DIMACS

Center for Discrete Mathematics & Theoretical Computer Science
Founded as a National Science Foundation Science and Technology Center

DIMACS Workshop on Clustering problems In BIological Networks (CIBIN)

May 9 – 11, 2006
DIMACS Center
Rutgers University, Busch Campus
Piscataway, New Jersey

Organized by:

Sergiy Butenko *(Texas A&M)*
W. Art Chaovalitwongse *(Rutgers)*
Panos M. Pardalos *(U of Florida)*

Program and Abstracts

National Science Foundation
WHERE DISCOVERIES BEGIN

CONFERENCE PROGRAM

TUESDAY, MAY 9

8:00 – 1:00 **Registration:** Conference Registration Desk

8:30 Opening Remarks from the DIMACS director and Conference Organizers

Session T1 Analyzing Metabolic and Protein Networks

Chair: Panos Pardalos

8:45 **Using Mathematical Programming for the Analysis and Redesign of Metabolic and Signaling Networks**
Costas Maranas

9:15 **Dynamic Modules in Metabolic Networks**
Eivind Almaas

9:45 **Analysis of Interaction Networks from Clusters of Co-Expressed Genes: A Case Study on Inflammation**
Ioannis Androulakis

10:15 – 10:45 **Coffee Break**

Session T2 Graph-Based Clustering Techniques

Chair: Art Chaovalitwongse

10:45 **Clique Relaxation Models of Clusters in Biological Networks**
Sergiy Butenko

11:15 **Maximal Hyperclique Pattern in Gene Profiles**
Weili Wu

11:45 **Practical Fixed-Parameter Algorithms for Graph-Modeled Data Clustering**
Sebastian Wernicke

12:15 – 1:45 **Lunch**

Session T3 Optimization Techniques for Biological Data Comparison

Chair: Sergiy Butenko

1:45 **Mathematical Programming Methods for Comparison Problems in Biocomputing**
Carlos Oliveira

2:15 **Using Formal Concept Analysis for Microarray Data Comparison**
Vicky Choi

2:45 **Stereotyped Activity Flow in Auditory Neocortical Microcircuits**
Kenneth Harris

3:15 – 3:45 **Coffee Break**

Session T4 Clusters and Dynamics of Biological Networks

Chair: Carlos Oliveira

3:45 **The Pure Parsimony Problem**
Allen Holder

4:15 **A Time Series Clustering Algorithm for Brain Network Analysis**
Art Chaovalitwongse

4:45 **Fast and Effective Clustering Very Large Networks Using Density-Based Clustering Algorithm**
Xiaowei Xu

WEDNESDAY, MAY 10

8:00 – 1:00 **Registration:** Conference Registration Desk

Session W1 Models and Algorithms for Genetic Expression Data

Chair: Sergiy Butenko

8:45 **Mining High-Throughput Biological Data: Methods, Algorithms and Applications**
Eytan Domany

9:15 **Clustering Algorithms for the Analysis of Type 1 Diabetes Data**
Michael Langston

9:45 **A Projected Clustering Algorithm for Biological Data Analysis**
Ping Deng

10:15 – 10:45 **Coffee Break**

Session W2 Coregulation in Metabolic Networks

Chair: Sungchul Ji

10:45 **Sequence-Based Predictive Modeling of Posttranscriptional Regulatory Networks**
Harmen Bussemaker

11:15 **Posttranscriptional Regulation within the Transcriptome of Human Tissues**
Gary Brewer

11:45 **Five Levels of Molecular Networks Underlying the Structure and Function of the Living Cell**
Sungchul Ji

12:15 – 1:45 **Lunch**

Session W3 Optimization-Based Clustering Techniques

Chair: Panos Pardalos

1:45 **A Novel Mixed-Integer Nonlinear Optimization-Based Clustering Approach: Global Optimum Search in Clustering with Enhanced Positioning (EP_GOS_Clust)**
Christodoulos Floudas

2:15 **Nonlinear Skeletons of Data Sets and Kernel Fuzzy Hyperplane Clustering Algorithm**
Pando Georgiev

2:45 **Consistent Biclustering via Fractional 0–1 Programming**
Stas Busygin

3:15 – 3:45 **Coffee Break**

Session W4 Clustering Microarray Data

Chair: Dhammika Amaratunga and Javier Cabrera

3:45 **Generalizations of the Topological Overlap Matrix for Module Detection in Gene and Protein Networks**
Steve Horvath

4:15 **Clustering as a Means of Identifying Co-Regulated Genes**
Jyotsna Kasturi

4:45 **Tuned Two-Way Bagging for Clustering Microarrays**
Vladimir Kovtun

5:15 **Data Analysis in Immune Response Profiling using Human Protein Microarrays**
Mariusz Lubomirski

Conference Dinner

7:30 **Conference Reception** - Dinner Banquet

Thursday, May 11 *(Joint with CSNA)*

8:00 – 1:00 **Registration:** Conference Registration Desk

Session H1 CSNA Invited Talks

Chair: Mel Janowitz

8:45 **Data Mining and Network Models of Massive Datasets**
Panos Pardalos

9:30 **Multi-Class Protein Classification Using String Kernels and Adaptive Codes**
Christina Leslie

10:15 – 10:45 **Coffee Break**

Session H2 Classification and Anomaly Detection in Biological Processes

Chair: Paul Kantor

10:45 **Comparison of Classification Methods to Predict Complications to Liver Surgery**
Leah Ben-Porat

11:15 **Self-Organizing Maps for Brain Electrical Activity Classification**
Wei Zeng

11:45 **Using Scan Statistics for Anomaly Detection in Genetic Networks**
Christopher Overall

12:15 – 1:30 **Lunch**

Session H3: Classification vs Clustering, Analyzing Gene Functionality

Chair: Claudia Perlich

1:30 **Protein Cluster Analysis via Directed Diffusion**
Yosi Keller

2:00 **Learning and Classification in Biological Data**
Sofus Macskassy

2:30 **Information-Based Clustering**
Gurinder Singh Atwal

3:00 **Classification vs Clustering, Analyzing Gene Functionality**
Claudia Perlich

3:30 – 3:45 **Coffee Break**

Session H4 Clustering and Sequencing Genomic Data

Chair: Panos Pardalos

3:45 **Using Cluster Analysis to Relate Subjective and Objective Pharmacovigilance Association Measures**
Ronald Pearson

4:15 **The Longest Ultraconserved Sequences and Evolution of Vertebrate Mitochondrial Genomes**
Weiwu Fang

4:45 **Distance Based Probabilistic Clustering of Data**
Cem Iyigun

5:15 **Closing Remarks from the Conference Organizers**

5:30 **Adjourn**

Conference Presentations

1. **Dynamic Modules in Metabolic Networks**
   Eivind Almaas (Lawrence Livermore National Laboratory)

2. **Analysis of Interaction Networks from Clusters of Co-Expressed Genes: A Case Study on Inflammation**
   Ioannis Androulakis (Rutgers University)

3. **Information-Based Clustering**
   Gurinder Singh Atwal (Princeton University)

4. **Comparison of Classification Methods to Predict Complications to Liver Surgery**
   Leah Ben-Porat (Memorial Sloan-Kettering Cancer Center)

5. **Posttranscriptional Regulation within the Transcriptome of Human Tissues**
   Gary Brewer (University of Medicine & Dentistry of New Jersey)

6. **Sequence-Based Predictive Modeling of Posttranscriptional Regulatory Networks**
   Harmen Bussemaker (Columbia University)

7. **Consistent Biclustering via Fractional 0–1 Programming**
   Stas Busygin (University of Florida)

8. **Clique Relaxation Models of Clusters in Biological Networks**
   Sergiy Butenko (Texas A&M)

9. **A Time Series Clustering Algorithm for Brain Network Analysis**
   Art Chaovalitwongse (Rutgers University)

10. **Using Formal Concept Analysis for Microarray Data Comparison**
    Vicky Choi (Virginia Tech)

11. **A Projected Clustering Algorithm for Biological Data Analysis**
    Ping Deng (University of Texas at Dallas)

12. **Mining High-Throughput Biological Data: Methods, Algorithms and Applications**
    Eytan Domany (Weizmann Institute of Science)

13. **The Longest Ultraconserved Sequences and Evolution of Vertebrate Mitochondrial Genomes**
    Weiwu Fang (Chinese Academy of Sciences, Beijing)

14. **A Novel Mixed-Integer Nonlinear Optimization-Based Clustering Approach: Global Optimum Search in Clustering with Enhanced Positioning (EP_GOS_Clust)**
    Christodoulos Floudas (Princeton University)

15. **Nonlinear Skeletons of Data Sets and Kernel Fuzzy Hyperplane Clustering Algorithm**
    Pando Georgiev (University of Cincinnati)

16. **Stereotyped Activity Flow in Auditory Neocortical Microcircuits**
    Kenneth Harris (Rutgers University)

17. **The Pure Parsimony Problem**
    Allen Holder (Trinity University)

18. **Generalizations of the Topological Overlap Matrix for Module Detection in Gene and Protein Networks**
    Steve Horvath (UCLA)

19. **Distance Based Probabilistic Clustering of Data**
    Cem Iyigun (Rutgers University)

20. **Five Levels of Molecular Networks Underlying the Structure and Function of the Living Cell**
Sungchul Ji (Rutgers University) - organizing a session Wednesday 10th morning

21. **Clustering as a Means of Identifying Co-Regulated Genes**
Jyotsna Kasturi (J&J Pharmaceutical Research & Development)

22. **Protein Cluster Analysis via Directed Diffusion**
Yosi Keller (Yale University)

23. **Tuned Two-Way Bagging for Clustering Microarrays**
Vladimir Kovtun (Rutgers University)

24. **Clustering Algorithms for the Analysis of Type 1 Diabetes Data**
Michael Langston (University of Tennessee)

25. **Multi-Class Protein Classification Using String Kernels and Adaptive Codes**
Christina Leslie (Columbia University)

26. **Data Analysis in Immune Response Profiling using Human Protein Microarrays**
Mariusz Lubomirski (J&J Pharmaceutical Research and Development)

27. **Using Mathematical Programming for the Analysis and Redesign of Metabolic and Signaling Networks**
Costas Maranas (Penn State University)

28. **Learning and Classification in Biological Data**
Sofus Macskassy (Fetch Technologies)

29. **Mathematical Programming Methods for Comparison Problems in Biocomputing**
Carlos Oliveira (Oklahoma State University)

30. **Using Scan Statistics for Anomaly Detection in Genetic Networks**
Christopher Overall (George Mason University)

31. **Data Mining and Network Models of Massive Datasets**
Panos Pardalos (University of Florida)

32. **Using Cluster Analysis to Relate Subjective and Objective Pharmacovigilance Association Measures**
Ronald Pearson (ProSanos Corporation)

33. **Classification vs Clustering, Analyzing Gene Functionality**
Claudia Perlich (IBM Research)

34. **Practical Fixed-Parameter Algorithms for Graph-Modeled Data Clustering**
Sebastian Wernicke (Institut f¨ur Informatik, Germany)

35. **Maximal Hyperclique Pattern in Gene Profiles**
Weili Wu (University of Texas at Dallas)

36. **Fast and Effective Clustering Very Large Networks Using Density-Based Clustering Algorithm**
Xiaowei Xu (University of Arkansas at Little Rock)

37. **Self-Organizing Maps for Brain Electrical Activity Classification**
Wei Zeng (Rutgers University)

Dynamic Modules in Metabolic Networks

Eivind Almaas
Microbial Systems Division
Lawrence Livermore National Laboratory

During the last few years, network approaches have shown great promise as a tool to both analyze and provide understanding of complex systems as disparate as the world-wide web and cellular metabolism. Much effort has been focused on characterizing topological properties of such systems. However, in order to develop detailed descriptions of complex networks, we need to look beyond their topology and incorporate dynamical aspects. The cellular metabolism, where nodes correspond to metabolites and links indicate chemical reactions, is an excellent model system where theoretical predictions can be compared with experimental results. I will present recent insights into the principles governing the modular utilization of the cellular metabolism. We find that, while most metabolic reactions have small fluxes, the metabolism's activity is dominated by an interconnected sub-network of reactions with very high fluxes. For the bacteria H. pylori and E. coli and the yeast S. cerevisiae, the metabolism responds to changes in growth conditions by reorganizing the rates of select reactions predominantly within this high-flux backbone. Furthermore, these networks are organized around the metabolic core -- a set of reactions that are always in use.

Strikingly, the activity of the metabolic core reactions is highly synchronized, and the core reactions are significantly more essential and evolutionary conserved than the non-core ones.

Analysis of Interaction Networks from Clusters of Co-Expressed Genes: A Case Study on Inflammation

A. Misra, T.J. Maguire, and I.P. Androulakis*
Department of Biomedical Engineering
Rutgers University

Extracting biological insight from high-throughput genomic studies of human diseases remains a major challenge, primarily due to our inability to recognize, evaluate and rationalize the relevant biological processes recorded from vast amounts of data.

We will discuss an integrated framework combining fine-grained clustering of temporal gene expression data, selection of maximally informative clusters, based of their ability to capture the underlying dynamic transcriptional response, and the subsequent analysis of the resulting network of interactions among genes in individual clusters. The latter are developed based on the identification of common regulators among the genes in each cluster through mining literature data.

We characterize the structure of the networks in terms of fundamental graph properties, and explore biologically the implications of the scale-free character of the resulting graphs. We demonstrate the biological importance of the highly connected hubs of the networks and show how these can be further exploited as targets for potential therapies during the early onset of inflammation. We conclude by identifying two possible challenges in network biology, namely, the nature of the interactions and the potentially limited information content of the temporal gene expression experiments, and discuss expected implications.

Information-Based Clustering

Gurinder Singh Atwal
Department of Physics, Lewis Sigler Institute for Integrative Genomics
Princeton University

Existing clustering methods in computational biology implicitly invoke several nontrivial assumptions about the structure of data. Thus the strength of a particular clustering technique depends on how well these assumptions match the true generative model of the data. Here, we address the clustering problem from an information theoretic perspective that avoids many of these assumptions. In particular, we motivate a cost function expressing the tradeoff between the average intra-cluster similarity and the compression of the data. Our formulation obviates the need for defining a cluster "prototype," does not require an a priori similarity metric, is invariant to changes in the representation of the data, and naturally captures nonlinear relations. We also address the mathematical problems associated with extracting information theoretic quantities from finitely sampled biological datasets. We apply this approach to different domains, including the yeast stress response module microarray expression profiles, and find that it consistently produces clusters that are more coherent than those extracted by existing algorithms. Finally, our approach provides a way of clustering based on collective notions of similarity rather than the traditional pairwise measures.

Comparison of Classification Methods to Predict Complications to Liver Surgery

Leah Ben-Porat*, Mithat Gonen, and William Jarnigan
Memorial Sloan-Kettering Cancer Center
New York, NY

Often in medical practice, it is of interest to identify patients that are at high risk for complications or death. As a result, prediction models are useful tools in medical decision making. There are many statistical methods available to build prediction models. The aim of this study is to compare several prediction methods using a large institutional database of patients undergoing liver surgery. The database under study includes 2002 consecutive patients who underwent liver resection at Memorial Sloan Kettering Cancer Center from 1991 to 2002 and captures information on more than thirty preoperative risk variables. The classification models were built to predict high grade complications following surgery. The strategies employed were logistic regression, generalized additive models, and classification trees. Within each strategy several approaches were explored. Graphical representations of the models were constructed so as to make the models user friendly for clinicians. All models were developed on a training set and evaluated on the test set. The performance of the models was compared by analyzing the ROC curves from the test set. The stepwise logistic regression model had similar predictive accuracy to the generalized additive model and the classification trees. Other non-linear machine learning methods such as neural networks will be applied to this dataset to determine their predictive power.

Posttranscriptional Regulation within the Transcriptome of Human Tissues

Gary Brewer

Department of Molecular Genetics, Microbiology, and Immunology
University of Medicine & Dentistry of New Jersey – Robert Wood Johnson Medical School

The 3'-untranslated regions of messenger RNAs (mRNAs) encoding many regulatory proteins contain sequence elements, known as A+U-rich elements or AREs, that promote rapid mRNA degradation both constitutively and in response to external stimuli. This process serves to limit levels of these mRNAs and hence proteins. We previously identified and molecularly cloned an ARE-binding factor, AUF1, which targets an ARE-bearing mRNA for degradation. Messenger RNA-binding proteins, such as AUF1, associate with mRNAs to form ribonucleoprotein particles (mRNPs). Microarray analyses have revealed that RNA-binding proteins associate with unique subsets of mRNAs to define posttranscriptional operons to coordinate expression of groups of mRNAs. Coordinately regulated mRNAs share untranslated sequence elements for regulation (USERs). The AUF1:ARE association fits nicely into this posttranscriptional operon model, since AUF1 serves as a linchpin to coordinate ARE-dependent gene expression, probably in conjunction with other RNA-binding proteins. Thus, AREs represent a family (or families) of USERs. Our hypothesis is that AUF1:mRNA associations within a tissue define posttranscriptional operons and that expression of these operons is altered in disease conditions. To define the landscape of AUF1 target mRNAs within specific tissues, we developed a procedure that utilizes AUF1 as an affinity ligand for purification of AUF1 target mRNAs from complex populations. This procedure combined with microarray analyses has permitted a global view of the transcriptome regulated by this key RNA-binding protein.

Sequence-Based Predictive Modeling of Posttranscriptional Regulatory Networks

Harmen Bussemaker
Department of Biological Sciences
Columbia University

It is the dynamic balance between transcription from DNA to messenger RNA and subsequent mRNA degradation that determines the steady-state mRNA abundance for each gene in a genome. However, while regulation of transcription rate by DNA binding transcription factors has been intensively studied, both experimentally and computationally, regulation of the transcript turnover rate by RNA binding factors has received far less attention. We took advantage of the fact that information about the condition-specific activity and sequence-specific affinity of RNA binding regulatory factors is implicitly represented in the steady-state mRNA abundances measured using DNA microarrays. Thus, by fitting a model based on a physical description of molecular interactions, we were able to gain quantitative insight into the mechanisms that underlie genome-wide regulatory networks. We developed a novel algorithm, MatrixREDUCE, that predicted the sequence-specific binding affinity of several known and unknown RNA-binding factors and their condition-specific activity, using only genomic sequence data and steady-state mRNA expression data as input.

We identified and computationally characterized the binding sites for six mRNA stability regulators in the yeast S. cerevisiae, which include two known RNA-binding proteins, Puf3p and Puf4p. We provide computational and experimental evidence that regulation of mRNA stability by the discovered factors is dynamic and responds to a variety of environmental stimuli. For example, little was previously known about the functional role of Puf3p, but our computational results suggest that Puf3p functions to destabilize mitochondrion-related transcripts when metabolite repressing sugars are present and in response to the drug rapamycin. We were able to experimentally confirm these predictions by growing a transformed strain expressing a hybrid mRNA designed to contain a functional Puf3p binding site in different culture conditions and measuring its half-life after a transcriptional shut-off.

Our work suggests that regulation of mRNA stability is not a special case phenomenon, but rather a pervasive regulatory mechanism that rapidly adapts cellular processes to a changing environment.

Consistent Biclustering via Fractional 0–1 Programming

Stas Busygin*, Oleg A. Prokopyev, Panos M. Pardalos
Department of Industrial and Systems Engineering
University of Florida

Biclustering consists in simultaneous partitioning of the set of samples and the set of their attributes (features) into subsets (classes). Samples and features classified together are supposed to have a high relevance to each other which can be observed by intensity of their expressions. We define the notion of consistency for biclustering using interrelation between centroids of sample and feature classes. We prove that consistent biclustering implies separability of the classes by convex cones.

We discuss both supervised and unsupervised learning algorithms, where the biclustering consistency is achieved by feature selection and outlier detection. The developed models involve solution of a fractional 0–1 programming problem. Encouraging computational results on DNA microarray data mining problems are reported.

Clique Relaxation Models of Clusters in Biological Networks

Sergiy Butenko*, Balabhaskar Balasundaram, and Svyatoslav Trukhanov
Department of Industrial and Systems Engineering
Texas A&M University

We introduce and study several clique relaxation models that originally appeared in social network analysis, but can also be used in several other important application areas, including the analysis of biological networks. We establish computational complexity results of the problems on arbitrary and restricted graph classes. Integer programming formulations are presented and basic polyhedral study of the problems is carried out. A branch-and-cut implementation is discussed and computational test results are provided. In particular, the results of application of proposed algorithms to biological networks are discussed.

A Time Series Clustering Algorithm for Brain Network Analysis

W. Art Chaovalitwongse
Department of Industrial and Systems Engineering
Rutgers University

We propose a novel optimization-based time series clustering algorithm for the connectivity analysis of brain networks. Specifically, we attempt to extract insightful characteristics of the brain connectivity from abnormal (epileptic) brains. The electroencephalogram (EEG) time series data from normal and abnormal (pre-seizure) states are analyzed. The experimental results from this study suggest that the proposed clustering algorithm may be able to differentiate normal and pre-seizure EEGs.

Using Formal Concept Analysis for Microarray Data Comparison

Vicky Choi[1]*, Yang Huang[2], Vy Lam[3], Dustin Potter[3], Reinhard Laubenbacher[3], and Karen Duca[3]

[1]*Department of Computer Science, Virginia Tech*
[2]*Department of Computer Science, Rutgers University*
[3]*Virginia Bioinformatics Institute, Virginia Tech*

Microarray and other high-throughput technologies have become common research methods in the life sciences. Tools to rapidly compare such data for biological discovery are needed. In this talk, we propose using Formal Concept Analysis (FCA) for this purpose. The method of FCA builds a (concept) lattice from the experimental data together with additional biological information. For microarray data, each vertex of the lattice corresponds to a subset of genes that are grouped together according to their expression values and some biological information relating to gene function. The lattice structure of these overlapping gene sets might reflect biological relationships in the dataset. Similarities and differences between experiments can then be investigated by comparing their corresponding lattices according to various graph measures. To this end, we have developed an efficient algorithm for building concept lattices and several measures for comparing lattices. We will show preliminary results applying our method to microarray data derived from influenza infected mouse lung tissue and healthy controls.

A Projected Clustering Algorithm for Biological Data Analysis

Ping Deng* and Weili Wu
Department of Computer Science
The University of Texas at Dallas

Projected clustering is designed for clustering data in high dimensional space. Since data may not be correlated in all dimensions, it is more meaningful to consider clusters in subspaces of full dimensions. Recently, several projected clustering algorithms that focus on finding specific projection for each cluster have been proposed. We find that, besides the distance, the closeness of points in different dimensions also depends on the distributions of data along those dimensions. Based on this, we propose a projected clustering algorithm, IPROCLUS (Improved PROCLUS), which is efficient and accurate in handling data in high dimensional space. We apply our algorithm on biological data and analyze the result.

A Projected Clustering Algorithm for Biological Data Analysis

Eytan Domany
Department of Physics of Complex Systems and Center for Systems Biology
Weizmann Institute of Science, Israel

DNA chips are novel experimental tools that have revolutionized research in molecular biology and generated considerable excitement. A single chip allows simultaneous measurement of the level at which thousands of genes are expressed. A typical experiment uses a few tens of such chips, each devoted to one sample - such as material extracted from a tumor. Hence the results of such an experiment consist of a table, of several thousand rows (one for each gene) and 50 - 100 columns (one for each sample). Extracting relevant information from such a large, complex and noisy data set requires development of novel methods of analysis.

I will briefly demonstrate how we combine standard statistical analysis with novel unsupervised methods (clustering[1], bi-clustering[2] and sorting[3]) to mine expression data obtained from leukemia[4], cervical[5] and colon cancer patients. If time permits, I will describe a novel experimental tool - antigen chip – and the manner, in which it can be used to predict whether an individual (mouse) will or will not become diabetic[6].

1. Blatt et al, Physical Review Letters 76, 3251 (1996)
2. Getz et al, PNAS 97, 12079 (2000).
3. Tsafrir et al, Bioinformatics 21, 2301 (2005)
4. Rozovskaia et al, PNAS 100, 7853 (2003)
5. Rosty et al, Oncogene 24, 6367 (2005)
6. Quintana et al, PNAS 101, 14615 (2004)

The Longest Ultraconserved Sequences and Evolution of Vertebrate Mitochondrial Genomes

Weiwu Fang[1]*, Sili Tu[1], Xu Cai[1], Yuncheng Wang[2], and Weixing Wu[3]

[1]Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China
[2]Shandong Agricultural University, Taian, China
[3]University of International Business and Economics, Beijing, China

Ultraconserved sequences and small RNA in genomes often play a special role in life and evolution of many species. We compared 753 genomes of bacteria, archaea, and mitochondria (more than 540M data) and found four unique ultraconserved sequences in 352 vertebrate mitochondrial genomes which are the longest or second longest or third longest ultraconserved subsequences in the vertebrate mitochondrial genomes, their lengths are about 22 -- 30 bp similar to those of small RNA , and only one copy can be found in each genome.

This is the first report on discovery of ultraconserved sequences in the genomes of so many different species of animals. Surprisingly, the classification and evolution relationship among some high-level categories of animals can be clearly reflected by their regularity of occurrence; moreover, these findings gave rise to some new ideas of evolution of mitochondria and living beings. For instance, the variations in mitochondrial genomes of animals may help clarify the evolution relationship between Aves and Reptile, and understand the fact that the origin of mitochondrion is at least not a simple copy of genomes of lower living things such as bacteria and archaea.

Finally, some classification and computation skills are also discussed.

A Novel Mixed-Integer Nonlinear Optimization-Based Clustering Approach: Global Optimum Search in Clustering with Enhanced Positioning (EP_GOS_Clust)

Meng Piao Tan and Christodoulos A. Floudas*
Department of Chemical Engineering, Princeton University
Princeton, NJ 08544-5263, USA

Cluster analysis of genome-wide expression data from DNA microarray hybridization studies has proven to be a useful tool for identifying biologically relevant groupings of genes, which can lead to interesting insights. Patterns seen in genome-wide expression experiments can give indications about unknown regulatory elements. Also, since genes with similar functions cluster together, grouping genes of known functions with poorly characterized genes may provide a simple means of gaining understanding into the functions of these uncharacterized genes. It is hence important to apply a rigorous yet intuitive clustering algorithm to uncover these genomic relationships. However, several of the popularly-used clustering algorithms demonstrate an uncomfortable level of sensitivity to the initialization point, as well as a broad level of latitude accorded to the user with regards to the optimal number of clusters. Furthermore, the quality of such clustering algorithms in finding groupings of data with the tightest possible clustering raises a number of issues.

In this presentation, a novel clustering algorithm framework is introduced. It is based on a variant of the Generalized Benders Decomposition, denoted as the Global Optimum Search, which includes a procedure to determine the optimal number of clusters to be used. As an investigative study, the proposed algorithm is applied to experimental DNA microarray data centered on the Ras signaling pathway in the yeast Saccharomyces Cerevisiae. The clustering results are compared to that obtained with existing popular clustering algorithms. The proposed approach outperforms these algorithms in both the areas of intra-cluster similarity and inter-cluster dissimilarity, often considered as the two key tenets of clustering. The proposed algorithm's implementation is also structured to expedite the solution for the determination of the optimal number of clusters.

Nonlinear Skeletons of Data Sets and Kernel Fuzzy Hyperplane Clustering Algorithm

Pando Georgiev* and Anca Ralescu
Department of Electrical & Computer Engineering and Computer Science
University of Cincinnati

We define a nonlinear skeleton of a data set (with respect to a given kernel) as a union of hyper-surfaces (defined by this kernel), which best approximates the given data set. We develop a kernel fuzzy hyper-plane clustering algorithm for finding such a skeleton. The resulting algorithm has a direct application to nonlinear blind signal separation problem, as well as to a nonlinear representation of the data set in terms of superposition of linear and nonlinear mappings, both defined by the given kernel. We consider two criteria for such representation: statistical independence and sparseness. They correspond, respectively, to a linear Independent Component Analysis problem and a linear Sparse Representation problem in a feature space. Some applications to data sets from biomedicine are considered.

Stereotyped Activity Flow in Auditory Neocortical Microcircuits

Kenneth Harris
Quantitative Neuroscience Laboratory, Center for Molecular and Behavioral Science
Rutgers University

Information is processed in the cortex by the parallel action of large numbers of neurons. To study the structure of ensemble activity during sensory stimulation, we recorded auditory cortical populations (40-100 cells) while presenting simple stimuli (tones, noise, and clicks) in a passive listening paradigm. We observed a diversity of stimulus tuning and temporal response profiles, even amongst neurons recorded on a single electrode. Neural firing rates varied substantially with stimuli; however, differences between the temporal structures of responses evoked by multiple stimuli in a single neuron were small, compared to differences between the temporal structures of multiple cells to a single stimulus.

At the population level, diverse yet reliable onset latencies revealed a stereotyped spread of activation through the recorded population. This activation sequence was similar across stimuli, and also for spontaneously occurring patterns associated with the start of cortical UP states. To investigate the consequences of this temporal structure for stimulus coding, we performed a population vector analysis. Population codes evolved with time, characterized by increasing sparseness and orthogonalization during the first few hundred milliseconds of stimulus presentation. We hypothesize that our observations may reflect the interplay of recurrent network activity and diverse cellular physiologies.

The Pure Parsimony Problem

P. Blain, A. Holder*, C. Davis, J. Silva, and C. Vinzant
Department of Mathematics
Trinity University

Haplotyping is the process of reconstructing the genetic information donated by a prior generation to form a current population and is important because it allows us to find genetic markers that describe a population's susceptibility to diseases. We address the pure parsimony problem, which is to find the minimum amount of genetic diversity in the previous generation that is capable of explaining the current population. Numerical techniques have met with limited success, and instead of attacking the problem computationally, we cast the problem into the realm of graph theory. This leads to the study of diversity graphs and a mathematical pursuit that highlights the underlying structure of the biological problem.

Two important results are discussed. First, we provide a characterization theorem that describes whether or not a graph is capable of representing an actual biological model. One of the characterizations shows that the pure parsimony problem is an instance of a list coloring problem. Such problems are currently being studied heavily by combinatorialists. Second, we show that in special situations the biology problem has attractive solutions.

Generalizations of the Topological Overlap Matrix for Module Detection in Gene and Protein Networks

Steve Horvath
Departments of Human Genetics and Biostatistics
University of California, Los Angeles

Systems biologic studies of gene and protein interaction networks have found that these networks are comprised of `modules' (groups of tightly interconnected nodes). Module identification is an essential step towards understanding the whole network architecture. Here we will focus on module identification methods that are based on using a node dissimilarity measure in conjunction with a clustering method. More specifically, we introduce a general class of node dissimilarity measures based on the notion of “topological” overlap, which has been found to be biologically meaningful in several applications. The resulting generalized topological overlap measure (GTOM) generalizes the standard topological overlap measure (TOM) introduced by Ravasz et al (2002). Specifically, the m-th order version of this family is constructed by (i) counting the number of m-step neighbors that a pair of nodes share and (ii) normalizing it to take a value between 0 and 1.

Next, we will discuss the multi-point topological overlap matrix for neighborhood analysis. We extend the traditional topological overlap measure which measures pair-wise relationship in network analysis, to multi-point TOM (MTOM), which reveals the relationship among multiple points directly. We show how this measure can be used to define the neighborhood of a candidate set of genes. We illustrate the use of these methods using yeast and cancer network data.

Distance Based Probabilistic Clustering of Data

Cem Iyigun and Adi Ben-Israel
RUTCOR – Rutgers Center for Operations Research
Rutgers University

The problem is to partition a given data set $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\} \subset \mathbb{R}^n$, into clusters $\{\mathcal{C}_i, \ldots, \mathcal{C}_m\}$, where the number m of clusters is either given, or to be determined by an optimality criterion.

Clusters consist of similar points, and are themselves dissimilar, where similarity is in a sense of a distance $d(\cdot, \cdot)$ on $\mathbb{R}^n$.

With a cluster $\mathcal{C}_i$, we associate a **center** $\mathbf{c}_i$, and for any data point $\mathbf{x} \in \mathcal{D}$ we then compute:
- a **distance** $d(\mathbf{x}, \mathbf{c}_i)$, denoted by $d_i(\mathbf{x})$, and
- a **probability** $p_i(\mathbf{x})$ of membership in $\mathcal{C}_i$.

We assume throughout that for all $\mathbf{x}$,

$$p_i(\mathbf{x})\, d_i(\mathbf{x}) = \text{constant, depending on } \mathbf{x}, \quad \text{for } i = 1, \ldots, m, \tag{1}$$

making membership in nearby clusters more probable. Since probabilities add to 1, assumption (1) implies

$$p_i(\mathbf{x}) = \frac{\prod_{j \neq i} d_j(\mathbf{x})}{\sum_{k=1}^{m} \prod_{j \neq k} d_j(\mathbf{x})}\ ,\ \ i = 1, \ldots, m\ , \tag{2}$$

in particular, for $m = 2$,

$$p_1(\mathbf{x}) = \frac{d_2(\mathbf{x})}{d_1(\mathbf{x}) + d_2(\mathbf{x})},\ p_2(\mathbf{x}) = \frac{d_1(\mathbf{x})}{d_1(\mathbf{x}) + d_2(\mathbf{x})}\ . \tag{3}$$

We present a new clustering algorithm that iterates on centers, distances and probabilities, compare it with existing methods, and illustrate its advantages.

Five Levels of Molecular Networks Underlying the Structure and Function of the Living Cell

Sungchul Ji
Department of Pharmacology and Toxicology
Rutgers University

Available experimental data on the living cell can be organized in terms of the 5 distinct classes of networks that are coupled in a hierarchical manner. There are two kinds of networks – “in-plane” and “between-plane” networks. The nodes in the i^th plane are connected not only to other nodes in the same plane but also the nodes located in the (i -1)^th and (i+ 1)^th planes in 1-to-1, 1-to-many, or many-to-many modes. The 5 classes of networks operating in the living cell are (i) genetic networks (GNs), (ii) mRNA networks (RNs), (iii) protein networks (PNs), (iv) biochemical networks (BNs), and (v) functional modular networks (FMNs). The first three networks are well known. The BNs (i.e., the various metabolic pathways such as glycolytic pathway, Krebs cycle, respiratory chain, oxidative phosphorylation pathway, apoptotic pathway, etc.) have been well worked out in the second half of the last century. The concept of ‘functional modular networks (FMNs)’ is invoked here for the first time as a means to integrate the lower-order molecular networks. It is postulated that the unique properties and behaviors of the living cell are direct consequences of the dynamics of interactions between FMNs and extracellular environment. For example, when budding yeast grown in the presence of glucose is suddenly exposed to a new growth medium in which glucose is replaced by galactose, the glycolysis biochemical pathway is suppressed (which is known as the Pasteur effect) and the respiratory biochemical pathway is activated (known as the Crabtree effect) in order to survive in the new nutritional environment. This characteristic response of the budding yeast to glucose-galactose shift (mediated by the interaction between glycolytic and respiratory functions) represents a 2-node FMN. It appears highly likely that other cellular responses elicited by more complex environmental stimuli (e.g., heat, dehydration, etc.) will activate FMNs that consist of more than two nodes.

Since the introduction in the mid-1990’s of the revolutionary DNA microarray technique into cell biology, many investigators have been using the method to measure genome-wide mRNA levels in whole cells. In INTERPRETING these mRNA levels, most workers assumed that mRNA level (TL, or transcript level, for short) changes accurately reflected changes in the rates of the expression of the genes (TR, or transcription rates) encoding the measured mRNA molecules. This is tantamount to assuming that RNs and GNs described above are superimposable or that all of the edges of the ‘between-plane’ network that is sandwiched by the gene and mRNA planes are vertical. However, when TL and TR are measured simultaneously from budding yeast under glucose-galactose shift [Garcia-Martinez et al., Mol. Cell 15:303-313 (2004)], it was found that TL and TR are linearly correlated only 52% of the time (over the period of 850 minutes of observation). Furthermore TL in fact decreased about 30% of the time even though TR increased, and TL increased 10% of the time despite the fact that TR decreased. To account for these findings, a simple mathematical equation was derived expressing TL as a function of TR and TD (transcript degradation rates), which, when applied to the experimentally measured TL and TR values, allowed us to calculate the corresponding TD values for glycolytic and respiratory mRNA molecules for the first time. The results demonstrate that TL values are controlled by the ratio of TR and TD, and not by TR alone, as has been widely assumed in the past.

Clustering as a Means of Identifying Co-Regulated Genes

Jyotsna Kasturi
Johnson & Johnson Pharmaceutical Research & Development

One of the main goals of post-genomic molecular biology is the systematic discovery of the biological mechanisms underlying the behavior of all genes in the human genome. High-throughput technologies such as microarrays have made it possible to study the behavior of thousands of genes and proteins simultaneously. Genes with similar expression profiles are often involved in similar cellular and biological processes, controlled by common regulators. Unsupervised cluster analysis methods provide a means for inferring functional relationships between genes by identifying cohorts based on their expression similarity. Biological data is complex – typically high dimensional, in small sample sizes, containing experimental noise, etc.; it poses new challenges to researchers analyzing such data. Further downstream analyses, such as the identification of common regulatory elements are directly impacted by the quality of the clusters obtained. A new way to measure expression profile similarities in noisy data through smoothing and the use of an information-theoretic similarity measure is presented. Further, incorporating other complementary sources of information into the clustering can aid in the identification of co-regulated genes. A clustering method based on the Self-Organizing Map algorithm, to simultaneously cluster data from diverse sources is also presented.

Protein Cluster Analysis via Directed Diffusion

Yosi Keller*, Stephane Lafon, and Michael Krauthammer
Department of Applied Mathematics
Yale University

Graph-theoretical approaches are useful for elucidating the modular compositions of protein-protein interaction networks, which are known to consist of regions of increased network connectivity (clusters) corresponding to known molecular complexes or functional pathways. In this work, we introduce the concept of local spectral search as a graph-based methodology for cluster analysis.

Based on a set of known samples within a target set, we aim to identify the complete target set. We derive both an expansion scheme (form the known set of samples tot the entire set) and a rigorous clustering criterion that allows us to identify the target cluster. We apply the proposed scheme to a protein interaction network, where we infer the set of proteins related to a particular function based on a small number of proteins in that set.

Tuned Two-Way Bagging for Clustering Microarrays

Vladimir Kovtun[1]*, Dhammika Amaratunga[2], and Javier Cabrera[1]
[1] *Rutgers University*
[2] *Johnson & Johnson Pharmaceutical Research & Development*

Classification methods, both supervised and unsupervised, are playing a crucial role in interpreting data from high throughput functional genomics experiments. The focus of this talk is unsupervised classification, which, in the microarray literature, has manifold applications: to discover novel subclasses among the samples, to assess proof of concept, to extract features of interest, to assess sample quality, and so on. DNA microarray data is characterized as having a huge number of variables and only a few samples. It has been demonstrated that, when working with data having such features, bagging (bootstrap aggregation) is a useful concept. In this talk, we will give an overview of microarray clustering, outline briefly the most commonly-used methods for clustering microarrays, and introduce a new clustering method that is based on tuning a bagging process and that greatly improves the resultant classification.

Clustering Algorithms for the Analysis of Type 1 Diabetes Data

John D. Eblen[1], Ivan C. Gerling[2], Michael A. Langston[1]*, Arnold M. Saxton[1] and Jay R. Snoddy[3]

[1] *University of Tennessee, Knoxville, TN 37996*
[2] *University of Tennessee Health Science Center, Memphis, TN 38163*
[3] *Vanderbilt University Medical Center, Nashville, TN 37235*

Combinatorics and graph theory have proven to be effective tools for analyzing a wide variety of biological data. These tools are employed and enhanced in the context of high throughput data taken from non-obese diabetic mice in order to improve our understanding of the causative mechanisms of type 1 diabetes. The recently-introduced paraclique algorithm [1] is applied in an effort to handle noise and elucidate putatively co-regulated genes. A variety of parameter values and correlation thresholds are studied. Dual thresholds are also considered in an effort to combine transcriptomic with proteomic data, a challenging synthesis problem of widespread potential significance.

[1] E. J. Chesler and M. A. Langston, Combinatorial Genetic Regulatory Network Analysis Tools for High Throughput Transcriptomic Data, Proceedings, RECOMB Satellite Workshop on Systems Biology and Regulatory Genomics, San Diego, 2005.

Multi-Class Protein Classification Using String Kernels and Adaptive Codes

Christina Leslie
Center for Computational Learning Systems and Center for Computational Biology and Bioinformatics
Columbia University

One of the central problems in computational biology is the classification of protein sequences into functional and structural families based on sequence homology. Traditional approaches to this task include alignment-based algorithms and probabilistic models like profile hidden Markov models for protein families. However, these methods are less successful for remote homology detection, where one wants to predict a structural relationship between sequences that have diverged over a long evolutionary distance.

Our approach to protein remote homology detection and fold recognition is to use modern discriminative methods from machine learning, namely support vector machine classifiers (SVMs), combined with kernels specialized for biological sequence data. To this end, we introduce novel and efficient-to-compute string kernels that incorporate biologically motivated notions of inexact string matching, based on shared approximate occurrences of short subsequences ("k-mers").

Extending these ideas to take advantage of abundant unlabeled data -- large databases of protein sequences whose structural class is unknown -- we then define cluster kernels and profile-based string kernels that outperform all currently available remote homology detection methods. In the case of profile kernels, we can also interpret the SVM classifier by extracting discriminative motif regions that suggest conserved structural subunits in a protein superfamily.

Finally, to deal with the large underlying multi-class problem -- there are hundreds of protein folds and over 1000 superfamilies in widely used structural classification systems -- we introduce an adaptive code approach for learning to weight prediction scores from different binary SVM classifiers. Our adaptive code learning approach significantly outperforms 1-vs-all for this problem and also allows us to use optimize with respect to different loss functions.

We are building and will soon release a full multi-class fold prediction server, called SVM-fold, which will incorporate both our string kernel and our adaptive code learning work.

Data Analysis in Immune Response Profiling using Human Protein Microarrays

Mariusz Lubomirski[1]*, Dhammika Amaratunga[1], Javier Cabrera[2], and Stan Belkowski[1]
[1] Johnson & Johnson Pharmaceutical Research & Development
[2] Rutgers University

DNA microarrays are well known and established technology in pharmaceutical research industry providing variety of essential information for drug development. Protein microarrays are fast emerging as a follow up technology and they begin to experience rapid growth as the challenges in protein to spot methodologies are overcome. Like DNA microarrays, their protein counterparts produce large amounts of data which must be suitably analyzed in order to yield meaningful targets. We describe statistical methodologies applied to the data from the immune response profiling assay using Human Protein Microarrays.

Learning and Classification in Biological Data

Sofus Macskassy
Fetch Technologies, Inc.

Biological data are often relational in nature and recent machine learning techniques in relation learning and network classification are therefore applicable to such data. In this talk, I explore classifying biological data using only the implicit and explicit relations between the instances, i.e. the network structure and ignoring any other attributes of the objects. I will introduce the concept of network learning and introduce NetKit, a network learning toolkit for statistical relational learning, and show its use on the biological data in a network classification framework.

Using Mathematical Programming for the Analysis and Redesign of Metabolic and Signaling Networks

Costas D. Maranas
Department of Chemical Engineering
The Pennsylvania State University

In this talk we will describe how optimization-based computational tools can be used to guide microbial strain redesign leading to targeted overproductions of desired chemical products. Using as a starting point stoichiometric models of metabolism, we will first explore how optimization can be used to pinpoint which new functionalities to add into a microbial production host to endow it with new capabilities extracted from a generated database of more than 5,700 reactions. Conversely, we will describe how to identify gene deletions leading to the coupling of growth with the production of the desired chemical product. Finally we will explore how optimization can be used to analyze the topological properties of metabolic networks, identify gaps and suggest ways of filling them.

A similar mathematical framework will also be briefly described for elucidating the input-output structure of signaling networks and for pinpointing targeted disruptions leading to the silencing of undesirable outputs in therapeutic interventions. The frameworks are demonstrated on a large-scale reconstruction of a signaling network composed of nine signaling pathways implicated in prostate cancer. The developed computational tools will be highlighted using a number of design case-studies and the predictions will be contrasted with experimental results.

Mathematical Programming Methods for Comparison Problems in Biocomputing

Carlos Oliveira
School of Industrial Engineering and Management
Oklahoma State University

Mathematical programming is a modeling and solving technique that has been used with success in several areas of application. More recently, the considerable increase of importance of genetics and biocomputing has provided new possibilities of application for linear programming concepts. In this talk we discuss formulations for comparison problems, such as the closest string problem, the furthest string problem, and the phylogenetic tree analysis problem. Our methods are illustrated using LP implementations that can be solved in reasonable time for the instances tested.

Using Scan Statistics for Anomaly Detection in Genetic Networks

Christopher C. Overall[1]*, Jeffrey L. Solka[2], Jennifer W. Weller[3], and Carey E. Priebe[4]

[1]George Mason University, Fairfax, VA
[2]Naval Surface Warfare Center (NSWC), Dahlgren, VA
[3]Virginia Polytechnic Institute, Blacksburg, VA
[4]Johns Hopkins University, Baltimore, MD

We have applied graph-based scan statistics to genetic networks in order to detect anomalies in gene activity over time. An anomaly may manifest as a single gene with excessive connections to other genes or it may manifest as chatter, in which a neighborhood of genes centered about a particular gene exhibit excessive connections with one another.

Scan statistics are commonly used in image analysis to detect the presence of an anomalous local signal in a time-series of images. Based upon these techniques, Priebe et al. (2005) developed a theory of scan statistics for graphs and then applied the new methodology to detect anomalies in a time-series social network that they developed from the publicly available Enron email dataset.

The current work has been adapted from that of Priebe et al. (2005) and applied to datasets generated from time-series transcription profiling experiments, such as those using microarrays and reverse transcription-coupled PCR. Timeseries gene expression datasets are typically represented as a matrix in which each column corresponds to a time point and each row corresponds to a gene. Hence, each element of the matrix represents the expression value of a single gene at a single time point. Prior to applying the graph-based scan statistic methodology to a gene expression dataset, each of the time points are transformed into a graph with the genes as vertices and an edge drawn between two genes if there is a relationship between them. Either univariate or multivariate model-based clustering is applied to the gene expression values at a particular time point and the posterior probabilities are then used to determine relationships. After each time point is transformed into a graph, graph-based scan statistics are then utilized to detect any anomalous gene activity. We applied this methodology to the following two datasets for analysis: the yeast Saccharomyces cerevisiae cell cycle dataset (microarray) from Spellman et al. (1998) and the rat central nervous system development dataset (reverse transcription-coupled PCR) of Wen et al. (1998).

Data Mining and Network Models of Massive Datasets

Panos Pardalos
Department of Industrial and Systems Engineering
University of Florida

In this talk we discuss new research directions in data mining - using network-based models for data analysis and decision making. In many practical situations, a real-world dataset can be represented as a large graph (network) with certain attributes associated with its vertices and edges. These attributes may contain specific information characterizing the given application, which often provides a new insight into the internal structure and patterns of the data. The considered examples include telecommunications, social networks, biomedicine, and finance.

Using Cluster Analysis to Relate Subjective and Objective Pharmacovigilance Association Measures

Ronald K. Pearson
ProSanos Corporation
San Diego, CA

The field of pharmacovigilance is concerned with the detection and interpretation of associations between drugs and adverse medical events that may or may not be caused or exacerbated by those drugs. In the U.S., the primary source of data for pharmacovigilance analysis is the FDA's Adverse Event Reporting System (AERS) database, which is organized by Individual Safety Reports (ISR's) listing the drugs a patient was taking, the adverse reactions they reported, and a limited amount of additional data (e.g., patient age, gender, reporting source information, etc.). Association between drugs and adverse events can be measured in a variety of ways, and this talk will consider three: two objective measures (the classical Reporting Ratio and the Statistical Unexpectedness, based on Fisher's exact test for association between drugs and adverse events), and one subjective measure (the Index of Suspicion, a numerical value between 0 and 1 computed from the AERS role code data, which classifies every drug appearing in an ISR as ``primary suspect,'' ``secondary suspect,'' ``interacting,'' or ``concomitant''). Examination of the AERS database for various combinations of drugs and adverse events suggests that these three association measures are related, but in a complicated manner. Thus, we are motivated to explore this relationship using partition-based cluster analysis. The talk will present a comprehensive assessment of clustering results obtained for representative sets of drug/adverse event pairs, examining the influence of variable transformations and scaling, the detection and treatment of outliers, and the use of different dissimilarity measures. Assessment of the number of clusters present in the data is based on a permutation reference strategy that has been discussed earlier (R.K. Pearson, T. Zylkin, J.S. Schwaber, and G.E. Gonye, ``Quantitative evaluation of clustering results using computational negative controls,'' Proc. 4th SIAM Intl. Conf. Data Mining, Lake Buena Vista, Fla., April 2004, pp. 188--199). Our ultimate objectives are first, to obtain useful insights into the reliability of spontaneous reporting systems like the AERS database and second, to construct an index that quantifies the tendency for some drugs to be subjectively blamed for adverse events even in the absence of objective evidence for an association with those events.

Classification vs Clustering, Analyzing Gene Functionality

Claudia Perlich
IBM T.J. Watson Research Center

In this work we contrast supervised relational learning and unsupervised clustering on the 2005 ILP challenge domain of classifying the yeast gene functionality. We represent the domains as a large network of genes and proteins, where the edges capture similarity information that was calculated using BLAST. We presents a statistical propositionalisation approach to relational classification and contrast the classification performance with clustering approaches, that use different pairwise similarities. We show that the advantage of the supervised approach is its ability to aggregate information from multiple sources and to optimize the relative weighting of the information.

Practical Fixed-Parameter Algorithms for Graph-Modeled Data Clustering

Sebastian Wernicke
Theoretical Computer Science,
Institut f¨ur Informatik, Friedrich-Schiller-Universit¨at Jena
Ernst-Abbe-Platz 2, D-07743 Jena, Germany

Most problems concerned with the detection of cluster structures in a graph are known to be NP-complete. For some NP-complete problems, the concept of *fixed-parameter tractability* (FPT) offers an exact alternative to commonly employed heuristic approaches because it is not the size of a problem that makes it hard, but the *structure* of an instance. This *structural hardness* is expressed as an integer variable k, the so-called parameter. A size-n instance of a fixed-parameter tractable problem can then be solved in time $f(k)\cdot p(n)$ where f is a function solely depending on k and $p(n)$ is a polynomial in n. Whenever k turns out to be small, fixed-parameter algorithms can solve the problem quite fast (sometimes even in linear time).

The purpose of the talk is twofold: First, we wish to provide a primer about practical FPT techniques. In our opinion, these belong into the standard toolkit when dealing with clustering problems. Second, we present three concrete case studies from the realm of graph-modeled clustering:

- *Clique.* Using techniques that were originally developed for the fixed-parameter tractable Vertex Cover problem, it is possible to detect a size-$(n-k)$ clique in an n-vertex graph in $O(1.3^k + kn + m)$ worst-case time.

- *Cluster Deletion and Cluster Editing.* Here, the assumption is that the underlying correct cluster structure is a graph that is a disjoint union of cliques, distorted by removing (*Cluster Deletion*) or removing and adding (*Cluster Editing*) at most k edges. The underlying cluster structure of an n-vertex graph can be found in $O(1.77^k + n^3)$ time and $O(2.27^k + n^3)$ time, respectively.

- *Clique Covering.* The assumed underlying cluster structure is an overlapping union of cliques; the task is to cover the edges of a graph with a minimum number of cliques. We present efficient algorithms that allow for *optimal* problem solutions in time competitive with common heuristics.

Practical experiments suggest that the algorithms perform much better on real-world data than the provable worst-case bounds suggest. Thus, for some NP-complete clustering problems, FPT offers algorithms which are both efficient and deliver optimal solutions.

Fast and Effective Clustering Very Large Networks Using Density-Based Clustering Algorithm

Xiaowei Xu[1]*, Zhidan Feng[1], and Tom Schweiger[2]

[1]Department of Information Science, University of Arkansas at Little Rock
[2]Acxiom Corporation, Fayetteville, AR 72701

Density-based clustering algorithm is originally designed to detect clusters of arbitrary shape as well as to distinguish noise in spatial and multi-dimensional databases. Technically, the algorithm is based on region queries, which can be supported efficiently by spatial index structures such as R-trees (at least, if the dimension of the data space is not too high). Since first introduced in 1996, density-based clustering algorithm has been successfully applied in a wide variety of applications, such as spatial data mining, image processing, web mining, text mining, and genomics. To our best knowledge, however, there is no report on its application to network clustering. In this paper we report our experiences of applying density-based approach for the analysis of very large networks including social networks, scientific collaboration networks, customer networks and biological networks. We first review the basic concepts of density-based clustering algorithm. Then a hierarchical density-based clustering algorithm, which is preferable because of the hierarchical structural nature of the networks, is presented. While a density defined by the number of points in a circle centered to the given point seems to be a natural definition for geometric data, it is obviously not suitable for the network data. This fact leads to a density concept based on the neighborhood of vertices and a similarity measure defined by the fraction of shared neighbors for the application of density-based clustering approach to the network data. Since the density-based algorithm needs exactly one neighborhood search for each vertex and the neighborhood search has a constant cost using the adjacency list for the sparse graphs, which are the most popular networks in the real world, the complexity of density-based clustering algorithm is linear to the number of vertices in the networks. We conducted extensive experiments to evaluate our algorithm by using both synthetic and real network datasets. The results show that our approach outperforms the modularity-based network clustering approach in both the speed and the accuracy.

Maximal Hyperclique Pattern in Gene Profiles

Yaochun Huang and Weili Wu*
Department of Computer Science
The University of Texas at Dallas

Different with the traditional data set, gene profiles are the datasets with continuous attributes. The relation values between genes and samples are not bit values. They could be any real number, no matter positive or negative. If we partition this kind of datasets into bit value datasets, there will be potential information loss. Here, by extending the work of Hans min-apriori approach, we develop a new normalization method for the bio datasets.

Hyperclique pattern are groups of objects which are strongly related to each other. Hyperclique patterns in bio dataset are very interesting since genes or samples will be very similar if they are in same patterns. We introduce algorithms for mining maximal hyperclique patterns in large datasets containing quantitative attributes. In addition, we adopt the algorithm structures of three popular association pattern mining algorithms and add a critical clique pruning technique. Finally, we mine quantitative maximal hyperclique patterns in gene profiles with the algorithms.

Self-Organizing Maps for Brain Electrical Activity Classification

Wei Zeng* and W. Art Chaovalitwongse
Department of Industrial and Systems Engineering
Rutgers University

Self-organizing map (SOM) is widely used as an unsupervised neural network. Introduced by in the 90’s, SOM has been successfully applied to research problems in many areas, including data visualization signal processing, image retrieval, speech recognition, etc. Having the ability to preserve inner structure of input data, SOM can be used as a clustering tool. However, most applications deal with static data and use Euclidian distance as a measure of similarity between input data and weight of neurons in SOM grid. However, this process incurs the loss of data characteristics when applied to time series data to find similar patterns or clusters. In this talk, we propose a Dynamic Time Warping (DTW) technique to be incorporated with SOM algorithm. DTW has a capability of capturing patterns in the time series data, which will be very useful in clustering and classification of time series. We will demonstrate that traditional SOM combined with DTW similarity measure can yield better performance than Euclidean distance. In addition, we will propose an implementation of SOM as a time series classifier. The main application of this technique lies on the classification of brain electrical activity, especially electroencephalogram (EEG). The brain networks are very complex and difficult to understand. In epilepsy research, there is a need of powerful techniques used to differentiate/classify normal and epileptic brain activities. The SOM clustering and classification techniques will be applied to EEG signals acquired from patients with epilepsy to study the brain networks for that purpose.