Conditional Probabilistic Modeling of Carbohydrate Metabolism

I. Barjis, M. Sierra
Department of Biological Sciences
New York City College of Technology
The City University of New York

ABSTRACT

The process of metabolic network is very complex, and consequently, difficult to understand and to teach; furthermore, it is impossible to predict and analyze it when unpredictable changes are made. Because of the complexity of metabolic networks and their regulation, formal modeling is a useful method to improve the understanding of these systems. To achieve our goal we've used probabilistic modeling methods to model, analyze, and simulate the process of carbohydrate metabolism in a very compact notation. In particular our research is directed to the development of new probabilistic model of complex biological process such as carbohydrate metabolism. In this paper we use Hidden Markov Models (HMMs) and conditional statistics to model and simulate the process of carbohydrate metabolism.

1INTRODUCTION

The continuing growth of amount of biological data acquired, the development of genome-wide measurement technologies, and the shift from the study of individual process to a systems view all contribute to the need to develop computational techniques for learning models from data. At the same time, continuous increase in computer power enables the simulation of complex biological and molecular process, composed of hundreds of transitions (reactions) (Molinero et al, 2005). The aim of this paper is to provide a broad look at state of the art techniques used in the probabilistic modeling methods involving biological structures and systems, and to bring together method developers and experimentalists working towards the same end. Carbohydrates are principal components of many natural products, and form structures ranging from monosaccharides (24 atoms, molecular weight 180) to complex polysaccharides composed of thousands of these units. In many relevant cases –as glycogen, cellulose, and their hydrolysis products - the polymers are composed of only one subunit (monosaccharide) type. The complexity of the natural product is due to the length, connectivity and branching of these chains.

Biological and molecular processes are complex, dynamic, and invisible. These processes can be investigated and studied from a very detailed perspective to a very high, abstract perspective. The aforementioned characteristics make these processes difficult to explain, teach, illustrate, and understand. Furthermore, any laboratory experiments of biological processes or reactions are time-consuming studies. The results of these experiments may take days, or even weeks before the dynamic behavior of the reaction of process can be observed. Even then, many biological experiments fail to get the desired or expected results. Consequently traditional lab experiments make them not only time consuming, but costly as well. Creating a biological simulation or model that promotes the development of hypotheses on interactions that occur with limited knowledge of dynamic inputs can lead to a better understanding of the behaviors of that system (Collins 2003, Ideker 2001, Palsson 2000, Segrè 2002) The importance of developing biological models has value in the field of healthcare. Furthermore the national Institute of Health released a statement indicating "the continued innovation and development in model systems is important to progress in improving the health of the nation." (NIH Online Statement, 1989)

HIDEN MARKOV MODEL

A Hidden Markov Model (HMM) is a statistical model where the system being modeled is assumed to be a process with unknown parameters, and the object is to determine the hidden parameters from the observable parameters [Figure 1] The extracted model parameters can then be used to perform further analysis of the subject matter. In the second half of the 1980s, HMMs began to be applied to the analysis of biological processes. Since then, they have become a vital and indispensable tool in the field of bioinformatics (Rabiner,1989). HMMs have predominantly been used to model processes in which there are naturally occurring sequences, such as DNA sequencing, gene location prediction, or protein structure prediction. HMMs operate with remarkable reliability and precision, with proven accuracies of over 90% (Krogh, 1994). HMMs are also widely used in speech recognition, body motion recognition, and optical character recognition.

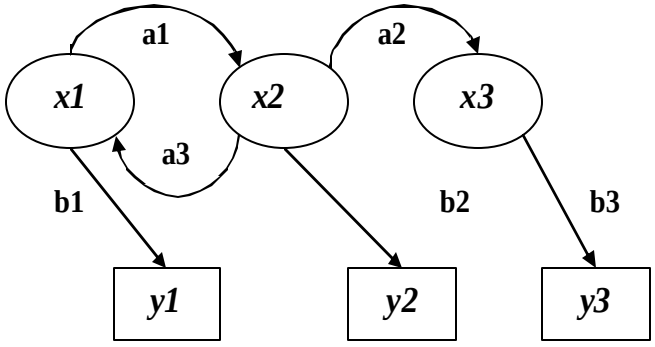

Figure 1: Sample Hidden Markov Model
x - hidden states | *y* - observable outputs | *a* - transition probabilities | *b* - output probabilities

NSTI-Nanotech 2007, www.nsti.org, ISBN 1420061828 Vol. 1, 2007

CARBOHYDRATE METABOLISM

We will create an HMM that will help us simulate the metabolic pathways that the human body uses throughout the day. Humans take in many types of compounds, such as carbohydrates, lipids, and proteins. These compounds are broken down and all converge into one Common Catabolic Pathway. The most readily metabolized compound in the body is Glucose, a 6-carbon monosaccharide. The first step involved in utilizing glucose for energy is known as glycolysis. Glycolysis, via multiple enzymatic reactions, converts glucose into pyruvate. Depending on various conditions, the pyruvate can either be converted to Acetyl Co-Enzyme A (CoA), or Lactic Acid. Under normal conditions, CoA is the dominant product, and is used in aerobic respiration, which occurs in the presence of oxygen. In the absence of oxygen, pyruvate is catabolized via anaerobic respiration into Lactic Acid, which yields 18-19 times less energy than aerobic respiration. During increased levels of physical exertion, such as vigorous exercise, the demand for energy exceeds the supply of oxygen. The body reverts to anaerobic respiration during these periods of increased oxygen demand. But it is not a requisite for our muscles to be oxygen deprived in order to produce lactic acid. Lactic acid can actually be created and transported to muscle groups undergoing high rates of metabolism from muscle groups undergoing little to no metabolism (Farham, 2000)

MODELING OF CARBOHYDRATE METABOLISM

Many significant applied and basic research questions in science today are interdisciplinary in nature, involving physical and/or biological sciences, mathematics, and computer science in an area called computational science. Frequently, a research project has a team of professionals from a variety of fields. The ability to understand various perspectives and perform interdisciplinary work can aid communication and speed the progress of a project. Moreover, the use of computers has become an essential ingredient to many of such projects

In the post-genome era, biopathway information processing is one of the most important research topics in Bioinformatics. Powerful analytical technologies are becoming ubiquitous in biology, which are characterized by high-throughput parallel measurement of large numbers of molecular species. Further more the increasing knowledge in biology and improved measurement methods allow to build detailed models of the cellular interior and molecular processes..

Biochemical network models can be used to predict, explain and hypothesize about phenomena. When made quantitative and implemented in computer software, models can be used to carry out large numbers of simulations that are designed to answer 'what-if' questions. There are currently several software applications that make the process of modeling biochemical networks and use them to simulate data for the purpose of assessing the efficiency of analysis algorithms (Mendes, 1997) Biochemical modeling and simulation are becoming an important method to study data analysis algorithms in systems biology (Fiehn , 2003).

In this paper we use Hidden Markov Modeling techniques and probabilistic statistics to models and simulate, and study the biochemical network known as Carbohydrate.

Modeling of metabolism serves various useful purposes. Firstly, it attempts to improve our understanding of homeostatic regulation by evaluating essential parts or aspects of the metabolic system. To construct a detailed metabolic model of known pathways, the kinetics of the involved enzymes must be calculated, or data may be recovered from the literature. The kinetic data together with data on the effects of co-factors, pH, and so on are used to parameterize the model. This straightforward type of modeling means translating biochemistry into mathematics (Giersch, 2000). In the absence of this data, or when developing a model on which other models are to be based, it is safe to assume values, as long as the values reflect typical biological or physiological response.

In this model, the hidden state is going to be the subject's level of physical exertion. The observable output will be Lactic Acid or Acetyl CoA. By looking at the outputs, we will be able to determine which metabolic pathway the body is using at a specific time. With this information, we can infer with some accuracy as to the level of exercise the subject is undergoing.

For the model, let us assume that throughout the course of the day, a human is in one of four levels, or *states* of physical exertion. The first is rest, which includes sleeping, sitting, or standing. The second is light activity, which includes paced walking, showering, or eating. The third is moderate physical activity, which includes cleaning the house, gardening, or walking up steps. And the last is vigorous physical activity or exercise, examples of which include jogging, running, or weight training. As explained earlier, each of these states place a different oxygen demand on the body, and subsequently, the end products produced by glycolysis vary from state to state. We will assign assumed values to each of these states based solely on the notion that the greater the level of exertion, the greater the amount of lactic acid produced.

An HMM is defined by an alphabet of emitted symbols ?, a set of hidden states *Q*, a matrix state of transition probabilities *A*, and a matrix of emission probabilities *E*, where

- ? ? is an alphabet of symbols .
- ? *Q* is a set of states, each of which will emit symbols from the alphabet ? .
- ? $A = (a_{kl})$ is a |Q| x |Q| matrix describing the probability of changing to state *l* after the HMM is in state *k*; and
- ? $E = (e_k(b))$ is a |*Q*| x |? | matrix describing the probability of emitting the symbol *b* during a step in which the HMM is in state *k*.

Metabolic Pathway products corresponds to the following HMM m

- ? ? = {0,1}, corresponding to Acetyl CoA (0), Lactic Acid(1),
- ? *Q* = {R,L,M,E}, corresponding to rest (*R*), light activity(*L*), moderate activity (*M*), or exercise(*E*)
- ? $a_{RR} = a_{LL} = a_{MM} = a_{EE} = 0.7$, $a_{RL} = a_{RM} = a_{RE} = a_{LR} = a_{LM} = a_{LE} = a_{MR} = a_{ML} = a_{ME} = a_{ER} = a_{EL} = a_{EM} = 0.1$
- ? $e_R(0) = ¾$, $e_R(1) = ¼$, $e_L(0) = 5/8$, $e_L(1) = 3/8$, $e_M(0) = 1/2$, $e_M(1) =$
- ? $1/2$, $e_E(0) = 1/4$, $e_E(1) = 3/4$

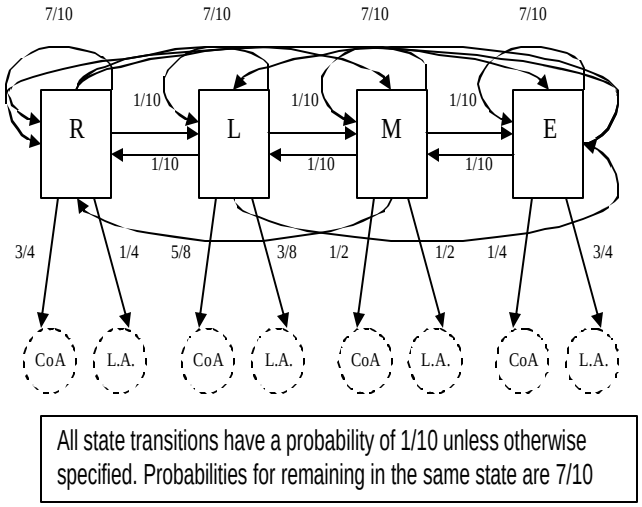

Figure 2: HMM showing State Transition Probabilities

A *path* $p = p_1 \ldots p_n$ in the HMM m is a sequence of states. Our subject is going to follow the routine of a typical weekend day. As he carries out various activities, his state changes. For the sake of uniformity, we will assume that the subject stays in a given state for at least one hour and that state transitions can only occur only on the hour. Following is a list of the activities and the state assigned to it (Table 1).

Table 1: Twenty four hour activity

Hour	Activity	State	Hour	Activity	State
0000	Sleep	*R*	1200	Garden	*M*
0100	Sleep	*R*	1300	Wash Car	*M*
0200	Sleep	*R*	1400	Lunch	*L*
0300	Sleep	*R*	1500	Gym	*E*
0400	Sleep	*R*	1600	Lift Weights	*E*
0500	Sleep	*R*	1700	Dinner	*L*
0600	Sleep	*R*	1800	Home Repairs	*M*
0700	Sleep	*R*	1900	Watch TV	*L*
0800	Shower/Eat	*L*	2000	Clean House	*M*
0900	Walk Dog	*L*	2100	Read	*L*
1000	Morn. Jog	*M*	2200	Shower	*L*
1100	Read Paper	*L*	2300	Sleep	*R*

The corresponding path *p* for this day is

p = RRRRRRRRLLMLMMLEELMLMLLR.

If the resulting metabolic products are:

000000000010010110100000, then the following shows the matching of *x* to *p* and the probability of x_1 being generated by p_1 during a given hour.

$P(x_i|p_i)$ denotes the probability that symbol x_1 was emitted from state p_1. $P(p_i \; ? \; p_{i+1})$ denotes the probability of the transition from state p_i to p_{i+1} (fig. 3 and 4).

x	0	0	0	0	0	0	0	0	0	0	1	(
? =	R	R	R	R	R	R	R	R	L	L	M	I
P(xi| p i)	3/4	3/4	3/4	3/4	3/4	3/4	3/4	3/4	5/8	5/8	1/2	5

Figure 3: Hours 12a-11a

x	0	0	0	0	0	0	0	0	0	0	1	0
? =	M	M	L	E	E	L	M	L	M	L	L	R
P(xi| p i)	1/2	5/8	3/4	3/4	5/8	5/8	1/2	5/8	1/2	5/8	5/8	3/4

Figure 4: Hours 12p-11p

The path *p* = RRRRRRRRLLMLMMLEELMLMLLR includes multiple state transitions. The subject occasionally remains in the same state as the previous hour. The state changes are not random; they are based on the activities that the subject carries out. Let us assume however, that the subject's activities were based on the probabilities of state changes that we previously assigned. For example, the probability of the switches $p_8 \; ? \; p_9$ and $p_{19} \; ? \; p_{20}$ is 1/10. The probability of the switches $p_3 \; ? \; p_4$ and $p_{16} \; ? \; p_{17}$ is 7/10. The probabilities of the state changes throughout the course of the day are shown below (fig. 5 and 6).

x	0	0	0	0	0	0	0	0	0	0	1	0
? =	R	R	R	R	R	R	R	R	L	L	M	L
P(xi| p i)	3/4	3/4	3/4	3/4	3/4	3/4	3/4	3/4	5/8	5/8	1/2	5/8
$P(p_i \; ? \; p_{i+1})$	7/10	7/10	7/10	7/10	7/10	7/10	7/10	1/10	7/10	1/10	1/10	1/10

Figure 5: Hours 12a-11a

x	0	0	0	0	0	0	0	0	0	0	1	0
? =	M	M	L	E	E	L	M	L	M	L	L	R
P(xi| p i)	1/2	5/8	3/4	3/4	5/8	5/8	1/2	5/8	1/2	5/8	5/8	3/4
$P(p_i \; ? \; p_{i+1})$	7/10	1/10	1/10	7/10	1/10	1/10	1/10	1/10	1/10	7/10	1/10	1/10

Figure 6: Hours 12p-11p

NSTI-Nanotech 2007, www.nsti.org, ISBN 1420061828 Vol. 1, 2007

The probability of generating ***x*** through path ***p*** is 5.32×10^{-19} and is computed as

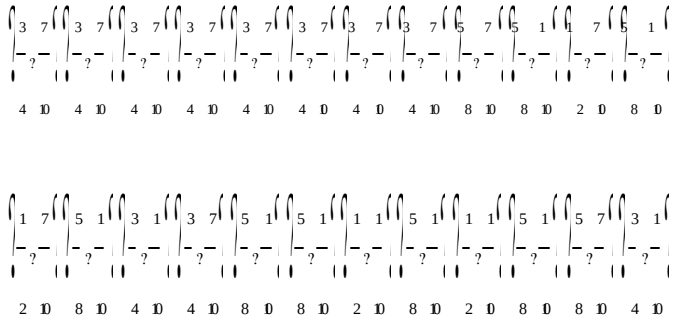

Figure 7: Probability of generating *x* through path pie

CONCLUSION

This paper defines the process of carbohydrate metabolism as a stochastic finite state automaton, using hidden Marko modeling. In the process of carbohydrate metabolism there are finite set of possible states and each of those states are associated with a specific probability distribution. Two areas could benefit from this type of research. First of all this approach is expected to be particularly important in comparing competing data analysis methods that require considerably different experimental setups. In this case, modeling may be the only way to be able to study their performance in a truly comparative way. Secondly the model could be used as a teaching tool by educational institutions. Models are powerful tools for understanding complex biological systems. Furthermore it helps students to be engaged in active learning. This paper as an initial step of ongoing research shows how probabilistic statistics and HMM can be used to model the process of carbohydrate metabolism.

REFERENCES

1. Collins FS , Green ED , Guttmacher AE , Guyer MS (2003) A vision for the future of genomics research. *Nature* 422: 835–847
2. Covert MW , Schilling CH , Famili I , Edwards JS , Goryanin II , Selkov E , Palsson BO (2001) Metabolic modeling of microbial strains *in silico*. *Trends Biochem Sci* 26: 179–186
3. Edwards JS , Palsson BO (2000) The *Escherichia coli* MG1655 *in silico* metabolic genotype: Its definition, characteristics, and capabilities. *Proc Natl Acad Sci USA* 97: 5528–5533
4. Farham, Bridget. (2000) Lactate: Myth and Facts. *SPORTS SCIENCE Training and Muscle Metabolism*, May
5. Fiehn, O. and Weckwerth, W. (2003) Eur. J. Biochem. 270, 579–588
6. Foster, D.M., and Hetenyi, G. Jr, Journal of Parenteral and Enteral Nutrition, Vol 15, Issue 3, 67S-71S
7. Giersch, C., (2000) Mathematical Modelling of Metabolism. Current Opinion in Plant Biology 2000, 3:249–253
8. Ideker T , Thorsson V , Ranish JA , Christmas R , Buhler J , Eng JK , Bumgarner R , Goodlett DR (2001) Integrated genomic and proteomic analyses of a systemically perturbed metabolic network. *Science* 292: 929–934
9. Kauffman KJ , Prakash P , Edwards JS (2003) Advances in flux balance analysis. *Curr Opin Biotechnol* 14: 491–496
10. Krogh, A., Mian I. S., and Haussler D.(1994) A hidden Markov model that finds genes in E. coli DNA. Nucleic Acids Res. November 11; 22(22): 4768–4778.
11. Mendes,P.(1997)Trends Biochem.Sci. 22 ,361 – 363
12. Nicholson JK , Holmes E , Lindon JC , Wilson ID (2004) The challenges of modeling mammalian biocomplexity. *Nat Biotechnol* 22: 1268–1274
13. Palsson BO (2000) The challenges of *in silico* biology. *Nat Biotechnol* 18: 1147–1150
14. Rabiner, Lawrence R.,(1989) A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, 77 (2), p. 257–286.
15. Segrè D , Vitkup D , Church GM (2002) Analysis of optimality in natural and perturbed metabolic networks. *Proc Natl Acad Sci USA* 99: 15112–15117
16. Stephanopoulos G , Alper H , Moxley J (2004) Exploiting biological complexity for strain improvement through systems biology. *Nat Biotechnol* 22: 1261–1267
17. Modeling in biomedical research: An assessment of current and potential approaches. NIH Technol. Assess Statement Online 1989 May 1-3, 2006 September 26; (4):19.