High-Dimensional Data in Biology

Chair: Yutaka Yasui (University of Alberta)

KAREN A KOPCIUK, Alberta Health Services - Cancer Care
Sample Size Estimation Methods for NMR Metabolomics Data

Typical projection-based methods used to analyze metabolomics data, such as Partial Least Squares Regression, do not have statistical inference methods or the corresponding sample size methods. To overcome this study planning limitation, we investigated sample size estimation approaches for high dimensional microarray data for their suitability with NMR metabolomics data. Simulation studies compared three methods across a number of relevant data features. Simulation study results and an application to a pancreatic cancer data set suggested two methods could be used for sample size estimation of NMR metabolomics data. Future work will investigate these methods with mass spectrometry metabolomics data.

TOBY KENNEY, Dalhousie University
A Generalized Codon-based Model of Nucleotide Substitution for Protein-coding DNA Sequences

A generalized codon-based model for protein-coding DNA sequences is proposed for phylogenetic analysis. This model framework provides a unified framework for existing codon (or DNA or amino acid) models. Furthermore, it offers greater flexibility in the choice of rate matrix, allowing existing models to be easily extended to incorporate more of the possible driving forces in molecular evolution, such as structure information and amino acid properties. We provide a software package called Codon Optimal Likelihood Discoverer (COLD) to implement these proposed generalized codon models. We demonstrate how our model framework allows model selection based on standard likelihood theory.

HONG GU, Dalhousie University
Estimation of Darwinian Positive Selection Using Generalized Codon-based Models

The generalized codon-based model for protein-coding DNA sequences offers great flexibility in the choice of rate matrix and significantly improves the goodness-of-fit. We further develop the mixed model approach for this type of model by allowing random effects on one or more parameters across the amino acid sites. By allowing random effects on the parameter that estimates the nonsynonymous change rate on different sites, generalized codon-based models include as special cases the M-series models in Yang et al (2000). We demonstrate the more accurate estimation of heterogeneous selection pressure at amino acid sites, comparing to the estimation by M-series models.

THIERRY CHEKOUO TEKOUGANG, Université de Montreal
Statistical View of Popular Biclustering Algorithms

Recently, several biclustering algorithms have been proposed to reveal submatrices of the data matrix whose rows exhibit similar behaviour across a set of columns, and conversely. These have important applications to gene expression analysis, for example, to find genes that are co-regulated across a subset of conditions. Few of these algorithms are based on explicit models. This work proposes some underlying statistical models associated with some of the most popular biclustering algorithms. It shows that these algorithms can be justified within a Bayesian framework and can be derived from Bayesian computational techniques.

DENA GIVARI, University of Guelph
On the Application of a Bayesian Infinite Mixture Modelling

In microarray analyses, genes are expressed to varying levels across different cellular conditions. It is believed that the identification of gene with similar expression patterns across conditions demonstrating variable temporal, developmental, topographical,

histological and physiological patterns will give insight to their biological functions. The aim is to create clusters which hold genes whose expression levels are inter-related at various conditions. We further develop a recent algorithm based on the Bayesian infinite mixture model; in particular we allow for non-isotropic covariance matrices in order to better reflect reality. Our algorithm is discussed and illustrated on simulated and real data.

VAHID PARTOVI NIA, McGill University
Agglomerative Bayesian Biclustering

In many biological domains such as metabolomics, proteomics, and genetics a data matrix with subjects in rows and variables in columns is produced. For such data simultaneous clustering of subjects and variables, called biclustering, is of interest. For instance, in gene expression data, clustering subjects reflects which subjects have similar genetic make up, and clustering genes reflects which genes might function similarly on the measured subjects. We suggest a fully-automatic biclustering algorithm using a Bayesian model which does not require knowledge about the number of partitions. Furthermore, agglomerative method produces a graphical representation of biclusters through dendrogram.