

Fuzzy Modeling of Metabolic Maps in MetNetDB*

Julie A. Dickerson and Pan Du

*Department of Electrical and Computer Engineering
Iowa State University
Ames, IA, 50011-3060, USA*

julied@iastate.edu; dupan@iastate.edu

Eve Wurtele and Jie Li

*Department of Botany
Iowa State University
Ames, IA, 50011, USA*

mash@iastate.edu; jieli@iastate.edu

Abstract - Metabolic networks combine metabolism and regulation. These complex networks are difficult to understand and visualize due to the diverse types of information that need to be represented. FCModeler, a publicly available software package is designed to enable the biologist to visualize and model metabolic and regulatory network maps in plants. It links to an interactions database (MetNetDB) containing information on regulatory and metabolic interactions derived from a combination of web databases and input from biologists in their area of expertise. FCModeler evaluates hypotheses and provides a graph-theoretic modeling framework for assessing the large amounts of data captured by high-throughput gene expression experiments. This leads to a new method for classifying gene expression data by combining network, gene ontology, and measured expression information in a fuzzy clustering framework.

I. INTRODUCTION

A major challenge in the post-genome era is to understand how interactions among molecules in a cell determine its form and function. With the help of transcriptomic, proteomic and metabolomic analysis technologies, biologists can obtain vast amounts of valuable data on metabolic network interactions, and many approaches are being developed to analyze the resultant data.

Three basic types of interactions are conversion, regulatory, and catalytic. In a conversion interaction, a node (typically a chemical(s)) is converted into another node, and used up in the process. A catalytic interaction represents an enzyme that enables a chemical conversion and does not get used up in the process. In a regulatory interaction, the entity activates or deactivates another node, and is not used up in the process.

The interactions (also referred to as edges or links) in the network are modeled as fuzzy functions depending on the detail known about the network. Modeling using fuzzy cognitive maps (FCMs) is performed in either R or Matlab™ analysis program and the results showing node activation levels are animated in FCModeler. Fuzzy cognitive maps are fuzzy digraphs that model causal flow between concepts [1] or, in this case, biomolecular entities [2, 3]. Entities stand for causal fuzzy sets where events occur to some degree. The entities are linked by interactions that show the degree to which these entities depend on each other. Interactions stand for causal flow. The sign of an interaction (+ or -) shows

causal increase or decrease between entities. The fuzzy structure allows the RNA, metabolite, or protein levels to be expressed as continuous values. This modeling has demonstrated regulation in the Arabidopsis network, in the case of gibberellin conversion from an inactive form to an active form[3].

Gene expression data in form of microarray data measures the amount of RNA associated with a particular gene present in a sample. Fuzzy clustering methods that combine information from metabolic networks can help test hypotheses on how genes regulate one another.

II. MODELING METABOLISM

A. Metabolic Networking Database

The Metabolic Networking DataBase (MetNetDB) contains a metabolic and regulatory map of Arabidopsis with a user-friendly JAVA interface for creating and searching the map. The map, together with gene expression data (metabolomics, proteomics, and microarray), can be transferred to FCModeler as an XML file, for use in data exploration.

The MetNetDB map is being assembled by biologists with expertise in specific areas of metabolism. It is composed of entities (genes, RNAs, polypeptides, protein complexes, metabolites, and environmental inputs) connected by interactions (conversion, catalytic, regulatory). Identities of the genes, RNAs, and polypeptides have been downloaded from TAIR (<http://www.arabidopsis.org/>). Protein complexes are currently added by expert users, as there is no adequate database of protein complexes in Arabidopsis. Identities of many metabolites have been downloaded from KEGG (<http://www.genome.ad.jp/kegg/>); metabolites not present in KEGG are manually added as new entities by expert users, based on their CAS registry number (<http://www.cas.org/EO/regsys.html>).

The metabolic reactions from the AraCyc database have been downloaded into MetNetDB. An important aspect of the map is the inclusion of information on subcellular location. This is critical, because particular entities can interact contingent on being located in the same subcellular compartment. A given entity may be present as separate pools in multiple compartments, for example citrate is present in the mitochondria (where it participates in the TCA cycle) and the

* Funding for this project was provided by grants from the National Science Foundation in the Arabidopsis 2010 (DBI-0209809) and Information Technology Research (IBN-0219366) Programs. Seed funding was also provided by the Iowa State University Plant Sciences Institute and the Roy J. Carver Foundation.

cytosol (where it is a substrate for cytosolic acetyl-CoA formation [4]).

B. Modeling Metabolic Relationships

Three basic types of interactions are conversion, regulatory, and catalytic. In a conversion interaction, a node (typically a chemical(s)) is converted into another node, and used up in the process. A catalytic interaction represents an enzyme that enables a chemical conversion and does not get used up in the process. In a regulatory interaction, the entity activates or deactivates another node, and is not used up in the process.

A wide variety of cellular processes can be represented, each occurring to entities in specified subcellular compartments. For example, to represent the reaction catalyzed by ATP citrate lyase (ACL), that generates cytosolic acetyl-CoA [4], two interactions are used. One is a conversion interaction; its inputs are $\text{citrate}_{\text{cytosol}} + \text{CoA}_{\text{cytosol}} + \text{ATP}_{\text{cytosol}}$ and its outputs are $\text{acetyl-CoA}_{\text{cytosol}} + \text{oxaloacetic acid}_{\text{cytosol}} + \text{ADP}_{\text{cytosol}} + \text{P0}_{\text{cytosol}}$. The second is a catalytic interaction; its input is ATP citrate lyase_{cytosol}. In another example, to represent the translocation of citrate from the mitochondrion to the cytosol, two entities and a single conversion interaction are used: $\text{citrate}_{\text{mitochondrion}}$ goes to $\text{citrate}_{\text{cytosol}}$. The formation or modification of a protein complex can be represented. For example, ACLA and ACLB are the subunits that compose the enzyme ACL. A single conversion interaction is used to represent the reaction; its inputs are $\text{ACLA}_{\text{cytosol}}$, and $\text{ACLB}_{\text{cytosol}}$. Its output is $\text{ACL}_{\text{cytosol}}$.

C. Graphing the Metabolic Map: FCModeler

The main goals of the FCModeler package are to capture the intuitions of biologists and provide a modeling framework for assessing large amounts of information and to test the effects of hypotheses. The tools that are being developed use graph theoretic approaches to analyze network structure and behavior and fuzzy methods that model changes in the network [2]. There are three parts of this system: a dynamic graph visualization package written in Java, graph-theoretic analysis to find critical paths, and modeling using fuzzy cognitive maps to capture uncertainty in the model. Figure 1 shows a sample sub-graph from the MetNetDB for OAA metabolism in Arabidopsis and highlights some of the visualization flexibility available in FCModeler.

III. MODELING METABOLIC NETWORKS USING FUZZY COGNITIVE MAPS

A. Metabolic Networking Database

The interactions (also referred to as edges or links) in the network are modeled as fuzzy functions depending on the detail known about the network. Modeling using fuzzy

cognitive maps (FCMs) is performed in the Matlab™ analysis program and the results showing node activation levels are animated in FCModeler. Fuzzy cognitive maps are fuzzy digraphs that model causal flow between concepts [1] or, in this case, biomolecular entities [2, 3]. Entities stand for causal fuzzy sets where events occur to some degree. The entities are linked by interactions that show the degree to which these entities depend on each other. Interactions stand for causal flow. The sign of an interaction (+ or -) shows causal increase or decrease between entities. The fuzzy structure allows the RNA, metabolite, or protein levels to be expressed as continuous values. This modeling has demonstrated regulation in the Arabidopsis network, in the case of gibberellin conversion from an inactive form to an active form[3].

Fuzzy cognitive maps (FCMs) have the potential to answer many of the concerns that arise from the existing models. Fuzzy logic allows a concept or gene expression to occur to a degree—it does not have to be either on or off [5]. FCMs have been successfully applied to systems that have uncertain and incomplete models that cannot be expressed compactly or conveniently in equations. Some examples are modeling human psychology [6], and on-line fault diagnosis at power plants [7]. All of these problems have some common features. The first is the lack of quantitative information on how different variables interact. The second is that the direction of causality is at least partly known and can be articulated by a domain expert. The third is that they link concepts from different domains together using arrows of causality. These features are shared by the problem of modeling the signal transduction and gene regulatory networks.

Simple or trivalent FCMs have causal edge weights in the set $\{-1,0,1\}$ and concept values in $\{0,1\}$ or $\{-1,1\}$. Simple FCMs give a quick approximation to an expert's causal knowledge. More detailed graphs can replace this link with a time-dependent and/or nonlinear function. The types of link models used in the current project are described below.

Regulatory Links: The regulatory edges are modeled using a simple FCM model that assumes binary connecting edges for the single edge case. When there are multiple excitatory or inhibitory connections, the weights are divided by the number of input connections in the absence of other information. As more information becomes known about details of the regulation, for example how RNA level affects the translation of the corresponding protein, the function of the link models will be updated. The regulatory nodes will also have self-feedback since the nodes stay on until they have been inhibited.

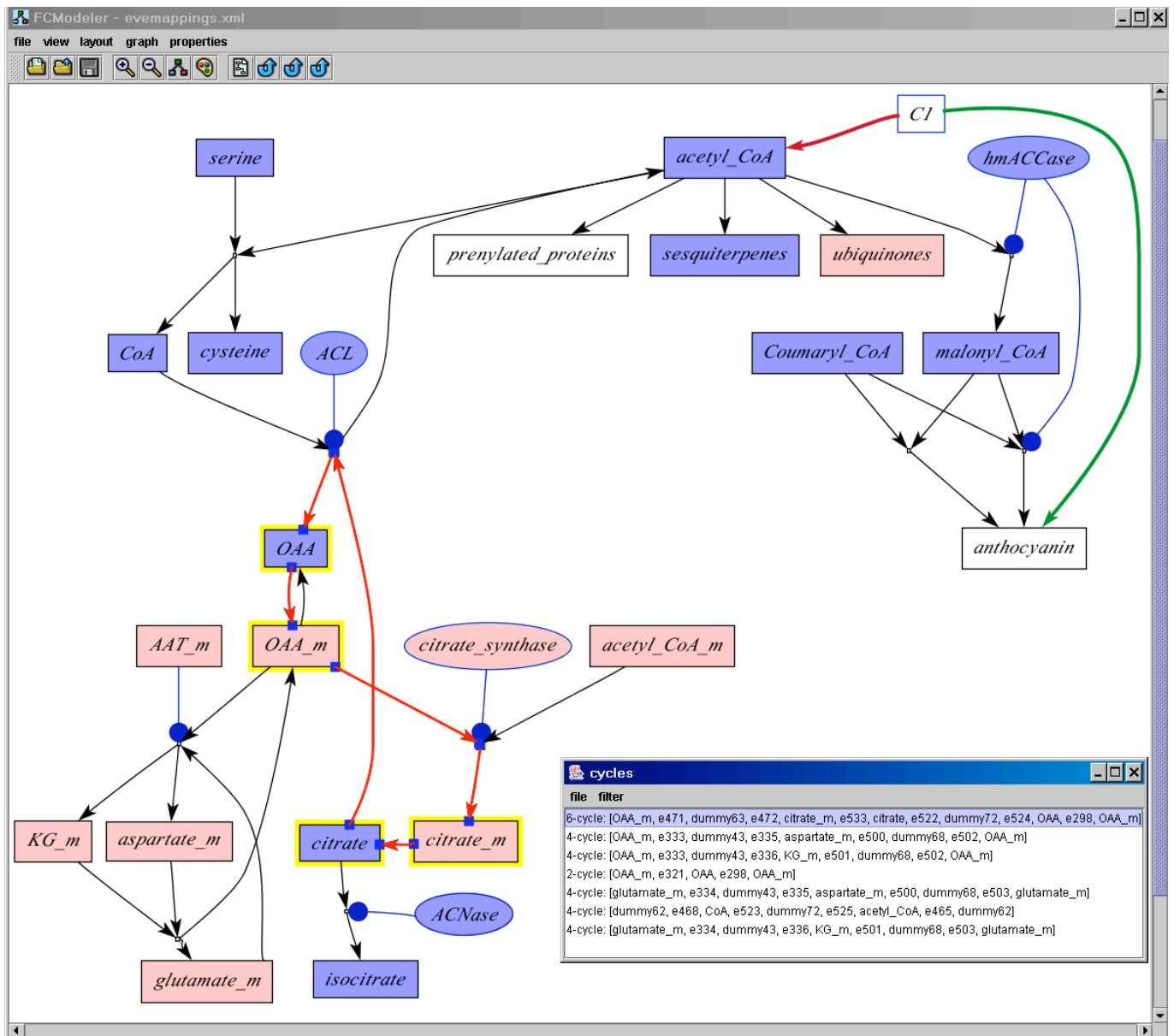


Figure 1. The highlighted nodes and links show a small cycle within the metabolic network. The textbox gives a list of all cycles found in the displayed graph. Colors and shapes of features can be user-designated: entities in mitochondria and cytosol are shaded; entities in unknown location, white; enzymes are shaped as ellipses.

Conversion Links: Conversion relationships are modeled in different ways depending on the goal of the simulation study. The first case corresponds to investigating causal relationships between nodes. The node is modeled in the same manner as a regulatory link in which the presence of one node causes presence at the next node. When information about the rate of change in a reaction is available, a simple difference equation can model the gradually rising and falling levels of the nodes.

Catalyzed Links: Catalyzed reactions add a dummy node that acts upon a conversion link. This allows one link to modify another link. In the current model, the catalyzed link is simulated by weighting the input node in such a way that both inputs must be present for the node to be active. Another

method of modeling catalyzed links is an augmented matrix that operates on the edges between the nodes. The catalyst node acts as a switch that allows a reaction to occur in the proper substrates are available.

B. Modeling using Fuzzy Cognitive Maps

Hypotheses about the metabolic networks can be tested using gene expression data from microarray experiments. Gene expression data measures gene activity. The standard working hypothesis is that genes which behave similarly are related and are in the same metabolic pathway [8]. This hypothesis is used to infer the function of unknown genes. The main problem with this analysis is that large numbers of genes can be selected with a high rate of false detections.

Correlating the microarray data to the extant knowledge base on metabolism allows researchers to validate that known genes are behaving consistently and adds another element into the modeling. Figure 2 shows the results for the biotin pathway in Arabidopsis. Expression of genes in the starch metabolic network of Arabidopsis during starch synthesis and degradation {Foster, 2003 #171}. The genes behave very differently

IV. FUZZY CLUSTERING

A. Clustering Gene Expression Data

In biological networks, many genes are co-regulated. A single gene may participate in different biological processes in different situations. A gene expression profile may behave similarly to several groups of genes. Hard clustering algorithms, e.g., hierarchical clustering or k-means clustering [9], give clusters will in which one gene can only belong to one cluster. These algorithms cannot extract the gene relations described above. Fuzzy C-means uses membership values to measure the relationship between a gene and its clusters [10]. As a result, a gene can belong to several clusters to a degree.

In gene expression data, large clusters tend to occur. As we more interested in finding clusters whose elements are highly co-regulated above a correlation threshold. Another problem of k-means or fuzzy k-means is that they have no guarantee of detecting outliers. Actually, these outliers often have biology significance. Here, we propose an improved fuzzy clustering algorithm to solve these problems.

B. Fuzzy C-Means Clustering with Thresholds

The Fuzzy C-means algorithm minimizes the objective function:

$$J = \sum_{i=1}^n \sum_{j=1}^k u_{ij}^m \|x_i - V_j\|^2$$

represents the data, represents the cluster centers. is the membership of to , and is the distance between and V_j . One commonly used fuzzy membership function is:

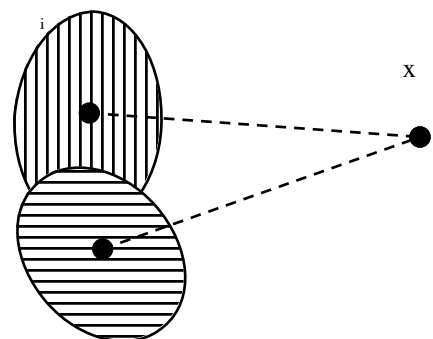
$$u_{ij} = \frac{1}{\sum_{k=1}^k \left(\frac{\|x_i - V_j\|}{\|x_i - V_k\|} \right)^{\frac{2}{m-1}}}$$

Adding a window function to the membership function limits the size of clusters. The modified membership function is:

$$u_{ij} = \frac{1}{\sum_{k=1}^k \left(\frac{\|x_i - V_j\|}{\|x_i - V_k\|} \right)^{\frac{2}{m-1}} + \frac{\|x_i - V_j\|^2}{\sigma^2}}$$

The window function needs to be centered at V_i and can take any form. This work uses truncated Gaussian windows with values outside the range of set to zero. By adding , genes with distances larger than will have no

effect on the cluster centers. The membership function without may result in genes far away from the cluster centers have high membership value. As shown in figure 2, the gene x is far away from the two clusters, but it has 0.5 membership values to both clusters.



C. Algorithm

The proposed algorithm is similar to the ISODATA algorithm with cluster splitting and merging[11, 12]. There are four parameters: k (initial cluster number), (scale of the window), T_{split} (split threshold), (combine threshold). Whenever there are genes further away from the cluster center than , the cluster is split and faraway genes become new cluster centers. Also, if two cluster centers are very close to each other (the distance between them is less than), combine these two clusters. Usually, is less than and . The algorithm is:

1. Initialize parameters: k , , and ;
2. Iterate using Fuzzy C-means until convergence to a given threshold ;
3. Split process: do split if there are elements farther away from cluster center than ;
4. Iterate using Fuzzy C-means until convergence to a given threshold ;
5. Combine Process: combine the clusters whose distance between cluster centers is less than $T_{combine}$. If the clusters after combining have elements far away from cluster center (distance larger than), stop combining.
6. Iterate until converging to a given threshold .

and are small numbers to determine whether the clustering converged. In step 5, we select as a criteria because the genes beyond have no effect to the cluster centers. If one cluster have elements far away from cluster center (distance larger than), this cluster should be split.

V. CONCLUSIONS

The FCModeler software is designed with a focus on understanding the complex molecular network in the model plant eukaryotic species, Arabidopsis. FCModeler enables biologists to capture relationships at different levels of detail, to integrate gene expression data, and to model these relationships. Because of an absence of knowledge about many biological interactions, the software is designed to model at many levels of detail.

ACKNOWLEDGMENT

We thank Lucas Mueller and TAIR for helpful advice and for making the AraCyc pathways publicly available. Thank you to our Metabolic Networking project collaborators: Dr. Dianne Cook, and Dr. Carolina Cruz-Neira. Thanks are also due to the dedicated group of students supporting the metabolic modeling efforts at Iowa State University: Yuting Yang, Joset Etzet, Adam Tomjack, Andres Reinot, and Paul Jennings.

REFERENCES

- [1] J. A. Dickerson and B. Kosko, "Virtual Worlds as Fuzzy Cognitive Maps," *Presence*, vol. 3, pp. 173-189, 1994.
- [2] J. A. Dickerson, D. Berleant, Z. Cox, D. Ashlock, A. W. Fulmer, and E. S. Wurtele, "Creating and Modeling Metabolic and Regulatory Networks Using Text Mining and Fuzzy Expert Systems," in *Computational Biology and Genome Informatics*, C. H. Wu, P. Wang, and J. T. L. Wang, Eds. Hong Kong: World Scientific, 2002.
- [3] J. A. Dickerson, Z. Cox, E. S. Wurtele, and A. W. Fulmer, "Creating Metabolic and Regulatory Network Models using Fuzzy Cognitive Maps," presented at North American Fuzzy Information Processing Conference (NAFIPS), Vancouver, B.C., 2001.
- [4] B. F. Fatland, J. Ke, M. Anderson, W. Mentzen, L. W. Cui, C. Allred, J. L. Johnston, B. J. Nikolau, and E. S. Wurtele, "Molecular Characterization of a Novel Heteromeric ATP-Citrate Lyase that Generates Cytosolic Acetyl-CoA in Arabidopsis," *Plant Physiology*, vol. 130, pp. 740-756, 2002.
- [5] B. Kosko, "Fuzzy Cognitive Maps," *International Journal Man-Machine Studies*, vol. 24, pp. 65-75, 1986.
- [6] M. Hagiwara, "Extended Fuzzy Cognitive Maps," presented at 92 IEEE Int Conf Fuzzy Syst FUZZ-IEEE, San Diego, 1992.
- [7] K. Lee, S. Kim, and M. Sakawa, "On-line fault diagnosis by using fuzzy cognitive maps," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E79-A., pp. 921-922, 1996.
- [8] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proceedings National Academy of Science*, vol. 95, pp. 14863-14868, 1998.
- [9] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: John Wiley & Sons, 1973.
- [10] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. New York: Plenum Press, 1981.
- [11] G. H. Ball, "Data analysis in the social sciences: what about the details," *AFIPS Proc. Cong. Fall Joint Comp.*, vol. 27, pp. 533-559, 1965.
- [12] G. H. Ball and D. J. Hall, "ISODATA, a novel method of data analysis and pattern classification," Stanford Research Institute, Technical Report 1965.