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

Abstract:

Key words: 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. We focus 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 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[1] 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 Pathways

Metabolic Pathway Databases There are a few major database projects designed to capture pathways: What Is That? (WIT) Project[2] (http://wit.mcs.anl.gov/WIT2/WIT), Kyoto Encyclopedia of Genes and Genomes (KEGG http://www.genome.ad.jp/kegg)[3], and EcoCyc/MetaCyc (http://ecocyc.DoubleTwist.com/ecocyc/)[4,5]. 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. WIT3, currently in a pre-alpha stage, focuses on metabolic reconstructions from sequence data, however its links are inactive so its current status is unknown. KEGG computerizes current knowledge of molecular and cellular biology in terms of the information pathways that consist of interacting genes or molecules and links individual components of the pathways with the gene catalogs being produced by the genome projects. Also, the drawings of

individual biochemical pathways in both KEGG and WIT/WIT2 are not created dynamically; rather each one is constructed *a priori* and stored in a database. 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[6,7]. The metabolic-pathway database, MetaCyc, describes pathways and enzymes for many different organisms (e.g. *Arabidopsis thaliana*, AraCyc), and combines information from sequences. Our prototype, MetNetDB, combines knowledge from experts, Aracyc and more specialized pathway sequence data, with experimental data from microarrays, proteomics and metabolomics and dynamically displays the results in FCModeler[8,9]. The database is designed to include information about subcellular location, and to handle both enzymatic and regulatory interactions.

Other database designs emphasize data visualization. Cytoscape visualizes existing molecular interaction networks and gene expression profiles and other state data using Java[10]. It has facilities for constructing networks and displaying annotations from fixed files. MetNet-FCModeler operates on stand-alone computers with well-defined XML file formats that allow users to easily import their own data into the model network. FCModeler also works with R[11] to add a generalized modeling framework.. MetNetDB is web-accessible and users can create their own custom-pathways, that can then be uses in analyses of expression data.

2.2 Network Modeling and Reconstruction

2.3 Extracting Biological Interactions from Text Corpora

Introduction. Mining of the biological "literaturome" is an important module in a comprehensive creation, representation, display, and simulation system of metabolic and regulatory networks. Without it, many biomolecular interactions archived in the literature remain accessible in principle but underutilized in practice. 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. Pragmatically, the two are "not necessarily convergent" [12,13], although clearly the intent is that eventually they will be. Competitions to test the performance of automatic annotation such as the BioCreative Workshop [14], 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.

Empirical facts about biological texts. A number of workers have investigated mining of biomolecular interactions from text[15-28]. However, reporting of empirical facts about interaction descriptions remains quite limited. Craven and Kumlien[29] provided a list 20 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.[17] gave a ranked list of 20 words found useful in identifying abstracts describing protein interactions. Results were derived from yeast-related abstracts and therefore may be yeast-specific. Ono et al.[19] 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.[21] proposed four categories of passages using a rule-based scoring strategy, and gave the information retrieval (IR) performance of each category. Sekimizu et al.[30] 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 results similar in spirit to those anticipated from the proposed work. These results concern 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[31]. Ding et al.[32,33] 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, one may expect researchers to focus increasing attention on this important gap.

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 can be rated, or sentences to be ranked for human curation. In the following paragraphs we compare two methods for evidence combination. One is the well-known Naïve Bayes model. The other is 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[34].

Evidence combination with the Naïve Bayes model. This standard method produces probability estimates that can used for categorization[35]. The formula is:

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

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 calculate the desired probability, at the cost of providing an estimated due to the assumption that the features occur independently of one another. A readable derivation is provided by Wikipedia[36].

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 is[34]:

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

where the odds that a sentence describes an *i*nteraction if it has features $f_1,...,f_n$ are $O(h|f_1,...,f_n)$, 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 equation $O(h|f_1,...,f_n)=O_1...O_n/(O_0)^{n-1}$ just given is in terms of odds, which are ratios of hits to misses. Thus, for example, 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

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 [9]. ***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: Visualizing and Modeling Metabolic Networks

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

Application of graph theoretic methods to analyze complex networks of data. 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 will 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 and will help. 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[37,38]. FCModeler also searches for alternate paths between two entities, to help find out all the ways one part of the graph can interact with another. Grouping cycles by similarity metrics may lead to simplifying the display of complex graphs [38-41] which states to what degree each cycle is contained in another. Cycles that are very similar form a "family" of cycles. The difference between the cycles could indicate that two

interacting pathways have the same effect or that there are two mechanisms for control of a process. The common areas among pathways may reflect critical paths in the network X_i.

Network Modeling using Fuzzy Cognitive Maps.

We are working on model validation using pathways developed by expert users [8,42]. We have designed an XML file format that accurately encodes the network topology information for MetNetDB. Automated checking of pathway information using data from expression studies can test the accuracy of the predictions and help determine the most predictive pathway model. The network modeling uses the R computing environment.

3.2.1 Multi-Resolution 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. The 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 [43]:

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

$F = \{X_i,\ i = 1,...,N\}$ 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. One commonly used fuzzy membership function is:

$$m_{ij} = \frac{1/d_{ij}^2}{\sum_{k=1}^{K} 1/d_{ik}^2} \quad (3)$$

Adding a window function $W(d)$ to the membership function limits the size of clusters. The modified membership function is:

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

The window function $W(d)$ is centered at V_j and can take any form. This work uses truncated Gaussian windows with values outside the range of 3σ set to zero:

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

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

3.2.2 Hierarchical Algorithm

The multi-resolution algorithm is similar to the ISODATA algorithm with cluster splitting and merging[44,45]. 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 1. ε_1 and ε_2 are small numbers to determine whether the clustering converged. 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.* HIERARCHICAL FUZZY K-MEANS ALGORITHM

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}.
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 converging to a given threshold ε_2.

3.2.3 Effects of window size

Changing the window size can affect the level of detail captured in the clusters. If $\sigma << 1$, 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 hierarchical fuzzy K-mean clustering ($\sigma = 0.1$, 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 $\sigma = 0.1$, 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 $\sigma = 0.2$, smaller clusters are combined with nearby clusters. Highly correlated profiles are detected. The $\sigma = 0.3$ level is the coarsest level.

3.2.4 Construction of gene 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 as in [46].

The gene regulatory model can be simplified as a linear model[47]:

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

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:

$$R^C_{AB}(\tau) = \sum_k w_k R^k_{AB}(\tau) \tag{8}$$

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 and the distance to cluster center. An additive fuzzy system is used to combine this information [48]. 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

Evidence Code	Meaning of the Evidence Code	Membership Value, w_{evi}
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 II, 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 [48]. 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} :

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

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 PathBinder components of varied design is illustrated by our systems at www.vrac.iastate.edu/~berleant/MedRep and www.plantgenomics.iastate.edu/PathBinderH. For the present system, an integrated PathBinder component, called PathBinderA, 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.

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 et al. 2002). The Table below (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 recalls and precisions 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[49].

Using the mining results, stage 3. While the mining algorithm, stages 1 and 2, extract biomolecular interactions from the literature, this phase is to integrate the extraction capability into the larger MetNetDB system. The integration is designed to provide 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 efficiency support for curation is an important need, both in general and for METNET in particular. PathBinderA is designed to support curation by presenting mined potential interactions to curators starting from the best, most likely interactions. The curators are members of the labs of the project team, continuing but making more efficient a curation process that has enabled constructing the current prototype. Additionally, the design has clickable links from a putative interaction to the bag of sentences relevant to it, from which sentences are presented starting from the ones with the highest likelihood of describing the interaction. Thus, if even one sentence is deemed to describe an interaction between a given pair of biomolecules by a curator, there is no longer any need for the curator to examine other sentences with respect to that pair. This 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 any such system can easily make the threshold adjustable, and should do so. Such likely interactions will thus be 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 of 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.

A key functionality we plan is to allow users to choose *species and other taxa* to view sentences about. This is feasible, as has been shown at www.plantgenomics.iastate.edu/PathBinderH. 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 1).

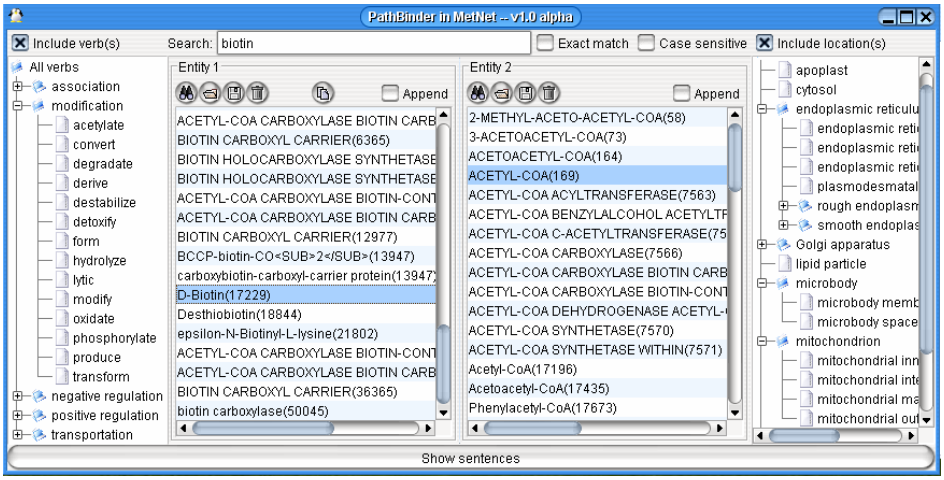

Figure #17-1. PathBinderA interface, showing four choices a user can make to choose two biomolecules, a subcellular location, and an interaction-relevant verb.

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.

The tested data set compared Arabidopsis thaliana plants, wild-type (WT) and transgenic plants containing antisense ACLA-1 behind the constitutive CaMV 35S promoter (referred to as aACLA-1). The microarray type was an Affymetrix GeneChip. 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), and changing from light (from 0 to 8 hours) to dark (from 8 to 20 hours)[50]. 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[51]. Both replicates of each gene expression profile are standardized to zero mean, one standard deviation. The data was filtered by comparing the expression values between the WT and ACLA1 gene mutated at 1, 8 and 12 hours, differentially expressed genes having larger than 2 fold changes at any time point were kept. There

are 484 genes in total after filtering. The gene expression patterns used for clustering are the time point measurements for the wild-type plant.

4.1 Construct the genetic network using time correlation

These relationships between clusters can be found by constructing the regulatory networks based on the cluster center profiles using a correlation threshold of T_R = 0.65. The strength of correlation is mapped into three categories: [0.65, 0.75), [0.75, 0.85), and [0.85, 1]. In figure 2 three types of line thickness from thin to thick to represent the strength of the correlation. Black dashed lines represent positive coregulation; green dashed lines represent negative coregulation; red solid lines with bar head represent negative regulation; blue solid lines with arrowheads represent positive regulation. Figure 2 shows that cluster 1 and 5 are highly coregulated (0 time delay), cluster 1 and 5 positively regulate cluster 4 with time delay 2.5h and 3h, and both negatively regulated cluster 3 with time delay 1.5h; cluster 4 is negatively regulated by cluster 3 with delay 1h, 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.

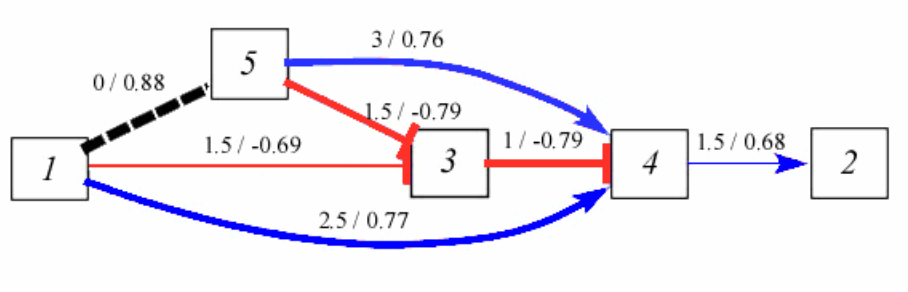

Figure #17-2. Gene regulatory networks inferred from the case with sigma equal to 0.2. The numbers on each link show the time delay for the interaction on top and the correlation coefficient of the interaction on the bottom.

Since the data were unequally sampled with 0.5h as minimum interval, we interpolated the gene expression profiles as equally sampled 41 time points with 0.5h interval. The time correlation of each replicate is computed using equation (7), then combined using equation (8). The time period is limited to the range of [-4h, 4h] because the light period only lasts 8 hours in this data set. Figure 3 shows the constructed regulatory networks of the 28 cluster centers at the σ= 0.1 level. The graph notations are the same as figure 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.

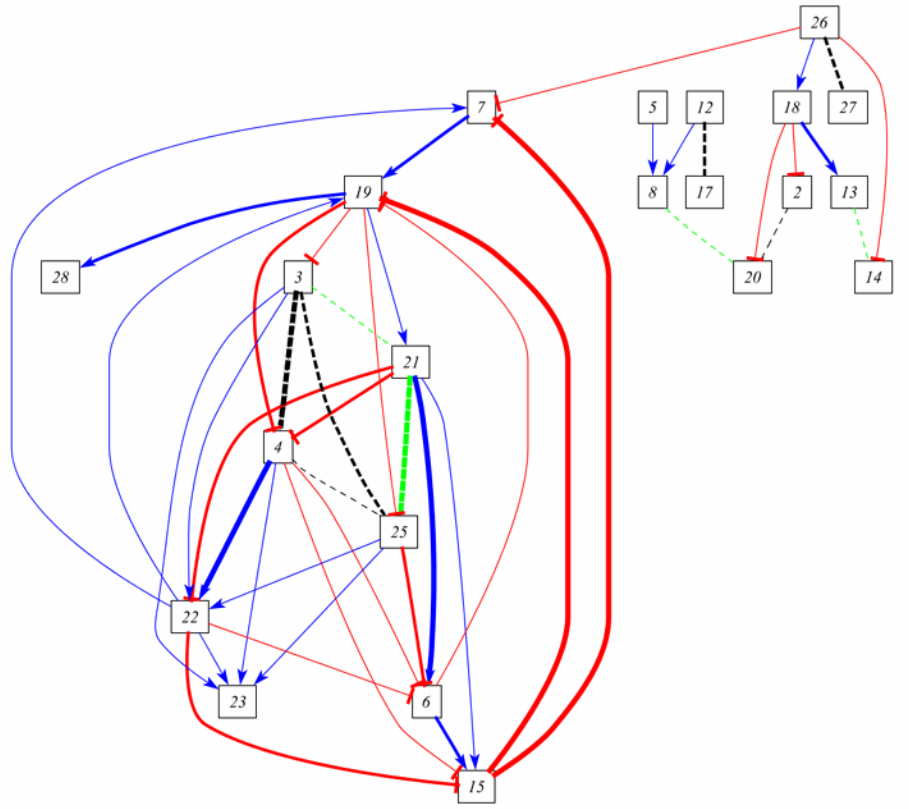

Figure #17-3. Regulatory networks among cluster centers at the window size sigma = 0.1 level.

Figure 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 4 shows part of the simplified graph.

4.2 Cluster and Network Validation

Cluster validation makes use of the available literature accessible through PathBinderA and GO information to find out what kind of functions or processes a cluster involves and to search for potential interactions. In figure 3, the graphs in the upper right corner are less connected. The Gene Ontology shows most of the genes clusters are not annotated. This means these clusters have no biological evidence of direct relation with the highly connected group. It also shows how the fuzzy hierarchical algorithm successfully separates those unrelated genes.

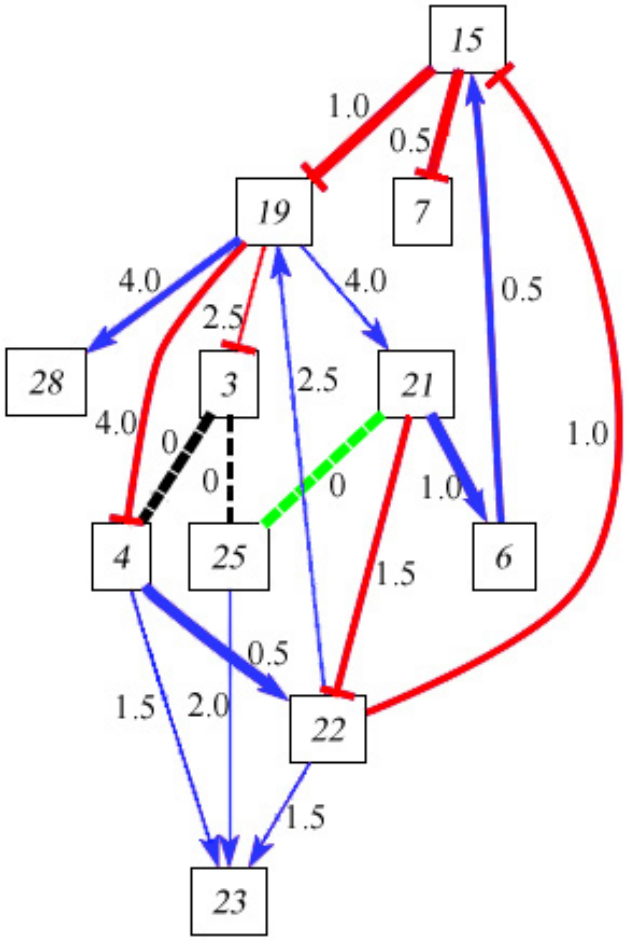

Figure #17-4. Simplified regulatory network with redundant edges removed for the window size $\sigma = 0.1$ level. The number on each link represents the estimated time delay.

Figure 4 shows that cluster 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 III shows the fuzzy weights for the GO terms in each cluster. The BP (Biological Process) GO annotations show that clusters 3 and 4 involve many similar biological processes. For example, both clusters involve "Carboxylic acid metabolism", "Regulation of transcription, DNA-dependent", and "Protein amino acid phosphorylation". Cluster 3 has more emphasis on "Regulation of transcription, DNA-dependent" and cluster 4 emphasizes "Protein amino acid phosphorylation". Also cluster 3 involves "water derivation", but cluster 4 mainly involves another BP "Response to desiccation, hyper osmotic salinity and temperature". Clusters 3 and 4 provide a good example of the overlapping of fuzzy clusters, while the separation of two clusters does make sense.

Clusters 21 and 25 are two highly negatively coregulated clusters. Cluster 21 involves "Photosynthesis, dark reaction" which is active at night, while cluster 25 mainly involves "Carboxylic acid metabolism" and other metabolism usually active in the day. Cluster 21 contains genes for "Trehalose biosynthesis". Trehalose plays a role in the regulation of sugar metabolism, which has just been identified for Arabidopsis[52]. Clusters 6 and 21 involve sugar metabolism (carbohydrate metabolism in GO term). This is a significant biological result for understanding regulation in this experiment.

Figures 3 and 4 show that cluster 19 regulates clusters 3, 4, 21, 22, 25 and 28. After checking the BP GO annotations, we found 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" are two major protein regulation mechanisms, while "Response to auxin stimulus" and "Cell-cell signaling" involve the processes of receiving stimulus or signals from others. Therefore these BP GO annotations match our network structures.

Clusters 23 and 28 have no out-going edges, which implies that they are not involved in regulatory activity. Clusters 3, 4, 6, 7, 15, 19 21 22, and 25 involve one or several of "Regulation of transcription, DNA-dependent", "Protein amino acid phosphorylation" or "N-terminal protein myristoylation" biological processes. The later two are two major protein regulation mechanisms. Also cluster 21 involves Trehalose regulation as shown earlier. The BP annotations for clusters 23 and 28 are "Response to stimulus" and "Carbohydrate metabolism" which are non-regulatory.

Cluster 19 contains the ethylene response gene "ethylene-induced esterase". Cluster 4 contains jasmonic acid response and several jasmonate

biosynthesis genes. The 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 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 6.

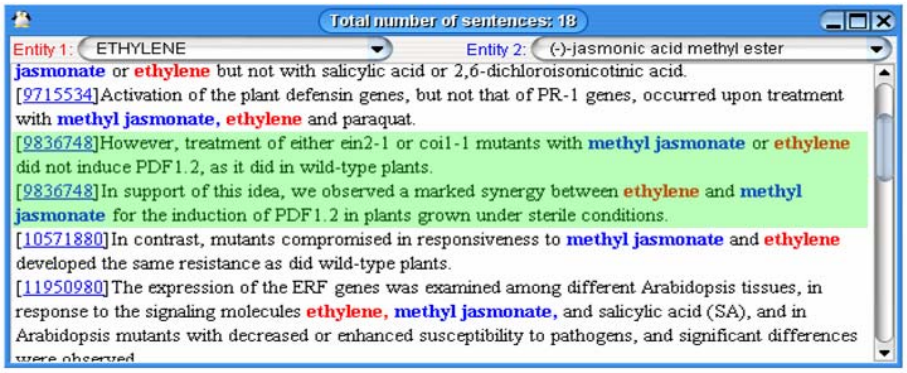

Figure #17-5. PathBinderA output for the terms ethylene and jasomonic acid methyl ester. The relevant sentences and the Medline identification number are given.

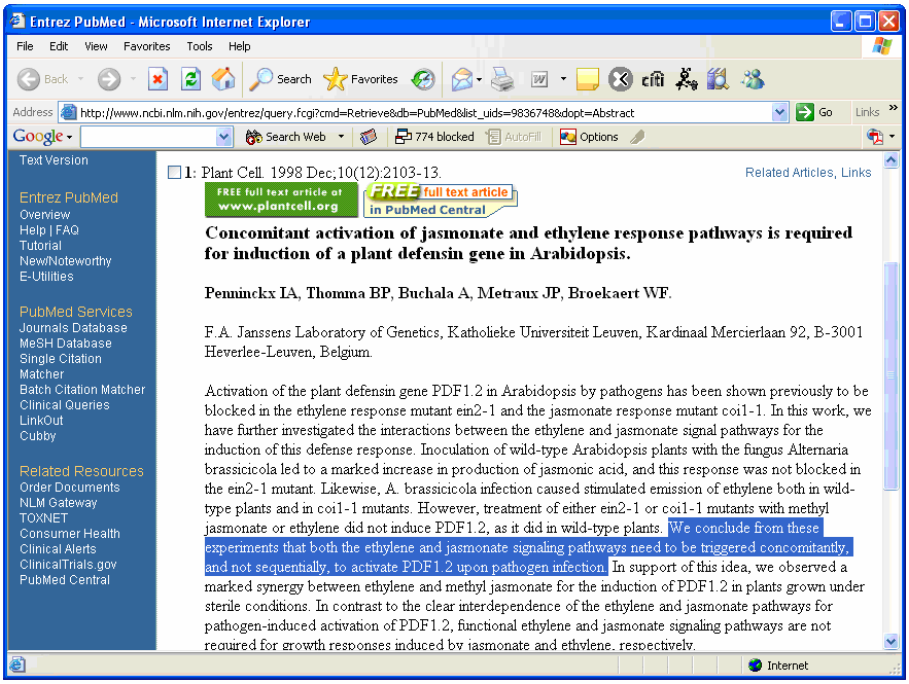

Figure #17-6. The complete abstract for the selection shown above gives more details on the relationship between the ethylene and jasmonate signaling pathways.

5. DISCUSSION

6. ACKNOWLEDGEMENT

The network visualization was performed using the facilities at the Virtual Reality Application Center at Iowa State University. The authors would like to thank Dr. Carol Foster, and Ling Li for kindly making their microarray expression data available for this work.

7. REFERENCES

1. Rhee, S.Y., W. Beavis, T.Z. Berardini, G. Chen, D. Dixon, A. Doyle, M. Garcia-Hernandez, E. Huala, G. Lander, M. Montoya, N. Miller, L.A. Mueller, S.

Mundodi, L. Reiser, J. Tacklind, D.C. Weems, Y. Wu, I. Xu, D. Yoo, J. Yoon, and P. Zhang, *The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community.* Nucl. Acids. Res., 2003. **31**(1): p. 224-228.

2. Overbeek, R., N. Larsen, G.D. Pusch, M. D'Souza, E.S. Jr, N. Kyrpides, M. Fonstein, N. Maltsev, and E. Selkov, *WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction.* Nucl. Acids. Res., 2000. **28**: p. 123-125.
3. Kanehisa, M. and S. Goto, *KEGG: Kyoto Encyclopedia of Genes and Genomes.* Nucleic Acids Research, 2000. **28**(1): p. 27-30.
4. Karp, P.D., M. Riley, M. Saier, I.T. Paulsen, S.M. Paley, and A. Pellegrini-Toole, *The EcoCyc and MetaCyc databases.* Nucleic Acids Research, 2000. **28**(1): p. 56-59.
5. Karp, P.D., M. Riley, S.M. Paley, and A. Pellegrini-Toole, *The MetaCyc Database.* Nucl. Acids. Res., 2002. **30**: p. 59-61.
6. Karp, P.D., M. Krummenacker, S. Paley, and J. Wagg, *Integrated pathway/genome databases and their role in drug discovery.* Trends in Biotechnology, 1999. **17**(7): p. 275-281.
7. Karp, P.D., *Pathway databases: a case study in computational symbolic theories.* Science, 2001. **293**(5537): p. 2040-4.
8. Dickerson, J.A., D. Berleant, Z. Cox, W. Qi, D. Ashlock, E.S. Wurtele, and A.W. Fulmer, *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, Editors. 2003, World Scientific Publishing: Singapore. p. 207-238.
9. Wurtele, E.S., J. Li, L. Diao, H. Zhang, C. Foster, B. Fatland, J.A. Dickerson, A. Brown, Z. Cox, D. Cook, E.-K. Lee, and H. Hofmann, *MetNet: software to build and model the biogenetic lattice of Arabidopsis.* Comparative and Functional Genomics, 2003. **4**: p. 239-245.
10. Shannon, P., A. Markiel, O. Ozier, N. Baliga, J. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, *Cytoscape: a software environment for integrated models of biomolecular interaction networks.* Genome Research, 2003. **13**(11): p. 2498-504.
11. Ihaka, R. and R. Gentleman, *R: A language for data analysis and graphics.* Journal of Computational and Graphical Statistics, 1996. **5**: p. 299-314.
12. Irving, R.W. and M.R. Jerrum, *Three-dimensional statistical data security problems.* SIAM Journal on Computing, 1994. **23**(February): p. 170-184.
13. Krallinger, M., *Biological Information, Information, and Knowledge, BioLINK home page*. (June. 2004), Retrieved from http://www.pdg.cnb.uam.es/BioLINK/.
14. EMBO BioCreative Workshop, *A Critical Assessment for Information Extraction in Biology" (BioCreative), at*. (June. 2004), Retrieved from http://www.pdg.cnb.uam.es/BioLINK/workshop_BioCreative_04/.
15. Blaschke, C., M. Andrade, C. Ouzounis, and A. Valencia. *Automatic Extraction of Biological Information from Scientific Text: Protein-Protein Interactions*. in *International Conference on Intelligent Systems for Molecular Biology*. 1999. Heidelberg.

16. Humphreys, K., G. Demetriou and R. Gaizauskas. *Two Applications of Information Extraction to Biological Science Journal Articles: Enzyme Interactions and Protein Structures*. in *Pacific Symposium on Biocomputing 5*. 2000.
17. Marcotte, E.M., I. Xenarios, and D. Eisenberg, *Mining literature for protein-protein interactions.* Bioinformatics, 2001. **17**(4): p. 359-63.
18. Ng, S.K. and M. Wong, *Toward Routine Automatic Pathway Discovery from On-line Scientific Text Abstracts.* Genome Inform Ser Workshop Genome Inform, 1999. **10**: p. 104-112.
19. Ono, T., H. Hishigaki, A. Tanigami, and T. Takagi, *Automated extraction of information on protein-protein interactions from the biological literature.* Bioinformatics, 2001. **17**(2): p. 155-61.
20. Park, J.C., H.S. Kim, and J.J. Kim, *Bidirectional incremental parsing for automatic pathway identification with combinatory categorial grammar.* Pac Symp Biocomput, 2001: p. 396-407.
21. Thomas, J., D. Milward, C. Ouzounis, S. Pulman, and M. Carroll, *Automatic extraction of protein interactions from scientific abstracts.* Pac Symp Biocomput, 2000: p. 541-52.
22. Wong, L., *PIES, a protein interaction extraction system.* Pac Symp Biocomput, 2001: p. 520-31.
23. Sugiyama, K., K. Hatano, M. Yoshiawa, and S. Uemura, *Extracting information on protein-protein interactions from biological literature based on machine learning approaches.* Genome Informatics, 2003. **14**: p. 699-700.
24. Temkin, J.M. and M.R. Gilder, *Extraction of protein interaction information from unstructured text using a context-free grammar.* Bioinformatics, 2003. **19**(16): p. 2046-53.
25. Daraselia, N., A. Yuryev, S. Egorov, S. Novichkova, A. Nikitin, and I. Mazo, *Extracting human protein interactions from MEDLINE using a full-sentence parser.* Bioinformatics, 2004. **20**(5): p. 604-11.
26. Wren, J.D. and H.R. Garner, *Shared relationship analysis: ranking set cohesion and commonalities within a literature-derived relationship network.* Bioinformatics, 2004. **20**(2): p. 191-8.
27. Wren, J.D., R. Bekeredjian, J.A. Stewart, R.V. Shohet, and H.R. Garner, *Knowledge discovery by automated identification and ranking of implicit relationships.* Bioinformatics, 2004. **20**(3): p. 389-98.
28. Donaldson, I., J. Martin, B.d. Bruijn, C. Wolting, B.T. V. Lay, B.B. S. Zhang, G.D. Bader, K. Michalickova, T. Pawson, and C.W.V. Hogue, *PreBIND and Textomy - mining the biomedical literature for protein-protein interactions using a support vector machine.* BMC Bioinformatics, 2003. **4**(11).
29. Craven, M. and J. Kumlien, *Constructing biological knowledge bases by extracting information from text sources.* Proc Int Conf Intell Syst Mol Biol, 1999: p. 77-86.
30. Sekimizu, T., H.S. Park, and J. Tsujii, *Identifying the Interaction between Genes and Gene Products Based on Frequently Seen Verbs in Medline Abstracts.* Genome Inform Ser Workshop Genome Inform, 1998. **9**: p. 62-71.
31. Ding, J., D. Berleant, D. Nettleton, and E. Wurtele. *Mining MEDLINE: Abstracts, Sentences, or Phrases?* in *Pacific Symposium on Biocomputing (PSB 2002)*. 2002. Kaua'i, Hawaii.

32. Ding, J., D. Berleant, J. Xu, and A.W. Fulmer. *Extracting biochemical interactions from MEDLINE using a link grammar parser*. in *Proceedings of the Fifteenth IEEE Conference on Tools with Artificial Intelligence (ICTAI 2003)*. 2003. Sacramento, CA, USA.
33. Ding, J., *PathBinder: a sentence repository of biochemical interactions extracted from MEDLINE*, in *Dept. of Electrical and Computer Engineering*. 2003, Iowa State University: Ames, IA.
34. Berleant, D., *Combining Evidence: the Naïve Bayes Model Vs. Semi-Naïve Evidence Combination*. 2004, Software Artifact Research and Development Laboratory, Iowa State University: Ames, IA.
35. Lewis, D. *Naïve Bayes at forty: the independence assumption in information retrieval*. in *Conf. Proc. European Conference on Machine Learning*. 1998. Chemnitz, Germany.
36. Wikipedia: the free encyclopedia, *Naive Bayesian classification*. (August 01. 2004), Retrieved from http://en.wikipedia.org/wiki/Naive_Bayesian_classification.
37. Cox, Z., A. Fulmer, and J.A. Dickerson. *Interactive Graphs for Exploring Metabolic Pathways*. in *ISMB, 2002*. 2002. Edmonton, CA.
38. Dickerson, J.A. and Z. Cox. *Using Fuzzy Measures to Group Cycles in Metabolic Networks*. in *North American Fuzzy Information Processing Society (NAFIPS) Annual Meeting*. 2003. Chicago, IL.
39. Bunke, H. and J. Csirik, *Parametric string edit distance and its application to pattern recognition.* IEEE Trans. Syst., Man, Cybern., 1993. **25**(1): p. 202-206.
40. Bunke, H., *On a relation between graph edit distance and maximum common subgraph.* Pattern Recognition Letters, 1997. **18**(8): p. 689-694.
41. Duda, R.O. and P.E. Hart, *Pattern Classification and Scene Analysis*. 1973, New York: John Wiley & Sons.
42. Dickerson, J.A., Z. Cox, E.S. Wurtele, and A.W. Fulmer. *Creating Metabolic and Regulatory Network Models using Fuzzy Cognitive Maps*. in *North American Fuzzy Information Processing Conference (NAFIPS)*. 2001. Vancouver, B.C.
43. Bezdek, J.C., *Pattern Recognition with Fuzzy Objective Function Algorithms*. Advanced Applications in Pattern Recognition, ed. M. Nadler. 1981, New York: Plenum Press.
44. Ball, G.H., *Data analysis in the social sciences: what about the details.* AFIPS Proc. Cong. Fall Joint Comp., 1965. **27**(1): p. 533-559.
45. Ball, G.H. and D.J. Hall, *ISODATA, a novel method of data analysis and pattern classification*. 1965, Stanford Research Institute.
46. D'Haeseleer, P., *Reconstructing Gene Networks from Large Scale Gene Expression Data*, in *Computer Science*. 2000, The University of New Mexico: Albuquerque, NM. p. 207.
47. D'Haeseleer, P., S. Liang, and R. Somogyi, *Gene expression analysis and modeling.* Pac Symp Biocomput, 1999(Tutorial).
48. Kosko, B., *Neural Networks and Fuzzy Systems*. 1992, Englewood Cliffs: Prentice Hall.
49. Skounakis, M. and M. Craven. *Evidence combination in biomedical natural-language processing, , .* in *3rd ACM SIGKDD Workshop on Data Mining in Bioinformatics*. 2003.

50. Foster, C.M., L. Ling, A.M. Myers, M.G. James, B.J. Nikolau, and E.S. Wurtele, *Expression of genes in the starch metabolic network of Arabidopsis during starch sythesis and degradation.* In Preparation, 2003.
51. Gautier, L., L. Cope, B. Bolstad, and R. Irizarry, *affy--analysis of Affymetrix GeneChip data at the probe level.* Bioinformatics, 2004. **20**(3): p. 307-315.
52. Eastmond, P.J. and I.A. Graham, *Trehalose metabolism: a regulatory role for trehalose-6-phosphate?* Curr Opin Plant Biol, 2003. **6**(3): p. 231-5.
53. Mueller, L.A., P. Zhang, and S.Y. Rhee, *AraCyc: A Biochemical Pathway Database for Arabidopsis.* Plant Physiol., 2003. **132**(2): p. 453–460.

8. SUGGESTED READINGS

On Information Retrieval: Modern Information Retrieval, by Ricardo Baeza-Yates, Berthier Ribiero-Neto, Berthier Ribeiro-Neto, Addison-Wesley Pub Co; 1st edition (May 15, 1999), ISBN: 020139829X.

9. 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 (www.arabidopsis.org) is a central clearinghouse for the model organism Arabidopsis. It contains extensive gene information???

Aracyc[53] (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 (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 (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.

10. QUESTIONS FOR DISCUSSION

Problem 1. 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.