若手研究者インターナショナル・トレーニング・プログラム(ITP)
バイオインフォマティクスとシステムズバイオロジーの国際連携教育研究プログラム　応募書類

Name: Timothy Hancock

Title: Machine Learning Approaches for Computation Systems Biology

Institute: 京都大学化学研究所バイオインフォマティクスセンター

Partner Institute of your choice: Theoretical Biophysics Laboratory, Institute of Biology, Humboldt-Universität zu Berlin

Duration of your choice: 21st October 2011 to 8th January 2012

Plan:

Any observed biological phenomenon is the result of interactions between extensive numbers of unknown elements. Systems biology describes the process of identifying these elements and piecing them together into a model which fully reconstructs the targeted observed phenomenon. There are two major philosophies for building such models, mechanism-driven physical models and data-driven statistical models. Both of these approaches have their own specific advantages. Physical models can accurately represent the dynamics of a system in order to simulate real world observations, and statistical models make explicit use of the available experimental data to uncover previously unknown relationships.

Historically there has always been a clear separation in the applicability of these two approaches. For example, physical models are used to model reaction systems because their known chemical properties provide an ideal starting point for such analyses. In contrast, statistical models are better used to represent protein interaction relationships whose specific physical properties are often unknown, but their potential existence can be experimentally predicted with high throughput experimentation.

Recently, separate advances in each field have had the effect of bringing these two distinct fields closer together. These advancements have been prompted by the requirement to overcome the shortcomings of each individual approach. For example, numerous heuristic approaches now exist to overlay experimental data into the standard physical model approach of flux balance analysis. Additionally in an effort to elucidate a more detailed physical description from statistical models, methods based on the stochastic differential equation representation of protein and kinetic activity are being investigated. These combined approaches are being complemented by rapid advances in experimental procedures which are now capable of simultaneous observations over multiple domains, such as transcriptomic and metabolomic data measurements over the same time course.

However, the concept of combing physical and statistical methodologies remains relatively new. Furthermore, many proposed methods are simple heuristics which make dramatic simplifications within one domain to enable the imposition of new information within the previously established methodologies. One common example of this simplification is the discretization of experimental data to define an active set of reactions for input into a physical model. Whilst these heuristic approaches provide a glimpse into the potential advantages that could be gained by combining these two methodologies, their heuristic nature raises several concerns about the validity of the biology they seek to represent. An ideal combination of these two philosophies is clearly one which does not reduce the information within one domain in order to force it into the established methodology of the other. One methodology, stochastic differential equation modeling, provides such flexibility by allowing for a compact representation of the relationships that exist across both domains.

Research Objectives

The goal of this project is to investigate the potential of stochastic differential equation modeling to accurately determine the physical mechanics of metabolic processes from the information contained within high-throughput experimental data. Specifically, we aim to express a metabolic network as a set of three coupled differential equations. These equations model the metabolite, enzyme and mRNA expression levels of a specified network. These equations are defined such that the concentration of a specific metabolite is assumed to be dependent on the enzyme concentration of the producing and consuming reactions. The enzyme concentration is subsequently assumed to be dependent on the mRNA expression of the genes which encode the enzyme which catalyzes these reactions. From these three equations that define our network model, observations from two sources, metabolite concentration and mRNA expression can be sourced by utilizing the data from simultaneous transcriptomic and metabolomic experiments. Then the missing enzyme concentration function can be inferred.

Our aim is to investigate the possibility of inferring the unknown enzyme concentration function by assuming it is a latent factor which transforms the expression signal of the reaction genes into metabolite concentration levels. Latent feature models are probabilistic methods which try to infer an unobservable underlying process which generates the observed data. For metabolic networks it is clear that one potential unobserved process involves the enzyme concentration levels which catalyze the reactions. A latent factor model which can infer the enzyme concentration levels from experimental observation of real metabolic networks could potentially overcome the limitations inherent in both physical and statistical approaches. This model would not only be of great interest to systems and synthetic biologists but also to the broader biological community.

The proposed research objectives falls strongly in line with the large scale objectives of the Theoretical Biophysics, Institute of Biology, Humboldt-Universität zu Berlin under the guidance of Professor Edda Klipp. Within the Theoretical Biophysics laboratory there is a wide variety of research spanning disciplines from experimental biology, theoretical physics to applied computer sciences. A unifying theme amongst these researchers is to combine their respective skills to systems biology. As the goal of this research is to combine ideas from machine learning and physical network models with an overall systems biology objective, a research visit to the Theoretical Biophysics Laboratory would clearly be beneficial. There are three clear objectives to the research plan:

1. To learn the current ways research from computer science and biology are combining their various skills and ideas to meet a systems biology objectives. Specifically, as my background is in computer science my goals will be focused on the understanding the biological objectives of the projects and understanding how machine learning algorithms are altered to meet these objectives.
2. To observe and participate in this collaborative process. This will be achieved by attending seminars and presenting my own research. The overall goal is to receive input on my current research and get new direction ideas such that future research better resembles the systems biological objectives.
3. To investigate the possibility of future Japanese-German collaborative efforts. To achieve this I will be discussing my ideas with the researchers in Professor Klipp's laboratory in an effort to garner interest in a combined effort towards the possible future research objectives.