Knowledge Representation for Systems Biology (*)

Su-Shing Chen
Department of Computer Engineering & Computer Science
University of Missouri-Columbia
Columbia, MO 65211
ChenS@missouri.edu
573-882-5176 (Tel); 573-882-8318 (Fax)

Keywords: Systematic Biology, Gene Function, Metabolic Pathway, Knowledge Representation, Cognitive Map, Hierarchical Cognitive Map

Abstract

The system-level understanding of various biological behaviors and phenomena requires several components, such as: gene sequences, protein structures, gene functions and metabolic pathways. A challenging problem is representing, learning and reasoning about these biochemical reactions, gene and protein structures, relationships between genotypes and phenotypes, and their interplay. Building such knowledge bases often integrates various different kinds of knowledge into a single hierarchical framework. On one hand, the knowledge of metabolic pathways consists of kinetic computation, graphical representation, and database. On the other hand, the functionality of genomes includes QTL mappings and higher-level data mining. This paper describes a hierarchical model of cognitive maps for representing gene and metabolic knowledge as well as genotype to phenotype mappings. Cognitive maps are bi-directional graphs that can learn and reason quantitatively and qualitatively. An example for maize hybrids resistant to maize earworm in an agri-ecosystem of the biosphere illustrates a hierarchical cognitive map of biological mappings and biochemical reactions.

1. Introduction

Systems biology is concerned with system-level understanding of biological systems, such as cells and organisms. A general survey of systems biology has included the structures (components and relationships), behaviors, control schemes, and design methods [6]. However the significant heterogeneity and complexity of biological systems prevents us from solving completely the problems through traditional computational methodologies. Our incremental approach is approximating and integrating various known biological component subsystems into a unified framework An appropriate knowledge representation scheme for this purpose is the hierarchical cognitive map [1]. We will demonstrate this scheme on complex genotype to phenotype mappings and cascading enzyme-catalyzed reactions. Other biological system issues can be similarly addressed. The associated learning and reasoning capabilities of cognitive maps further enhance knowledge discovery about systems biology.

(*) This research is partially supported by the NSF Plant Genome Initiative.

In functional genomics, genotype-to-phenotype mappings are complex relationships between markers and traits in chromosome regions. We represent gene markers and traits by nodes and their relationships by directed links of cognitive maps. We implement on these cognitive maps other data mining algorithms (e.g., neural network and relaxation labeling in section 6) needed to relate various trait expressions and phenotypes. In living cells, metabolism (enzyme-catalyzed reactions) is a highly coordinated and optimally controlled process of cascading (or hierarchical) activities. One important question is representing them by graphical methods. In [7], graph theoretical analysis was given for metabolic regulation. In [14], Petri nets were used to represent metabolic pathways. However both methods do not support quantitative and qualitative learning and reasoning well. We demonstrate that cognitive maps employing graph theoretical and logical reasoning analysis can represent metabolic pathways. In this paper, we develop further hierarchical cognitive maps to represent both genotype-to-phenotype mappings and metabolic pathways, providing an incremental approach to systems biology. Moreover, cognitive maps are extended to neural network, probabilistic or fuzzy cognitive maps so that metabolite concentrations become node values and enzyme signal concentrations become link weights of such a map [8,16,17].

2. Functional Genomics

In functional genomics, a basic problem is constructing genetic linkage maps from phenotypic variations at multiple genetic loci. Since genotype can not be determined uniquely from phenotype, various maximum likelihood and regression methods have been used to determine the recombination fraction between a pair of genetic loci. The individual loci controlling a quantitative trait are referred to as quantitative trait loci (QTLs). The problem is stated simply as follows: Given m genetic loci, M_1, ..., M_m, in chromosomal order, and phenotype information P_1, ..., P_n about members of several pedigrees, construct the optimal genetic map or find the maximizing recombination fractions θ_1, ..., θ_{m-1} between these genetic loci. This is a multidimensional search problem [10]. Recently, new technology in molecular biology and computation is providing complete sequences of genomes and expression patterns of their genes. We are able to retrieve genes with specific expression patterns and discover target drugs and gene products. This basic problem has become one of the most important problems.

Our general problem of functional genomics is stated as finding cognitive maps with nodes M_1, ..., M_m and P_1, ..., P_n. The relationships connecting various nodes are functions to be determined. As the basic functional genomics problem is formulated a multidimensional search problem, this general problem is also. Its representation is a cognitive map, perhaps even a hierarchical cognitive map. A computational approach to multidimensional search problem is through learning and reasoning. Our learning and reasoning methods are described in later sections. Recently, functional genomics is taking a new direction of higher ordered structures of organisms and behaviors, leading to the concept of "information flow in biological systems, in particular from genotype to phenotype." [15]. There, the central dogma is the mapping (e.g., cognitive maps) that determines how information in one domain (e.g., genotype) is transformed to information in another domain (e.g., phenotype). The thrust of this approach is captured (with small modification) in the extended genetic network of [15]. This coding theoretic approach fits very well the concept of hierarchical cognitive maps. Due to the hierarchical levels of proximal network, intracellular signaling, and intercellular signaling, the information flow in genetic networks is a hierarchical model of cognitive maps.

3. Metabolic Pathways

In biochemistry, enzyme cascades may transduce metabolic pathways of several levels. They must represent hierarchical cognitive maps. The reason that it must be hierarchical is simply illustrated by the following diagram where several layers of biological mappings and biochemical reactions are present for maize hybrids resistant to maize (corn) earworm in an agri-ecosystem of the biosphere. This summarizing result is contributed by Deboo, Albertsen, Taylor, Marrs, Alfenito, Lloyd, Walbot, Styles, Ceska and others [Byrne et al, 1996]. Higher level nodes A, B, C represent cognitive maps, which are biochemical reactions themselves on lower level maps.

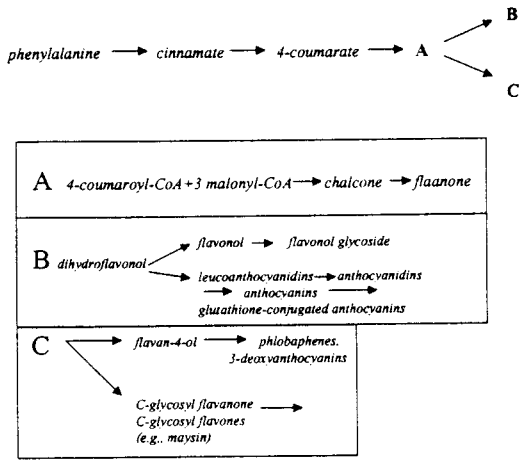

Figure 1. The Maize Phenylpropanoid / Flavonoid Pathway [Byrne et al, 1996]

In these maps, links represent the metabolic reactions between metabolite nodes. The node values are updated from some initial input nodes representing certain substrates. While metabolism is represented as cognitive maps, compartments of metabolic pathways will become hierarchical cognitive maps. Metabolism is an optimally controlled (or regulated) feedback process. Any kinetic reaction between two metabolites A and B can be simulated for metabolite concentrations and reaction fluxes which reflect the node values and link weights of a neural network, probabilistic, or fuzzy cognitive map [12,13]

4. Hierarchical Cognitive Maps

Cognitive maps are directed graphs representing relations (by links) among concepts/attributes (by nodes). Cognitive maps include several knowledge representation schemes. Semantic networks or frames form a special class of cognitive maps. Inference networks and causal networks form other classes of cognitive maps. In cognitive maps, link weights may be assigned to relations representing their compatibility degrees, and node values may be assigned to concepts and attributes representing relevance factors.

A hierarchical cognitive map consists of several cognitive maps, each of which represents gene network interaction or metabolic pathway. The knowledge bases of hierarchical cognitive maps will effectively

capture the complex behavior of biological systems. A hierarchical cognitive map is alternatively represented as a large cognitive map combining several individual ones in the following diagram:

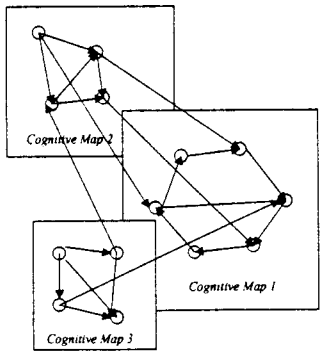

Figure 2. Hierarchical Cognitive Maps

The hierarchical cognitive map of maize hybrids resistant to maize (corn) earworm in an agri-ecosystem consists of five cognitive maps in white circles. In the biosphere, the falvonoid pathway is another hierarchical cognitive map, which may contain multiple cognitive maps itself.

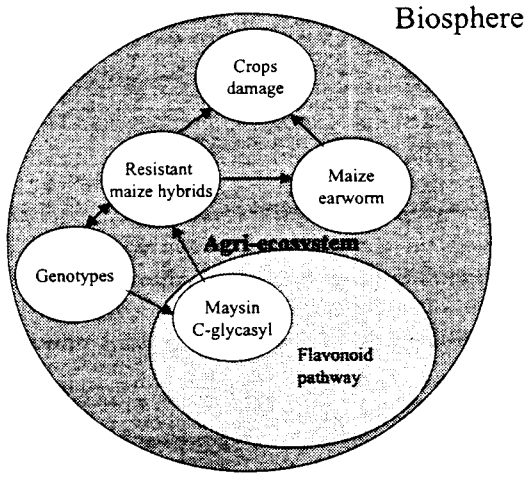

Figure 3. A hierarchical cognitive map for the maize ecosystem.

5. Neural Network, Probabilistic, or Fuzzy Cognitive Maps

Cognitive maps can extend to probabilistic, or fuzzy cognitive maps [8], and further to neural network learning maps [17]. These numerically enabled cognitive maps can be interfaced with other numerical simulation packages in biology [12,13]. We have developed the cognitive map software, called Pool2 [16]. Furthermore we have extended it to labeling processes.

Now we briefly describe learning and reasoning as a labeling process [3,4,5]. Let Σ be a collection of biological objects $\{x_1,...,x_n\}$ (e.g., gene sequences, protein structures, metabolites, genotypes, and phenotypes), and let Λ be a collection of labels $\{\lambda_1,...,\lambda_m\}$ with any mathematical structure (e.g., concentrations and intensities). The labeling problem is to find a consistent labeling of biological objects in Σ by Λ, given a set of relations among objects and a set of constraints among objects and their labels. For each x_i, let Λ_i be a subset of Λ that is compatible with x_i. For any pair $\{x_i,x_j\}$ of objects (i,j distinct), let Λ_{ij} be a subset of compatible pairs of labels in $\Lambda_i \times \Lambda_j$. A labeling $L = \{L_1,...,L_n\}$ is an assignment of a set of labels Λ_i in Λ to each x_i. L is consistent if for each i,j and all λ in Λ_i, $(\{\lambda\} \times \Lambda_j)$ intersects with Λ_{ij}. L is unambiguous if it is consistent and assigns only a single label to each object. The semantic labeling of cognitive maps is described as follows.

The semantic labeling is to assign a measure $m_i(\lambda)$ to the statement "λ is the correct label of x_i". An arbitrary labeling of a knowledge base may not be consistent and unambiguous, because the constraint satisfaction is required among either objects in the knowledge base or a combination of new input evidences with the knowledge base. The interaction with external users and systems is through a query system. At the initial stage, the $m_i(\lambda)$ is either estimated by the user or is provided by another cognitive map or simulation tool. Now the initial measures go through a constraint satisfaction checking by the label relaxation, which iterates the process until the convergence to final measures is reached. The final measures are sent back to the query subsystem for either interaction with external worlds or further reasoning. Thus, human group behavior is assisted by the label relaxation.

The relaxation scheme is mathematically described as follows. An initial assignment of measures $\{m_i(0)(\lambda)\}$ to $\{x_i\}$ is given at time 0. A relaxation operator R is defined to transform one set $\{m_i(k)(\lambda)\}$ of measures to another set $\{m_i(k+1)(\lambda)\}$. The limit $\{m_i^*(\lambda)\}$ of $\{m_i(k)(\lambda)\}$ gives the unambiguous labeling under compatibility constraints, as k approaches to infinity. In reality, we expect the limit to be attained after a finite number of iterations. In practice, the limit $\{m_i^*(\lambda)\}$ may not be unique (we are not always getting an unambiguous labeling). The multiple labelings are sent back to the users so that they can select an appropriate result for further reasoning.

There are several ways to define the relaxation operator R. A relaxation operator R should produce $m_i(k+1)(\lambda)$ from the combination of $m_i(k)(\lambda)$ and support $s_i(k)(\lambda)$ by some update equations, where $s_i(k)(\lambda) = \Sigma\, r_{ij}(\lambda,\lambda')m_j(k)(\lambda')$, where $r_{ij}(\lambda,\lambda')$ is the compatibility function of "label λ is assigned to x_i and label λ' is assigned to x_j", and j-indices are indices of all source nodes leading to the i-th node. A relaxation operator R is defined by the following update equations:

$$m_i(k+1)(\lambda) = \min[1, \max(0, m_i(k)(\lambda)+s_i(k)(\lambda))],$$
$$s_i(k)(\lambda) = \Sigma(r_{ij}(k)(\lambda,\lambda')+\Delta r_{ij}(k)(\lambda,\lambda'))m_j(k)(\lambda'),$$
$$\Delta r_{ij}(k+1)(\lambda,\lambda') = a_{ij}\Delta r_{ij}(k)(\lambda,\lambda')+b_{ij}m_i(k+1)(\lambda)m_j(k)(\lambda'),$$

where a_{ij} and b_{ij} are learning parameters. The first equation makes sure that $m_i(k+1)(\lambda)$ stays between 0 and 1. The second equation provides the network input to the (i,λ)th node. The third equation includes the Hebbian learning rule.

References

1. R. Axelrod, Structure of Decision, Princeton University Press, 1976.

2. P. F. Byrne, M. D. McMullen, M. E. Snook, T. A. Musket, J. M. Theuri, N. W. Widstrom, B. R. Wiseman, and E. H. Coe, Quantitative trait loci and metabolic pathways: Genetic control of the concentration of maysin, acorn earworm resistance factor in maize silks, Proc. National Acad. Sci., USA, 93, 1996, pp. 8820-8825.
3. S. Chen, Some extensions of probabilistic logic, Proc. AAAI Workshop on Uncertainty in Artificial Intelligence, Philadelphia, Penn., August 8-10, 1986, 43-48; An extended version appeared in Uncertainty in Artificial Intelligence, Vol. 2, edited by L. N. Kanal and J. F. Lemmer, North-Holland.
4. S. Chen, Automated reasoning on neural networks: A probabilistic approach, IEEE First International Conference on Neural Networks, San Diego CA, June 21-24, 1987.
5. S. Chen, Knowledge acquisition on neural networks, Uncertainty and Intelligent Systems, edited by B. Bouchon, L. Saitta and R. R. Yager, Lecture Notes in Computer Science, Springer-Verlag, Vol. 313, 1988, pp. 281-289.
6. H. Kitano, Perspectives on systems biology, "New Generation Computing", Vol.18, No.3, Ohm-sha, Springer-Verlag, New York Inc., 2000.
7. M. C. Kohn and W. J. Letzkus, A graph theoretical analysis of metabolic regulation, Journal of Theoretical Biology, 100, 1983, pp. 293-304.
8. B. Kosko, Fuzzy cognitive maps, Int. Journal of Man-Machine Studies, 24, 1986, pp. 65-75.
9. E. S. Lander and P. Green, Construction of multilocus genetic linkage maps in humans, Proc. Nat. Acad. Sci. USA, 84, 1987, pp. 1-5.
10. E. S. Lander and D. Botstein, Strategies for studying heterogeneous genetic traits in humans by using a linkage map of restriction fragment length polymorphisms, Proc. Nat. Acad. Sci. USA, 83, 1986, pp. 7353-7357.
11. A. L. Lehninger, Principles of Biochemistry, Worth Publishers, 1982.
12. P. Mendes, GEPASI: A software package for modeling the dynamics, steady states and control of biochemical and other systems, Computer Applications Biosciences, 9, 1993, pp. 563-571.
13. P. Mendes, Biochemistry by numbers: simulation of biochemical pathways with Gepasi 3, Trends in Biochemical Sciences, 22, 1997, pp. 361-363.
14. V. N. Reddy, M. L. Mavrovouniotis, and M. N. Liebman, Petri net representations in metabolic pathways, Proc. ISMB, 1993, pp. 328-336.
15. R. Somogyi and C. A. Sniegoski, Modeling the complexity of genetic networks: understanding multigenic and pleiotropic regulation, Complexity, 1, 1996, pp. 45-63.
16. W. R. Zhang, S. Chen, and J. C. Bezdek, Pool2: A generic system for cognitive map development and decision analysis, IEEE Trans. SMC, 19, 1989, pp. 31-39.
17. W. R. Zhang, S. Chen, W. Wang, and R. S. King, A cognitive map based approach to the coordination of distributed cooperative agents, IEEE Trans. SMC, 22, 1992, pp. 103-114.
18. Z. B. Zheng, Precision mapping of quantitative trait loci, Genetics, 136, 1994, pp. 1457-1468.