relative effectiveness of each algorithm, including the methodology used to arrive at results.

Data Analysis Tools for DNA Microarray, by Sorin Draghici, Chapman & Hall/CRC, 2003, ISBN: 1584883154.
This text on microarray analysis describes complex data analysis techniques, emphasizing specific data analysis issues characteristic of microarray data.

ON-LINE RESOURCES

MetNet, (http://www.public.iastate.edu/~mash/MetNet/) contains links to the websites for the MetNetDB, FCModeler and PathBinder tools mentioned in this chapter.

PathBinderH (www.plantgenomics.iastate.edu/PathBinderH) is a large database of sentences drawn from MEDLINE containing co-occurring terms from a large dictionary. It allows queries to be qualified by biological taxa. It is provided by the Center for Plant Genomics at Iowa State University.

The Arabidopsis Information Resource, TAIR (http://www.arabidopsis.org) is a central clearinghouse for the model organism, Arabidopsis.

Aracyc (Mueller, Zhang et al., 2003) (http://www.arabidopsis.org/tools/aracyc) AraCyc is a database containing biochemical pathways of Arabidopsis, developed at The Arabidopsis Information Resource. The aim of AraCyc is to represent Arabidopsis metabolism as completely as possible. It presently features more than 170 pathways that include information on compounds, intermediates, cofactors, reactions, genes, proteins, and protein subcellular locations.

KEGG: Kyoto Encyclopedia of Genes and Genomes (http://www.genome.jp/kegg/) KEGG is a comprehensive bioinformatics resource developed by the Kanehisa Laboratory of Kyoto University Bioinformatics Center. It contains information about genes and gene products, chemical compounds and pathway information.

Brenda: (http://www.brenda.uni-koeln.de/) is a repository for enzyme information.

R (http://www.r-project.org) is an Open Source language and environment for statistical computing and graphics. R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering) and graphical techniques, and is highly extensible.

Bioconductor (http://www.bioconductor.org) is an open source and open development software project for the analysis and comprehension of genomic data. The project was started in the Fall of 2001. The Bioconductor core team is based primarily at the Biostatistics Unit of the Dana Farber Cancer Institute at the Harvard Medical School/Harvard School of Public Health. Other members come from various US and international institutions.

PubMed (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi) is a search interface provided by the U.S. National Library of Medicine to a large database of biological texts, mostly but not exclusively from the MEDLINE database.

Agricola (http://agricola.nal.usda.gov/) is a database of article citations and abstracts in the agriculture field, provided by the U.S. National Agricultural Library.

Arrowsmith (kiwi.uchicago.edu) is a system for generating hypotheses about interactions from texts in MEDLINE. (The name is from Sinclair Lewis' novel Martin Arrowsmith.) Provided by the University of Chicago.

MedMiner (http://discover.nci.nih.gov/textmining/main.jsp) is a sentence retrieval system provided by the U.S. National Library of Medicine. It integrates GeneCards and PubMed.

PreBind (http://www.blueprint.org/products/prebind/prebind_about.html) is a database of sentences potentially describing biomolecular interactions. Uncurated, it feeds the Bind database, which is curated. Provided in affiliation with the University of Toronto.

QUESTIONS FOR DISCUSSION

Suppose there is a set of 8 sentences, 4 of which are hits and 4 of which are not. Feature 1 is present in all 4 hits and in 2 non-hits. Feature 2 also occurs in 4 hits and 2 non-hits. There is 1 non-hit with both features. What is the probability estimated by the Naïve Bayes formula that a sentence with both features is a hit? What are the odds for this estimated by the formula for semi-naïve evidence combination? What is the probability implied by these odds? What is the true probability? Repeat this process for the non-hit category. Discuss the results.