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: NGUYEN HAO CANH

Title: Latent Feature Models for Biological Knowledge Discovery

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： Centre for Computational Biology, Mines ParisTech

Duration of your choice: Dec 15th, 2012 ~ Mar 14th 2013

Background

It is the objective to find the common latent features that generate patterns from different data types of biological entities (genes, proteins, chemical compounds, networks). These latent features are candidates for explaining biological mechanisms. However, due to its constraints, the problem of finding latent features is computationally infeasible. The aim of this proposal is to contribute to the process of making it possible to construct and train the models approximately or accurately on large biological data. The research falls in the categories of Machine Learning/Data Mining and Bioinformatics. The final target is to use the models to understand biological processes such as protein-protein interactions, gene regulations, subnetwork, pathway formations, gene/chemical compound relationships.

These days, a *massive* amount of biological data is available from high throughput experiments such as sequencing and microarray. Traditional experimental data are also collected and organized in comprehensive databases. This is the chance to study Biology from systems level, to look at biological mechanisms involving multiple entities at the same time. This poses many problems in computational sciences, statistical sciences and so on. This research is in this area, trying to make sense out of the vast biological data.

The collected biological data are of different natures, making traditional analyzing methods not readily usable. Sequential data, for example, needs to be transformed into Euclidean representations, inducing *extremely high dimensions* while most of them are unrelated to the problem. To make the matter worse, spurious knowledge may be inferred due to the dimensionality problem. This is also the problem of time series of microarray data, selecting and combining appropriate features are usually very hard computationally. Data from interactomes (set of all biological interactions), on the other hand, are *relational* data. While traditional experiments only give individual interactions (such as from one protein to another), having a large number of interactions opens challenging problems. That is, by looking at the whole interactomes as networks, it gives rise to network analysis problems that infer additional knowledge that cannot be induced otherwise.

Biological processes usually involve multiple steps, such as pathways, requiring a throughout investigation of many entities at different times. To understand the whole processes, integrated models are required. There are needs to construct models for *complexly structured* data, involving many data types into the same framework. As the entities are diverse, we have to deal with very *sparse* data. As data are collected from independent experiments for different needs, from the whole systems' viewpoint, the data are generated with *unknown distributions* and statistical models need to take this fact into account.

The final goal is to present to biologists comprehensible explanations of biological mechanisms for verification. To this end, one must be able to describe the common mechanisms through different data types involved. For example, the phenomenon of a protein interacting to another protein (PPI) may involve some network substructures, requiring subsequences of the DNA and having certain changes in gene expressions. These are the patterns uniquely associated to the phenomenon. It is common in Statistics that these patterns are hypothesized to be generated by certain *latent features*. A typical example is a PPI is determined by a pair of interacting sites that interact to each other, and the interacting sites generates the patterns in PPI networks and expression profiles. Our research covers three aspects: statistical models for biological data, optimization methods for these models and biological validation of the computed results.

Plan

Our aim is to develop different statistical models specifically for the nature of biological data. As the nature of collected data are sequences, time series or networks, the statistical models of the biological systems have particular characteristics that need to take into account. First, they are of very high dimension. However, a biological phenomenon is usually indicated by a combination of observed features (a latent feature). This results in large structured models in the family of Bayesian or Markovian models. These models are difficult to train as they need combinatorial optimization algorithms. General methods of learning these models such as structured sparse approximation, graphical lasso are general-purposed and require an exponentially increasing amount of data. To address these issues, we specifically plan to investigate the following topics.

1. Latent features transfer: Since the models involve many variables, maximum likelihood or general Bayesian estimations may overfit the training data. It is the problem to incorporate suitable prior knowledge into the models. The priors are biological knowledge such as known relationships between entities or from closely-related species. For this purpose, we plan to *design graphical models of biological processes that reflect the partial knowledge* such as known subnetworks of the models or similar networks for closely-related species on the phylogenetic trees, i.e., the networks that are close to the known networks of closely related species. These problems belong to the categories of predicting missing edges or transfer learning. In these models, the known subnetworks or the parts of networks, that are similar to other species, become the latent features of the biological processes.

2. Inducing networks with latent features: Given the problem of high dimension, statistical estimations need to be regularized. Sparsity is one of the common regularizers that induce reliable estimators. To apply to biological processes, sparsity is usually in form of sparse networks. However, the usual sparsity-inducing methods do not impose any conditions on the inducing models. In biological networks, the networks are known to follow special patterns. This is a gap between general methods of structured sparse estimation and biological networks estimation. This poses different problems. One is the problem of *designing regularizers for graphical models that induce the networks that are close to known ones*. Another problem is to *design regularizers for graphical models that induce networks having the properties of biological networks.*

3. Integrated models for latent features: It is generally a dilemma on latent features of the two representations: networks and Euclidean data types (such as transformed from sequences or time series data). On the one hand, features of network structures are used to predict relationship of expressions of genes. On the other hand, sequences or gene expressions are used to predict network structures. In the case of similarity networks, it is to put both representations into the same space. However, it is extremely difficult to put them into the same space for non-similarity networks. Depending strongly on the type of the networks, the relationship between the two representations must be defined accordingly. It is our aim here to *provide an integrated models of latent features of networks and of Euclidean data for general network types*.