Center for the Science of Information

Seed Grant Proposal (Tier I)

Investigation of Metabolic Phenomena Using Information Theory

Frank DeVilbiss / fdevilbi@purdue.edu / Doraiswami Ramkrishna / Purdue University
Pablo Robles-Granda / problesg@purdue.edu / Jennifer Neville / Purdue University
Mohan Gopaladesikan / mgopalad@purdue.edu / Mark Daniel Ward / Purdue University

Adviser:
Doraiswami Ramkrishna / ramkrish@ecn.purdue.edu / Purdue University
Maxim Raginsky / maxim@illinois.edu / University of Illinois at Urbana-Champaign

10/19/2012

Problem Statement

Intuitively, an organism performs various metabolic reactions with the goal of survival, but the overarching control goal driving metabolism, programmed into said organism's DNA, is not fully understood at this juncture. Modeling metabolism has relied on steady state concepts without a full dynamic description of system-wide behavior. The predictive power of such approaches is limited. However, a cybernetic mathematical theory, based on viewing the regulation of metabolism as motivated by a survival goal, has been recently shown to successfully anticipate many metabolic phenomena. These include diverse uptake patterns of nutrients as well as a multiplicity of metabolic states at particular growth rates. These results show that metabolism's control objective may be much different from other constraint-based models that are widely used which claim that the cells regulate their metabolic activity towards the maximization of growth. Cybernetic models, on the other hand, have revealed that metabolism may function by dynamic optimal goals such as maximizing the carbon uptake rate at each instant. Understanding what control goal is programmed into DNA is essential to effective metabolic engineering efforts and, more fundamentally, to the study of biology itself.

Main question: Does the sum of metabolic regulation converge upon the control goal carbon uptake rate maximization?

Proposed Activity

Together, we will attempt to look at how the dynamic control goals that guide metabolic function are represented in complex sets of bioinformatic data. To do so, we will need to understand the informational *context* of biological data. Additionally, the isolation of trends that demonstrate dynamic, cellular optimization to qualitatively evaluate the validity of the idea of carbon uptake rate maximization is important. In the short term, we would like to collaborate together to determine how to best analyze gene expression data as a *big data problem*. Large vectors of data that represent cellular mRNA, protein and metabolite profiles taken at different growth conditions will demonstrate the complex interaction of biological functions which will be best investigated by a multi-disciplinary team that has expertise in the fields of biology, machine learning and statistics. Frank has a strong background on metabolic modeling. Pablo has written a thesis on machine learning techniques and Mohan has much experience doing applied statistical analysis. The types of information that will be analyzed conjointly will be gene expression changes in the context of protein reaction rates, network structure of metabolic pathways, and metabolic fluxes. For example, if genes with low protein kinetics are upregulated dramatically at higher growth rate conditions, the cell is trying to compensate and increase pathway throughput to maximize carbon uptake rate. By looking at these data types together, a clearer picture of the theory driving metabolic states to shift from one to another will be yielded.

Expected Outcomes

A seed grant from the CSoI would enable better communication amongst this team on both a short and long term basis. To enumerate short term achievements provided the bestowal of this grant, we plan to:

1) Develop an effective knowledge acquisition framework to most effectively analyze bioinformatics data. This analysis will attempt to yield a broader understanding of the coordinated regulation of many genes towards certain metabolic outcomes.
2) Formulate a unique method to analyze bioinformatics data as a complement to regulatory database information provided for organisms like *E. coli* in RegulonDB.

As for long term goals, this group would like to validate a theory of metabolism that may surpass our current knowledge of biological systems. The interaction of many biological functions at the genetic level represents an algorithm that guides metabolic states towards specific "optimal" configurations. Knowledge of this algorithm distilled from large states of data is of significant impact for the field of biology. It ascribes a mechanistic causality to the guided actions of the genetic circuit in biology determined by a necessity for timely interaction of the cell with its environment for the purposes of an organism's survival. While intuitive, a complete understanding of metabolic regulation in this context remains unknown.

Proposed Work Statement

To see an outline of the work proposed, look at figure 1 attached at the end of this narrative. The overall goal is model validation. Two alleyways will be taken to accomplish this task. The first is looking at the statistical agreement of cybernetic models with experimental data. Cybernetic models represent the theory of carbon uptake maximization mathematically. Being able to simulate dynamically shifting metabolic systems accurately represents a triumph in the theory's ability to describe metabolic systems. To analyze the models in a statistical sense, they will be compared to other metabolic models for the same systems using two methods: Kullback-Leibler Divergence and an analysis on the effects of noise upon model output. The other alleyway will be towards gauging how sound the theory is in terms of what complex bioinformatic data is saying about the model predictions. To do this, both large sets of bioinformatic data sets will be examined as well as regulation databases to extract the idea of carbon uptake rate maximization.

The division of activities is as follows:

- Frank DeVilbiss – Will take the lead in the proposed activities. He has much background in using cybernetic models of metabolism as well as experience with bioinformatic data sets. This work will also apply to his thesis.
- Pablo Robles-Granda – Will provide his expertise on machine learning to extract relevant trends in large sets of data.
- Mohan Gopaladesikan – Will add his incite as a statistician towards the analysis of data.

Team meeting frequency will be at least once every two weeks and will occur in one of two ways. The first is just a meeting of group members from Purdue. The second is a meeting of Purdue members and Maxim Raginsky, a faculty member at UIUC who is interested in collaborating in this research project. To do so, both meetings in person (necessitating travel) and meetings on Skype are proposed.

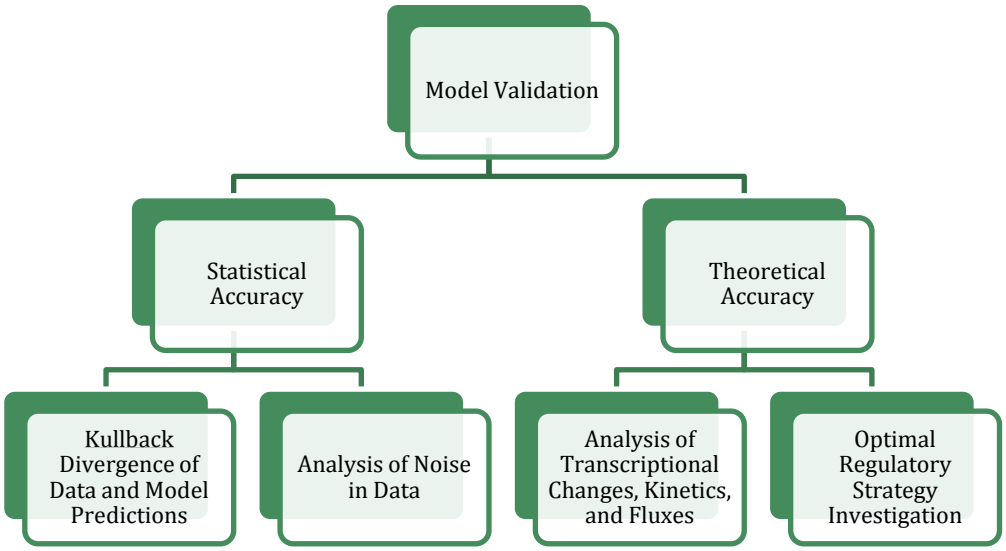

Figure 1

References

Some relevant references to this research task are as follows:

Heinemann, M., & Sauer, U. (2010). Systems biology of microbial metabolism. *Current Opinion in Microbiology* , 337-343.

Ishii, N., Nakahigashi, K., Baba, T., Robert, M., Soga, T., Kanai, A., et al. (2007). Multiple High-Throughput Analyses Monitor the Response of E. coli to Perturbations. *Science* , 593-597.

Kompala, D. S., Ramkrishna, D., Jansen, N. B., & Tsao, G. T. (1986). Investigation of Bacterial Growth on Mixed Substrates: Experimental Evaluation of Cybernetic Models. *Biotechnology and Bioengineering* , 1044-1055.

Song, H.-S., Ramkrishna, D., Pinchuk, G. E., Beliaev, A. S., Konopka, A. E., & Fredrickson, J. K. (Accepted). Dynamic Modeling of Aerobic Growth of Shewanella oneidensis. Predicting Triauxic Growth, Flux. *Metabolic Engineering* .

Wessely, F., Bartl, M., Guthke, R., Li, P., Schuster, S., & Kaleta, C. (2011). Optimal regulatory strategies for metabolic pathways in Escherichia coli depending on protein costs. *Molecular Systems Biology* , 1-13.

Young, J. D., & Ramkrishna, D. (2007). On the Matching and Proportional Laws of Cybernetic Models. *Biotechnology Progress* , 83-99.

Budget & Justification

Budget section removed

PI Research Statements

Frank DeVilbiss – School of Chemical Engineering, Purdue University

My personal research is tied intimately with this project.

The goal of my project is the validation of cybernetic metabolic models and the control objectives that drive them. Given success in predicting multiple steady states in a multi-substrate chemostat reactor, the idea that cells control their metabolic machinery in order to maximize their carbon uptake rate has much more credibility (Kim, Song, Sunkara, Lali, & Ramkrishna). Going beyond extracellular measurements, cybernetic models have been used to predict the internal metabolic fluxes of cells better than constraint based metabolic models for the bacteria *Shewanella oneidensis* (Song H.-S. , Ramkrishna, Pinchuk, Beliaev, Konopka, & Fredrickson, Accepted). These studies suggest that the maximization of carbon uptake rate may be an accurate summary of a fundamental regulatory structure in nature.

With great claims also comes the necessity to extensively test their veracity. While no mean exists to "prove" a theory, it is possible to generate a body of evidence that supports one. To accomplish this, two thrusts of research are being developed. One is the extensive analysis of the statistical accuracy of cybernetic models while the other is an investigation of the theoretical accuracy of the proposed models. The statistical validation of the cybernetic theory will involve analyzing models on the basis of the errors associated with their parameterization. The theoretical validation of the cybernetic models will involve the extraction of regulatory patterns from large sets of bioinformatic data that are logically consistent with the cybernetic approach.

Collaboration Benefits

In this proposed seed grant project, it is my hope that I will get insight from outside of my scope of research experience. I wish to collaborate with other researchers who are proficient at machine learning methods and statistics. While I have an understanding of some methods of analysis of large sets of bioinformatic data, I would like to develop better data examination techniques that are appropriate for the needs of my research. I know that through collaboration, I will have a much greater chance of success in this project.

Pablo Robles Granda – Department of Computer Science, Purdue University

My research interest lays in the intersection of network analysis, statistical analysis, operations research, and artificial intelligence. In particular, I am interested on how to develop methods for hypothesis testing of complex networks with particular emphasis in problems in the social and biological sciences.

Towards this goal, I am currently working on the development of both descriptive and predictive models of big networks. One of the questions that arises in network analysis is related to how different characteristics of the network are studied. Some examples of these characteristics are the structure, node distribution, geodesic distance, directionality of the edges, weight of edges, etc. The emphasis of some characteristics over others highly depend on the context of the problem being studied. In the case of metabolic networks, one important question is how to determine the accuracy of a prediction when compared to noisy or erroneous experimental data.

Through my work I expect to develop tools to analyze the complex characteristics of networks that arise in cellular metabolism. Although a full theory has been developed about the internal mechanism that the cell follows in order to process external substances and how this affects the cell's environment in terms of the changes of both substrates and biomass, the current limitation is related to the lack of information-theoretical, statistical, and computational methods for a full prove of this theory of cell metabolism. This problem is particularly affected by the size of metabolic networks, the complexity of the characteristics of these networks, and limitations related to "purity" of experimental data.

I believe that this research is of key importance in current times where the amount of unstructured relational and network data and the lack of mathematical tools to analyze it make it very difficult for social and biological scientist to test hypotheses. In consequence, development and verification of theories in both areas are difficult tasks. Addressing this problem is a necessity that can be overcome through interdisciplinary work where people with expertise in biological, mathematical, and computational domains can contribute to solve the problem. The development of methodologies, and the development of techniques, to analyze data in the domain of metabolic networks will benefit not only our research team but the scientific community as a whole through new insights about hypothesis testing.

Collaboration Benefits

My expectation for the work in this project is to gain insights on the behavior of biological networks, particularly of metabolic ones. I also believe that my expertise in computational tools will positively impact the work in our team. The multi-disciplinary nature of the analysis of metabolism will allow me to investigate how to create better statistical tools for hypothesis testing of networks in the biological domain.

Mohan Gopaladesikan - Department of Statistics, Purdue University

My primary research is broadly in analysis of algorithms and random structures. More specifically the use of analytic (complex-valued) methods to determine the precise asymptotic growth rate of various parameters in combinatorics and probability. These methods are very powerful and non-traditional in probability theory. However I have background from my masters in operation research and mathematical modeling of the real world problems. I have done a lot of applied statistic classes which coupled with my mathematical modelling skills would be very useful in cybernetic modeling of metabolism. Though this project is not directly related to my primary research area(which is very theoretical), being involved in this project would give me an opportunity to diversify my research and put to use my applied statistics knowledge. This project is very aligned with the Center's thrust area of life sciences, with imperative collaboration of computer science, statistics, engineering and biology. I think this is a great opportunity for me to be involved in this interdisciplinary projects involving big data.