Edward Chang (Google China)

Title: Parallel Algorithms for Mining Large-Scale Data

Abstract: In this talk, I will first describe both computational and storage challenges to traditional data mining algorithms brought about by information explosion. To deal with huge amount of data that expand continuously, an effective algorithm should be designed to 1) run on thousands of parallel machines for sharing storage and speeding up computation, 2) perform incremental retraining and updates for attaining online performance, and 3) fuse information from multiple sources in order to alleviate information sparseness. I will present algorithms we recently developed including parallel PF-Growth, parallel combinational collaborative filtering, parallel LDA, parallel spectral clustering [4], and parallel Support Vector Machines.

Francis Chin (The University of Hong Kong)

Title: Predicting Protein Complexes from PPI Data
(Other Coauthors: Henry C.M. Leung, Qian Xiang, S.M. Yiu)

Abstract:
Protein complexes play a critical role in many biological processes. Identifying the component proteins in a protein complex is an important step in understanding the complex as well as the related biological activities. In this talk, we address the problem of predicting protein complexes from protein-protein interaction (PPI) network of one species using a computational approach. Most of the previous methods rely on the assumption that proteins within the same complex would have relatively more interactions. This translates into dense subgraphs in the PPI network. However, the existing software tools have limited success. Recently, [Gavin et al. 2006] provided a detailed study on the organization of protein complexes and suggested that a complex consists of two parts: a core and an attachment. Based on this core-attachment concept, we developed a novel approach to identify complexes from PPI network by identifying their cores and attachments separately. We evaluated the effectiveness of our proposed approach using three different datasets and compared the quality of our predicted complexes with three existing tools. The evaluation results show that we can predict many more complexes and with higher accuracy than these tools with an improvement of over 30%. To verify the cores we identified in each complex, we compared our cores with the mediators produced by [Andreopoulos et al. 2007], which were claimed to be the cores, based on the benchmark result produced by [Gavin et al. 2006]. We found that the cores we produced are of much higher quality ranging from 10-fold to 30-fold more correctly predicted cores and with better accuracy.

Luonan Chen (Osaka Sangyo University)

Title: Inferring Transcriptional Interactions and Regulator Activities from Experimental Data

Abstract: Gene regulation is a fundamental process in biological systems, where transcription factors (TFs) play crucial roles. Inferring transcriptional interactions between TFs and their target genes has utmost importance for understanding the complex regulatory mechanisms in cellular systems. On one hand, with the rapid progress of various high-throughput experiment techniques, more and more biological data become available, which makes it possible to quantitatively study gene regulation in a systematic manner. On the other hand, transcription regulation is a complex biological process mediated by many events such as post-translational modifications, degradation, and competitive binding of multiple TFs. In this talk, with a particular emphasis on computational methods, I report the recent advances of the research topics related to gene regulatory networks, transcriptional regulatory networks, including how to infer transcriptional interactions among TF-Gene and among MicroRNA-mRNA, reveal combinatorial regulation mechanisms, and reconstruct TF activity profiles.

Yuehui Chen (University of Jinan)

Title: Microarray Data Classification using Flexible Neural Tree

Abstract: Microarray data are often extremely asymmetric in dimensionality, such as thousands or even tens of thousands of genes and a few hundreds of samples. Such extreme asymmetry between the dimensionality of genes and samples presents several challenges to conventional clustering and classification methods. The usually used methods for this kind of problem are extracting the informative genes firstly and then putting the extracted genes into a selected classifier. In this talk, a Flexible Neural Tree (FNT) model is proposed for informative gene selection and microarray data classification, simultaneously. Based on the pre-defined instruction/operator sets, a flexible neural tree model can be created and evolved. This framework allows input variables selection (feature extraction), over-layer connections and different activation functions for the various nodes involved. The FNT structure can be developed by using tree-structure based evolutionary algorithms, i.e., Genetic Programming (GP), Probabilistic Incremental Program Evolution (PIPE) etc., and the free parameters embedded in the neural tree can be optimized by any kinds of global search algorithms, i.e., Genetic Algorithm (GA), Particle Swarm Optimization (PSO) and so on. A comparison is also presented between the informative genes extracted by FNT model and the genes selected by other feature extraction methods. Experimental results show that the proposed method produces the better informative genes and recognition rates on four microarray benchmark datasets.

Katsuhisa Horimoto (National Institute of Advanced Industrial Science and Technology, Japan)

Title: Trace of Network Structure Changes by Two Approaches

Abstract: One of the most characteristic features in biological molecular network is that the network structure itself changes, depending on the cellular environments. Indeed, the activated molecules show variety with response to the distinctive cell conditions, and subsequently the network structures of active molecules also change. Here we present two approaches to trace the network structure change by using the microarray data. One is an application of the graphical chain model to the data measured in different cell stages, and another is an estimation of graph structure consistency with the measured data. The applications of the two approaches to the microarray data measured in the distinctive stages of liver cancer progression are presented, and their merits and pitfalls are discussed.

Tony Hu (Drexel University)

Title: Data Mining in Bioinformatics

Abstract: Despite an influx of molecular data in the form of sequences, structure, transcription profiles etc., most of the protein interaction information relevant to cell biology research still exists strictly in the scientific literature which is written in a natural language that computers cannot easily manipulate. Automatically mining and extracting information from biomedical text holds the promise of easily consolidating large amounts of biological knowledge in computer-accessible form. In this talk, we present a novel approach Bio-IEDM (Biomedical Information Extraction and Data Mining) to integrate text mining and predictive modeling to analyze biomolecular network from biomedical literature databases. Our method consists of two phases. In phase 1, we discuss a semi-supervised efficient learning approach to automatically extract biological relationships such as protein-protein interaction, protein-gene interaction from the biomedical literature databases to construct the biomolecular network. In phase 2, we present a novel clustering algorithm to analyze the biomolecular network graph to identify biologically meaningful subnetworks (communities). The clustering algorithm considers the characteristics of the scale-free network graphs and is based on the local density of the vertex and its neighborhood functions that can be used to find more meaningful clusters with different density level. The experimental results indicate our approach is very effective in extracting biological knowledge from a huge collection of biomedical literatures. The integration of data mining and information extraction provides a promising direction for analyzing the biomolecular network.

Michael Ng (Hong Kong Baptist University)

Title: SNP Markers Detection Method

Abstract: SKM-SNP, SNP markers detection method, is proposed to identify a set of relevant SNPs for the association between a disease and multiple marker genotypes. We employ a subspace categorical clustering algorithm to compute a weight for each SNP in the group of patient samples and the group of normal samples, and use the weights to identify the subsets of relevant SNPs that categorize these two groups. The experiment on a Parkinson disease data set containing genome-wide SNPs is reported to demonstrate the program.

Lei-Han Tang (Hong Kong Baptist University)

Title: Organization, Simplification, and Regulation of Metabolic Networks

Abstract: Through evolution living organisms have developed an elaborate network of enzyme-facilitated reactions and transport to process and cycle biochemical compounds for cell growth. The sheer complexity of such networks presents a great challenge to modellers. To prepare for systematic integration and dynamic modeling of experimental data on the metabolic flow and regulation under various growth conditions, we have developed a scheme to simplify the network based on relevant biochemical knowledge. Basic ideas of the simplification are reported here. Essentially, compounds that play only a facilitating role in the reactions (e.g., energy carrier, electron donor/acceptor, ammonia/phosphate/sulfate) are considered as horizontal nodes and are dealt with separately from the carbon flow which define the vertical links. The resulting network is essentially tree-like with only a few loops whose biochemical function can be clearly identified. Reactions at the branch points of the simplified network, which are often the entry points to major biochemical pathways, are key nodes for regulation. Examples of carrier compound regulation and end-product inhibition of pathways will be given to illustrate the utility of the simplified network.

Stephen Tsui (The Chinese University of Hong Kong)

Title: Construction of an Algorithm for the Prediction of Diabetic Nephropathy Using a Computational Approach

Limsoon Wong (National University of Singapore)

Title: Identifying Protein Complexes from Protein Interactome Maps

Abstract: Protein complexes are fundamental for understanding principles of cellular organizations. However, most protein interactome maps are still essentially an in vitro scaffold. Further these protein interactome maps contain a significant amount of noise interactions, as well as missing many real interactions. It is thus an important challenge to reliably deduce in vivo protein interactions and to identify membership in the same protein complexes. In this talk, we describe recent progress in computational techniques for protein complex prediction from noisy protein interaction network data.

Hong Yan (City University of Hong Kong)

Title: Spectral Estimation Methods for DNA Sequence and Microarray Data Analysis

Abstract: Spectral estimation techniques are widely used in signal processing systems. We can consider DNA sequences and microarray time series data as digital signals and use spectral estimation methods to extract latent oscillatory patterns in these data. This talk will present our recent work on parametric spectral analysis models and their applications to gene recognition, tandem repeat detection, microarray missing value estimation, and periodically expressed gene identification. We will demonstrate how to solve several difficult data analysis problems, including strong noise and weak signal, small time series length, non-uniform sampling and data distortions. Our methods will be compared with existing ones based on the experiment results obtained from a number of datasets.

Weichuan Yu (The Hong Kong University of Science and Technology)

Title: Optimization-Based Peptide Mass Fingerprinting for Protein Mixture Identification

Abstract: In current proteome research, the most widely used method for protein mixture identification is probably peptide sequencing. Peptide sequencing is based on tandem Mass Spectrometry (MS/MS) data. The disadvantage is that MS/MS data only sequences a limited number of peptides and leaves many more peptides uncovered. Peptide Mass Fingerprinting (PMF) has been widely used to identify single purified proteins from single-stage MS data. Unfortunately, this technique is less accurate than the peptide sequencing method and can not handle protein mixtures, which hampers the widespread use of PMF technique.

In this talk, we tackle the problem of protein mixture identification from an optimization point of view. We show that some simple heuristics can find good solutions to the optimization problem. As a result, we obtain much better identification results than previous methods. Through a comprehensive simulation study, we identify a set of limiting factors that hinder the performance of PMF-based protein mixture identification. We argue that it is feasible to remove these limitations and PMF can be a powerful tool in the analysis of protein mixtures, especially in the identification of low-abundance proteins which are less likely to be sequenced by MS/MS scanning.

Bo Zhang (Tsinghua University)

Title: Multi-granular Computing and Its Applications

Abstract: One of the basic characteristics in human problem solving is the ability to conceptualize the world at different granularities and translate from one abstraction level to the others easily. But so far computers can only deal with one abstraction level in problem solving generally. The aim of multi-granular computing is intended to endow the computers with the same capacity. This idea has widely been used to problem solving in Artificial Intelligence (AI) such as hierarchical planning and heuristic search. In the talk, we will introduce one of its basic theories, the quotient space based theory. In the theory, we will show how a problem (object) be represented at different grain-size worlds and how the multi-granular computing be used to reduce the computational complexity and enhance processing performances. In the application side, we will apply the model to both top-down problem solving and bottom-up information fusion. In information fusion, two basic problems are addressed, multi-granular/multimodal representation and fusion. Our related research works are introduced. We have participated in the annual TREC (Text Retrieval Conference) video retrieval evaluation (TRECVID) from 2005. Some evaluation results are presented as well. The initial granular computing theory we proposed was defined on equivalence relations. Then, we will extend it to tolerance relations and the future research direction is discussed.